8th Mediterranean Morphology Meeting

Proceedings

"Morphology and the architecture of the grammar"

Editors: A. Ralli, G. Booij, S. Scalise & A. Karasimos

University of Cagliari
University of Bologna
University of Patras
University of Leiden

M*orphology* and the **A***rchitecture* of **G***rammar*

On-line Proceedings of the
E*ighth* **M***editerranean* **M***orphology* **M***eeting*
(MMM8)
Cagliari, Italy, 14-17 September 2011

E*dited by*:
A*ngela* **R***alli,* **G***eert* **B***ooij,* **S***ergio* **S***calise* and
A*thanasios* **K***arasimos*

URL: http://morbo.lingue.unibo.it/mmm

2012 University of Patras, Greece
On-line Proceedings of the Eighth Mediterranean Morphology Meeting (MMM8)
ISSN: 1826-7491

E-Proceedings Board

Contact Information:

Centre of Modern Greek Dialects
Department of Philology
University of Patras
26504 Rio, Greece
http://www.philology-upatras.gr
Tel/Fax: +30 2610 969375
email: lmgd@upatras.gr (Centre of Modern Greek Dialects)
ralli@upatras.gr (Angela Ralli)
akarasimos@upatras.gr (Athanasios Karasimos)

ISSN: 1826-7491

Contents

CONTENTS 4

FOREWORD 6

ARTEMIS ALEXIADOU & FABIENNE MARTIN
Competing affixes as aspectual morphemes:
The case of deadjectival nominalizations 8

RAFFAELA BAECHLER & GUIDO SEILER
Simplification, complexification, and microvariation:
Towards a quantification of inflectional complexity in closely related varieties 23

ALEXANDRA BAGASHEVA
On the classification of compound verbs 42

MELANIE BELL
The English noun-noun construct: a morphological and syntactic object 60

MARIS CAMILLERI
Morphological Complexity in Maltese:
A divergence from canonicity 92

CLAUDIA MARZI, MARCELLO FERRO & VITO PIRRELLI
Prediction and Generalisation in Word Processing and Storage 114

ALICE HARRIS
Meaning resides in fully inflected forms:
The Georgian "unwillingness" construction 132

ATHANASIOS KARASIMOS
'One Allomorph to rule them all':
The Single Allomorph Selection Constraint in Greek 143

HIDEKI KISHIMOTO
Complex adjectives at the morphology-syntax interface 151

EWA KONIECZNA
Analogical modelling and paradigmatic word-formation
as attention-seeking devices 167

MARK LINDSAY
Rival suffixes:
synonymy, competition, and the emergence of productivity 190

HEIKE NECKER & LIANA TRONCI
Italian -ismo/-ista and Ancient Greek -ismos/-istes formations.
Morphological processes and diachronic relationships 202

CHIARA SEMPLICINI
Gender selection and syntactic constructions:
the case of Dutch double gender nouns 234

GREGORY STUMP
The formal and functional architecture of inflectional morphology 254

NIINA NING ZHANG
Encoding Unexpectedness by Aspect Inflection 269

GÉRALDINE WALTHER
Fitting into morphological structure:
accounting for Sorani Kurdish endoclitics 300

CAROLA TRIPS
Empirical and theoretical aspects of phrasal compounds:
against the "syntax explains it all" attitude 322

Foreword

The Mediterranean Morphology Meetings (MMM) are organised by Geert Booij, (University of Leiden), Angela Ralli (University of Patras), and Sergio Scalise (University of Bologna). For each meeting they cooperate with a local organiser. In 2011, the 8th Mediterranean Morphology Meeting was organised by the University of Cagliari, and the local organiser was Ignazio Putzu.

The aim of MMM is to bring together linguists who work on morphology in an informal setting, which guarantees maximal interaction between researchers and gives young linguists the chance to present their work at a conference of moderate size, where fruitful contacts with senior linguists can be established.

MMM meetings traditionally comprise one theme-free day and one day devoted to a special theme, which in 2011 was: Morphology and the Architecture of Grammar.

Previous meetings:
MMM1 1997 - MYTILENE, ISLAND OF LESBOS, GREECE
Topics: *Allomorphy, Compounding, Inflection*
Invited Speakers: Anna Anastassiadis-Simeonidis, Mark Aronoff, Andrew Spencer
Proceedings published in paper by the University of Patras.
Editors: Geert Booij, Angela Ralli, Sergio Scalise. Patras: University of Patras, 1998

MMM2 1999 – LIJA, MALTA
Topics: *The role of lexical categories versus non-lexical categories in morphology* || *The interface of morphology and phonology*
Invited Speakers: Greville G. Corbett, Ferenc Kiefer, Marianne Mithun.

MMM3 2001 – BARCELONA, SPAIN
Topics: *The borderline between syntax and morphology* || *The role of prosodic constraints in morphology*
Invited Speakers: Geert Booij, Anna-Maria Di Sciullo, Soledad Varela
Proceedings published in paper by the Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.
Editors: Geert Booij, Janet DeCesaris, Angela Ralli, Sergio Scalise. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra 2003. ISBN: 84 477 0857-8. MMM3 also paid a tribute to the late Danielle Corbin

MMM4 2003 – CATANIA, ISLAND OF SICILY, ITALY
Topic: *Morphology and Linguistic Typology*
Invited Speakers: Wolfgang Dressler, Paul Kiparsky, Franz Rainer
Proceedings published on the web: http://mmm.lingue.unibo.it/.

MMM5 2005 – FRÉJUS, FRANCE
Topic: *Lexical Integrity Hypothesis*
Invited Speakers: Denis Creissels, Brian D. Joseph, Rochelle Lieber & Sergio Scalise
Proceedings published on the web: http://mmm.lingue.unibo.it/.

MMM6 2007 – ITHACA, GREECE
Topic: *Morphology and Dialectology*
Invited Speakers: Taro Kageyama, Ingo Plag, Angela Ralli, Peter Trudgill
Proceedings published on the web: http://www.philology.upatras.gr/LMGD/el/

research/downloads/MMM6_Proceedings.pdf
Selected papers are also published in Morphology (Special issue: Morphology meets Dialectology, edited by Geert Booij, Angela Ralli and Sergio Scalise)

MMM7 2008 – NICOSIA, CYPRUS
Topic: *Morphology and Diachrony*
Invited Speakers: Geert Booij, Östen Dahl, Nigel Vincent
Proceedings published on the web: http://www.philology.upatras.gr/LMGD/el/research/downloads/MMM7_Proceedings.pdf

MMM8 – Cagliari Sardegna, Italy
Topic: *Morphology and the Architecture of Grammar*
Special Topic: *The Morphology of Sardenian*
Invited Speakers: Farrell Ackerman, Angela Ralli, Gregory Stump
Proceedings published on the web: http://www.philology.upatras.gr/LMGD/el/research/downloads/MMM8_Proceedings.pdf.
Selected papers of the Special Topic appear in *Lingue e Linguaggio* (2012, number I)

The permanent scientific committee

Geert Booij
Angela Ralli
Sergio Scalise

Competing affixes as aspectual morphemes: The case of deadjectival nominalizations

Artemis Alexiadou and Fabienne Martin
University of Stuttgart
artemis@ifla.uni-stuttgart.de fabienne.martin@ling.uni-stuttgart.de

1. Introduction

The phenomenon of deriving sets of two or more (near) synonymous words from the same stem with different affixes is rather common crosslinguistically (Booij 1977, Scalise 1984). The relation between rival morphological processes can be complex and very diverse (see e.g. van Marle 1985, 1986) and raises the following concern. Given the Blocking Effect (Aronoff 1976) and that competing affixes regularly differ regarding their productivity and distribution (Corbin 1984, van Marle 1985), the occurrence of doublets or triplets of this type is in principle predicted to be marginal. This paper focuses on one counter-example to this prediction, namely deadjectival nominalizations in French. Our main hypothesis is that the most productive deadjectival suffixes in French, namely *-ité*, *-tude*, *-erie* and *-isme*, functionally differ from each other in terms of their respective aspectual values: they play with respect to the adjectival stem a role similar to inflectional aspectual morphemes.

We will focus on the distribution of suffixes among *dispositional nouns* derived from evaluative adjectives like *stupide* 'stupid', and compare the aspectual readings of the adjectival stem with those of the derived noun.[1] Doublets or triplets derived from the same stem with different suffixes will be the object of particular attention, since the suffix should be the only element responsible of potential switchings in the interpretation. We choose to focus on dispositional nouns because their adjectival counterparts display a rich aspectual polysemy (cf. Fernald 1999, Geuder 2000, Martin 2008) – and thus allow to test the aforementioned hypothesis – and because doublets and triplets are quite frequent in this lexical domain. However, despite of this specific focus, we will take into account nouns from other lexical domains for the generalisations proposed.

If productive deadjectival suffixes differ from each other by their aspectual value, we can better explain the existence of dictionary[2]doublets or triplets derived from the same stem, cf. (1).

(1) a. coquetterie/ coquettisme (>coquet 'coquettish')
b. drôlerie/ drôlisme (>drôle 'funny')
c. fanfaronnerie/ fanfaronisme (>fanfaron 'swanky')
d. crapulerie/crapulisme (>crapule 'scoundrel')
e. crétinerie/ crétinisme (>crétin 'moron')
f. importunité/ importunisme (>importun 'out of place')

The same hypothesis also explains the high number of neologisms in the field of dispositional nouns. Incompetence cannot explain it alone, because neologisms are often used in the neighbourhood of the competing dictionary variant. If deadjectival suffixes

[1] We will call nouns derived from evaluative adjectives *dispositional nouns* rather than *quality nouns*, because "quality" has a larger extension than "disposition" (there are qualities, like e.g. colors, that are not dispositionsà.

2 We will call *dictionary words* words which are recorded in dictionaries.

differ by their aspectual value, it is easier to account for the creation of new forms and to explain which suffix is selected for which needs.

That competing suffixes functionally differ by their aspectual profile has already been argued for other kinds of nominalisations, cf. e.g. Alexiadou (2001) about *-ing* nominalisations in English and their counterparts in Greek, Dubois (1962), Martin (2010) or Uth (2011) about *-age, -ment* and *-ion* in French, as well as Ferret, Soare and Villoing (2011) about *-ée* and *-age* in French. In the field of deadjectival nouns, the fine grained descriptive study of Daude (2002) of Latin *nomina qualitatis* already suggests that the Latin ancestors of the French suffixes at study also compete by their aspectual value, and it will be shown in Section 3 that French reflects some aspects of the interplay between Latin competing suffixes.

The present study makes use of two types of data. Firstly, the different readings of 170 deadjectival dispositional nouns have been manually identified and classified on the basis of several tests presented in Section 2. Secondly, the productivity of the suffixes we are interested in was roughly appreciated through the use on neologisms. Dispositional nouns presented on the Internet but not stored in dictionaries were automatically collected by L. Tanguy at the ERSS Laboratory of the University of Toulouse. For each item of a list of 1000 evaluative adjectives, a list of nine possible nouns combining the adjectival stem and one of the suffixes at study was automatically generated following Hathout's technique (Tanguy & Hathout 2007). From the generated forms were automatically discarded all nouns present in le *Lexique des formes fléchies du français* or le *Trésor de la langue française.* The remaining forms which occur 1 to 200 times on the Internet (in presence of the adjectival stem) were collected with the help of Webbafix (Hathout and Tanguy 2002). A part of the output list has been cleaned manually. We discarded non French words (or produced by speakers which are not native speakers of French), non nominal forms, mispelled words, typos, hapax as well as words judged unacceptable by three native speakers to which I submitted a pre-cleaned list. For 110 dictionary words analysed for the study, 159 neologisms were identified. The table below summarises the distribution of suffixes among them. It shows that 34,3% of the neologisms are built with *-itude*, 25,6% with *-ité*, 18,1% with *-isme* and 13,7% with *-erie*. The other suffixes are hardly used to create new words.[3] Although these data remain to be confirmed by a research on a larger scale, we provisionally conclude that *-itude*, *-ité, -isme* and *-erie* are the main productive deadjectival suffixes in French. All dictionary words and neologisms used in this study are given in the Appendix.

Suffixes	***itude***	***ité***	**--**	***erie***	***isme***	***eur***	***ion***	***ance***	***ie***	***ise***	***esse***	***ice***
Dictionary words (total:110)	0	32	19	15	10	7	7	6	6	4	3	1
	0%	29%	17,2%	13,6%	9%	6,3%	6,3%	5,4%	5,4%	3,6%	2,7%	0,9%
Neologisms (total:159)	55	29	--	22	41	0	0	6	1	3	2	0
	34,3%	18,2%	--	13,7%	25,6%	0%	0%	3,7%	0,6%	1,8%	1,2%	0%

The paper is structured as follows. Section 2 focuses on the aspectual value of the suffixes *-isme*, *-erie*, *-itude* and *-ité.* Section 3 addresses their morpho-syntactic properties. It presents data that suggest that these four suffixes differ from the other non-productive suffixes by their level of attachment (Kiparsky 1982, Marantz 2001) and show how the differences in their morpho-syntactic composition can account for the differences in their interpretation.

2. The aspectual value of deadjectival suffixes

[3] *-ance* might be an exception, see Dal & Namer (2010).

As already mentioned in the introduction, the hypothesis we argue for is that the most productive deadjectival suffixes have each a specific aspectual value by which they functionally differ from each other and by which they contribute to the aspectual interpretation of the nouns containing them. Before presenting data in favour of this claim, we briefly delineate the different readings exhibited by dispositional nouns (see Martin 2012 for details).

A noun like *stupidité* can denote eventualities, dispositions or entities (i.e. a stupid book). When used as eventuality-denoting predicate, it can have individual-level (permanent) or stage-level (transient) readings. In the former case, it denotes dispositions, habits or stable tendencies. In the latter, it describes either an event (a stupid act) or a transient state (the state the individual is in when acting stupidly).

The range of readings a dispositional noun can have significantly varies with the suffix chosen. Among the 170 DNs analysed manually, all have an individual level reading. With respect to the stage level readings (stative or eventive), all *-erie* DNs have it, while 68% of *-ité* DNs get it, and 20% of *-isme* DNs. Among these stage-level DNs, 100% of *-erie* DNs have the eventive reading, while only 20,7% of *-ité* DNs get it, and none of the *-isme* DNs. Of course, all nouns which have the eventive reading have a temporary (stage-level) reading, but the reverse is not true; some nouns have a temporary reading, but no eventive one.

Tests used for the classification were, for the eventive reading, (i) the compatibility with *faire* 'do' or *commettre* 'commit' or (ii) the compatibility with *avoir lieu/ prendre place* 'take place' + spatial PP. For the temporary readings in general (stative or eventive), the tests used were (i) the embeddability in episodic perception reports, (ii) the availability of the iterative interpretation and (iii) the possibility to denote eventualities whose temporal trace equals the one of an event (see Martin 2012 for details).

We take these data to go against the claim that DNs are all aspectually underspecified at the lexical level (as e.g. by Beauseroy 2009:129 for French), since the suffix plays a role in the range of aspectual values a QN can have.

The next subsections are devoted to the aspectual value of each of the most productive deadjectival suffixes, namely *-isme*, *-erie*, *-itude* and *-ité.*

2.1. -isme

As a rule, DNs built with the suffix *-isme* exclusively denote dispositions, habits or tendencies and thus only get individual-level readings.[4] They therefore have a strong bias towards permanency, since properties of this kind are by default conceived as permanent. If they can nevertheless get an episodic interpretation, they still have to denote a (transient) disposition, see Martin 2012 for details.

A first piece of evidence for this claim is that in general, *-isme* DNs cannot get an iterative interpretation: firstly, virtually no *-isme* DN can be pluralized, cf. (2a); secondly, only few of them (e.g. *optimisme, angélisme,* or *héroïsme*) can be modified by *répété*, cf. (2b) vs (2c):

(2) (a) * Les despotismes/ infantilismes/ frénétismes/ érotismes/ égoïsmes/ diabolismes/ cynismes/ drôlismes/ crétinismes/ crapulismes/ coquettismes/ arrivismes/ optimismes/ héroïsmes/ angélismes...de Pierre.
'The despotisms/ childishnesses/ frenetisms/ erotisms/ égoïsms/ diabolisms/ cynicisms/ funny-isms/ cretinisms/ scoundrel-isms/ coquettish-isms/ carrierisms/ optimisms/ héroïsms/ angelisms... of Pierre.'

[4] According to Rita Manzini (p.c.), what we claim for French *-isme* DNs seems to be true for Italian *-ismo* DNs.

(b) ??Son despotisme / infantilisme/ frénétisme/ érotisme/ égoïsme/ drôlisme/ diabolisme/ cynisme/ crétinisme/ crapulisme/ coquettisme/ arrivisme ... répété.
'His repeated despotism, ...'
(c) votre optimisme répété concernant les recettes prévisionnelles (Internet)
'your repeated optimism with regard to the projected takings'

The rule of suffixation must be responsible of this constraint, since often, competing DNs derived from the same adjectival stems but with another suffix accept the iterative interpretation. For instance, DNs built with the suffixes *-erie* or *-age* in (3) can be pluralized and modified by *répété*.

(3) (a) Ses crétin**eries**/ coquette**ries**/ drôl**eries/** enfantill**ages**
'Her scoundrel-**eries**/ coquettish-**eries**/ funny-**eries**/ childish- **eries**.'
(b) Sa crétin**erie**/ coquette**rie**/ drôl**erie/** son enfantill**age** répété(e)
'Her repeated scoundrel-**erie**/ coquettish-**erie**/ funny-**erie**/ childish- **erie**.'

A second piece of evidence for the claim that *-isme* DNs are always individual-level and thus cannot denote transient states is that they are strange in episodic perception reports, cf. (4b) vs (4a).

(4) (a) Ce matin, j'ai assisté à sa crétin**erie**/ coquette**rie**/ drôl**erie/** son enfantill**age**.
'This morning, I witnessed her scoundrel-**erie**/ coquettish-**erie**/ funny-**erie**/ childish-**erie**.'
(b) ??Ce matin, j'ai assisté à son crétini**sme**/ coquett**isme**/ drôl**isme**/ infantil**isme**.
'This morning, I witnessed her scoundrel-**isme**/ coquettish-**isme**/ funny-**isme**/ childish-**isme**.'

Thirdly, when attached to adjectival stems which preferentially have a temporary reading like *nu* 'naked', *-isme* makes the individual-level reading of the derived noun compulsory. For instance, *nudisme* 'nudism' has to refer to a disposition rather than to a particular state, which is well rendered by its German translation *Nacktkultur*.

In conclusion, *-isme* tends to univocally attribute the individual-level reading to the deadjectival noun, including with adjectival stems which could in principle be attributed other readings, or preferentially have a stage-level reading. If-*isme* DNs nevertheless get transient interpretations, they still have to denote dispositions, habits or stable tendencies.

The exact scope of the claim remains to be evaluated though. We already pointed out that DNs like *optimisme* falsify what seems to us to be the general rule (they *can* be used to denote transient states). A more systematic analysis of *-isme* DNs remains to be done to evaluate how exceptional are these nouns.

2.2. -erie

As a rule, dispositionals noun composed with the 'abstract' *-erie* suffix (as opposed to the locative one found e.g. in *brasserie* 'brassery') have an eventive reading.[5] This suffix can not only attach to adjectival stems, but also to nominal ones as in *ânerie* 'stupidity/ rubbish', derived from *âne* 'donkey'. The eventive reading is also the only reading that *any -erie* noun can have, included neologisms. We take this as an indication of the fact that this reading is the basic one.

[5] There might be some exceptions to this, as e.g. *gloutonnerie* 'gluttony'.

Given that most stems of *-erie* nouns do not have an eventive reading by themselves – evaluative adjectives do not have it, cf. Martin 2008, and nouns like *âne* certainly do not either – the rule of suffixation in *-erie* must be the element that introduces the event argument.

Interestingly, the stem cannot systematically predicate an event, cf. (5).

(5) (a) Son acte/ ce qu'il a fait est ??gredin (OK est une gredinerie)
'His act/ what he did is mischievous (is a mischievous-erie).'
(b) Son acte/ ce qu'il a fait est ??fainéant (OK est une fainéanterie)
'His act/ what he did is lazy (is a lazy-erie).'

This suggests that *-erie* nouns on the eventive reading cannot be paraphrased by the string "the event which is P", P being the stem (this is one of the possible meanings discussed by Drapeau and Boulanger 1982). A more appropriate paraphrase is something like 'the event involving an entity *x* which is P'.

That *-erie* works as an "eventizer" is perhaps related to the presence of the infinitival morpheme *-er* in its composition. In fact, according to some authors, *-ie* is post-verbal, and *-er* is etymologically present in *-erie* (cf. Moldenhauer 1934, Bécherel 1981), but this uncorrectly predicts that there is a verbal form for each *-erie* noun. However, whatever its etymology is, it might be that *-er* is reanalysed as the infinitival morpheme in the suffix by speakers of French.[6]

Nouns in *-erie* also have an individual-level reading that we assume to be derived from the basic eventive reading by a mechanism like coercion. That the permanent reading is not basic is confirmed by the fact that the eventivity of *-erie* still surfaces when it is selected. Indeed, when *-erie* Ns are used to describe a permanent property, there is a tendency to assume that this property is actualised through concrete events. In other words, under their individual-level reading, *-erie* Ns are rather interpreted as habit-denoting than disposition-denoting nouns. This is not the case of the permanent reading expressed by *-ité* or *-isme* nouns. This intuition is difficult to illustrate, but the contrasts in (6) and (7) serve as an attempt: (6a) and (7a) are not contradictory because they denote 'classical' dispositions, that individuals can have without instantiating them in concrete manifestations, while (6b) and (7b) are.

(6) (a) Sa sensibil**ité** n'a jamais vraiment l'occasion de se manifester.
'His sensible-ité never really has the opportunity to manifest itself.'
(b) ??Sa sensibl**erie** n'a jamais vraiment l'occasion de se manifester.
'His sensible-erie never really has the opportunity to manifest itself.'
(7) (a) Heureusement, sa tendance maniaque n'a jamais l'occasion de se manifester.
''Fortunately, his maniac tendency never has the opportunity to manifest itself.'
(b) ??Heureusement, sa maniaqu**erie** n'a jamais l'occasion de se manifester.
''Fortunately, his maniac-erie tendency never has the opportunity to manifest itself.'

Interestingly, Daude (2002) already contrasts the Latin suffixes *-ia* from which *-erie* is derived and *-tas* (the ancestor of *-ité*) in a similar way. He claims that *-ia* dispositional nouns "are not completely abstract" because they "summarise a set of manifestations" of the quality, and are typically attributed to the individual on the basis of these concrete instantiations. This is how the habitual reading is distinguished from the dispositional one: it is an inductive generalisation inferred from observed behaviors.

2.3. -itude

[6] On the history of *-erie*, see e.g. the in-depth study of Hüning (1999).

According to our searches in corpora, *-itude* is one of the most productive deadjectival suffixes, which suggests that Bécherel (1981)'s claim that this suffix is unproductive is not correct or at least no longer valid.[7] The idea that *-itude* enjoys a revival is also argued by Koehl (2012a, 2012b), on the basis of much larger corpora than ours.

A specificity of *-itude* DNs is that they must denote a property of animates (as opposed to objects or events). This has already been observed by Rainer (1989:312) for the Italian suffix *-itudine*, cf. his examples (8).

(8) (a) La grati**tudin**e di Paolo verso Paola/??della tua visita.
La grat**itude** de Paolo envers Paola/??de ta visite.
'The grateful-itude of Paolo towards Paola/of your visit.'
(b) La rett**itudina** di Paolo/ ??della tua pronuncia.
La rect**itude** de Paolo/ ??de ta prononciation.
'The straight-itude of Paolo/ of your pronunciation.'

Besides, like *-isme* Ns, *-itude* ones do not easily get stage-level readings (eventive or stative), cf. (9)-(10). Observe that this is not always true of the corresponding adjectival stem (for instance, *être con* 'to be stupid' certainly has a stage-level reading).

(9) (a) Sa conn**itude** est sans bornes. (Internet)
'His stupid-**i**tude is infinite.'
(b) J'ai vu/ assisté à sa conn**erie**/# conn**itude**.
'I saw/ witnessed his stupid -erie/ stupid -itude.'
(c) Il a fait une conn**erie**/ # une conn**itude.**
'He made a stupid -erie/ a stupid -itude.'

(10) (a) Notre potentiel d'inhuman**itude** me dégoûte. (Internet)
'Our potential of inhuman-itude disgusts me.'
(b) Il a commis une inhuman**ité** (#inhuman**itude**) et une injustice plus grande que celle du précédent. (Internet)
'He committed a inhuman-ité (inhuman-itude) and an injustice greater than the former.'

On the individual-level reading, *-itude* DNs have a particular flavour which distinguishes them from *-ité* or *-erie* Ns on the same reading. This specificity of *-itude* nouns has been observed at several places. Bécherel (1981) claims that *-itude* Ns are more 'concrete' than corresponding *-ité* ones; the same intuition is reflected in Senghor's words about Césaire's use of *négritude*, which he finds 'more concrete' than *négrité* (Senghor 1977). For Latin, Daude (2002) claims that *-tudo* (from which *-itude* is derived) actualises more than *-tas* (which is more abstract), and suggests an exercise, a putting into a practice: "*nomina qualitatis* in *-tudo* tend to express traits of character or dispositions *determining a behaviour*" (italics and translation mines). Daude's claim about Latin is echoed in the Wikipedia entry devoted to the French suffix: '*-itude* serves to form words implying the idea of an attitude, a pose explicitly adopted, in opposition to the intrinsic […] quality designated by the noun or adjectival stem' (translation mine).

[7] We discarded from the counting neologisms in *-itude* like *intelligentitude*, *constantitiude* or *incongruitude*, rejected by my informants, and exclusively used as a parody of the politician Ségolène Royal in her use of *bravitude* instead of *bravoure* in 2007 and at the source of a big media buzz (one million hits on Google, cf. http;//en.wikinews.org/wiki/Bravitude_climbing_fast_on_Google). One could argue that Royal's neologism emulates the use of this suffix, but the fact she produced it could also be seen as the reflection of the productivity of *-itude* at that time.

We claim that this value of *-itude* comes from the fact that the semantics of the words *habitude* and *attitude* is transferred to the *-itude* suffixation process, so that these Ns end up with denoting habits/ ways of being/ regular behaviours, that is more concrete entities than dispositions. In other words, the idea is that we deal here with an instance of what Rainer (2005) calls *irradiation* after Bréal (1892:20), that is a transfer of a semantic feature from a word meaning to a word formation meaning.[8] A first indication of this is that users of neologisms in *-itude* sometimes overtly link their lexical creation to the words of *attitude* or *habitude*, cf. e.g. (11):

(11) "L'human**itude**", une attitude partagée, comme une nouvelle façon de vivre ensemble. (Internet)
"The human-itude", a shared attitude, like a new way to live together.

Moreover, *-itude* DNs are more appropriate than *-ité* ones to univocally denote behaviours or habits. For instance, *belgitude* translates the concept of *Belgisch Sein* – Belgian ways of being, of behaving, Belgian habits, etc. – much better than *belgicité*, which can also simply describe the property of having the Belgian nationality. This second reading is not available to *belgitude.* For instance, *belgitude* cannot be attributed to somebody who has the Belgian nationality but never lived in Belgium and doesn't know anything about Belgian habits and culture.

To summarise, *-itude* DNs denote habits and attitudes. This explains why they are [+ANIM] (inanimates do not have habits or attitudes). It also explains why these DNs are preferably interpreted as individual-level predicates – habits and attitudes are properties rather than instantiations of properties.

2.4. -ité

Although *-ité* is less used than *-itude* among neologisms, it is the most frequent one if one considers dictionary words and neologisms altogether, (with a total of 22,5% of all Ns examined). This confirms Koehl's (2009) study on the distribution of deadjectival suffixes. It is also the most underspecified suffix, since *-ité* nouns can in principle have any of the readings delineated in the previous section. What is specific to *-ité* compared to the three other productive suffixes analysed above is that it does not seem to contribute by itself to the aspectual value of the created noun. The aspectual readings of the derived noun is much more dependent from the readings displayed by the adjectival base. The permanent reading of dispositional nouns is always salient because the adjectival stem from which they derive systematically have a dispositional reading. For instance, *un homme loyal* 'a loyal man' preferably denotes a man which has the disposition to be loyal. And *loyauté* 'loyalty' unsurprisingly has a preference for the dispositional reading. For instance, *la loyauté de cet homme m'a surpris* 'the loyalty of this man surprised me' is by default understood as the expression of surprise about an individual-level property. But facts differ if we take an adjectival root that preferably selects the stage-level reading. For instance, *un homme nu* 'a naked man' preferably describes a man which is temporarily naked. And *la nudité de cet homme m'a surpris* 'the nakedness of this man surprised me' is by default understood as the expression of surprise about a stage-level property. (That *nudité* has difficulties to get the individual

[8] Rainer (2005) also claims that *-itude* nouns are an example of irradiation. But according to him, the 'irradiating' words are not *habitude* and *attitude*, but rather *nègre* 'neger' and *servitude*, whose semantics 'contaminates' the *-itude* word process formation through the word *négritude*, so that *-itude* Ns ended up designating oppressed social groups and their emancipatory aspirations. However, although this semantic path might be correct for a subset of *-itude* nouns directly created on Césaire's model, it does not capture the differences between ethnic nouns like *belgitude/ belgicité*, nor between Ns like *humanitude* and *humanité.*

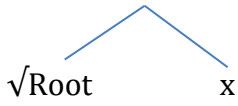
level reading is also confirmed by the oddity of *une manifestation de nudité* 'a manifestation of nakedness'.) In other words, *-ité* does not seem to carry its own aspectual feature. It is the unmarked form chosen for unmarked situations, while *-itude*, *-erie* and *-isme* are marked forms for marked situations – an example of a more general tendency that Horn (1984) calls "division of pragmatic labour".

3. Morpho-syntactical decomposition

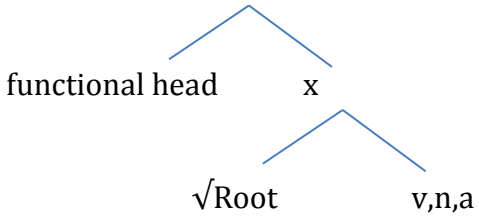
3.1. Root-attached vs. word-attached suffixes

Standardly, whenever two kinds of suffixes attach to the same root to derive a new word of the same category, they are taken to be distinguished according to the cycle of word-formation they take place in (Kiparsky 1982, Marantz 2001). From the perspective of Distributed Morphology (Marantz 2001, Embick 2010), suffixes that merge with roots are (i) idiosyncratic (no clear transparency in the composition of meaning), (ii) structure-changing (phonological changes are induced) and (iii) less productive. These morphemes attach within the root-cycle of word formation, and will be called 'root-attached' morphemes. On the other hand, suffixes that merge above functional heads are (i) semantically transparent (the compositional meaning can be predicted from the meaning of the parts), (ii) less structure changing and (iii) productive. These attach in the outer-cycle, and will be called 'word- attached suffixes', following Arad (2003).

(I) root-cycle



(II) outer-cycle attachment



word formation from roots *word formation from words*

Some suffixes always attach within the same cycle, cf. e.g. *-less*, *-ship, -ness, -hood*, which are univocally word-attached (Kiparsky 1982, Marantz 2001). Other suffixes can enter both cycles, cf. e.g *-able* as analysed by Aronoff 1976, but see Kiparsky 1982 about *-ability*. As a rule, a root-attached suffix cannot be attached to a stem containing a word-attached suffix (see e.g. Kiparsky's example **mongolismize*). But the reverse is possible (cf. *mongolianize*).

The fact that *-erie, -ité, -itude* and *-isme* are more productive and semantically transparent than the other deadjectival suffixes suggests that they are word-attached suffixes. Pseudo-words combining root-attached and word-attached suffixes point to the same conclusion. Indeed, pseudo-nominalizations built with one of these suffixes and one of the non-productive ones are judged more French when the non-productive suffix is attached first, cf. (12). Words built with suffixes attached in the reverse order sound less acceptable, cf. (13).

(12) °bénévolancerie °bouffonnessitude °agilancitude °fêtardisitude

(13) °bénévoleriance °bouffonitudesse °agilitudance °fêtarditudise

Note that if indeed *-ité* is word-attached in French, it is different from *-ity* in English, which has been argued to be root-attached by Kiparsky (1982) and Embick & Marantz (2008).[9] But this could be due to a difference in productivity between *-ité* and *-ity*: the former is the most used deadjectival suffix in French (Koehl 2009), while *-ity* is less productive than *-ness* in English (Bauer 2001).

3.2. Pluralization and episodic interpretation

The discussion in the previous sections can be summarised as follows. Dispositional nouns and the adjectival stems from which they derive do not have exactly the same aspectual readings. The affix is partly responsible for this semantic switch. The differences in the denotation between the four types of nouns can be summarised as follows. The suffix *-ité* is the unmarked productive suffix and can form Ns with any kind of aspectual interpretation. The suffix *-isme* tends to force the deadjectival noun to have a dispositional reading (but it remains to be evaluated how frequent are the exceptions to this rule). The suffix *-erie* imposes a preference for the eventive reading, but is compatible with any other readings, although, under the permanent reading, *-erie* nouns tend to denote habits rather than dispositions. The suffix *-itude* forces the noun to denote habits or attitudes and thereby imposes the feature of animacy and the individual-level reading.

In this section, we turn our attention to the availability of pluralization and iterative interpretation with deadjectival nominalizations. As mentioned, *-isme* DNs like *infantilisme* 'childishness' can be pluralized cf. (2). This is also generally true for *-itude* DNS (*habitude, attitude* and *aptitude* are among the few exceptions). On the other hand, DNs built with the suffix *-erie* can be pluralized, cf. (3). The empirical picture is more complicated with *-ité* DNs, and will not be investigated in detail here. In addition, we noted that in general, those nouns that can pluralize can have an event reading, while those that cannot are interpreted as individual level. The most prominent exceptions to this rule are nouns like *habitude* 'habit', *aptitude* 'aptitude', *capacité* 'ability'. These nouns exceptionally allow pluralization under the individual-level reading because the same individual can be ascribed several of the properties they denote (one can have several habits, but one normally does not have several dispositions to be childish).

Since pluralization is a nominal property, it will have to apply after the nominalization of the adjective has taken place. Still, however, we noted that pluralization is sensitive to the interpretation of the nominal, ±eventive. Following some of the recent literature (Borer 2005, Alexiadou, Iordăchioaia & Soare 2010, Alexiadou, Iordăchioaia & Schäfer 2011, Alexiadou 2011, Arche & Marín 2012), we take these differences to follow from the different morpho-syntactic structures associated with the various DNs. First, we take the ability of DNs to pluralize as evidence of the presence of a classifier projection (ClassP) within the nominal structure. From the perspective of Borer (2005), ClassP divides "stuff", and is the equivalent of Quantity projections in the verbal domain that make events heterogenous. ClassP is the input to NumberP. This projection introduces elements that count and can be seen as corresponding to the functional projection of Aspect (outer-Aspect) in the verbal domain, Alexiadou (2001), Arche & Marin (2012).

Building on Arche & Marín (2012) and Alexiadou (2011), we would like to propose that what is taken as the complement of a classifier is a structure that involves by default episodic eventualities (events or transient states) so that the classifier can denote different instantiations of eventualities. The fact that -isme or -itude DNs that are

[9]In English, German or Dutch, words in *-ity*, *-iteit* and *-ität* have been firstly borrowed from French and then went their own ways; now, these languages have a number of nouns built with these suffixes which do not have a French counterpart (cf. e.g.Booij 2009).

individual level do not pluralize suggests that when there is no transient eventuality, division introduced by the classifier is also *prima facie* not possible. The only way out is to pluralize a permanent property, which generally generates a pragmatic problem, except for nouns general enough like *habitude*. However, whenever there is an event or a transient state, the classifier projection can then introduce different instantiations of this eventuality.

The question that arises is what is the source of the transient reading. More particularly, one would like to know why the event reading is systematically available with *-erie* DNs. On the one hand, one could claim, following Ippolito's (1999) analysis of *-ata* nominalizations in Italian, that the morphological similarity between the infinitive and the *-erie* nominalization (highlighted by the presence of *–er* in both the verbal and the nominal environment) is due to the presence of the same syntactic structure. From this perspective, whener an event reading is available, this makes reference to the presence of a verbal stem in the morphological structure. The logic here would be that in view of the fact that the roots themselves do not have an event implication, this must be introduced by verbalizing the roots. The problem, though, is that without further refinements, this analysis predicts that there is a verbal form for every *-erie* noun, contrary to fact.

On the other hand, one could argue that the systematic availability of the transient reading with *-erie* DNs is due to the collective or frequentative value of this suffix, observed a.o Spitzer 1931:30 (cf. also the entry devoted to *-erie* in the *TLF*). Indeed, if *-erie* encodes a plural operator, it brings about on its own the interpretation which can fit this operator in the domain of eventualities, namely the transient reading. However,while this account explains why *-erie* DNs systematically have a *transient* reading, it doesn't explain why the *event* reading is also systematically available.

In favour of the idea that *-erie* DNs involves a kind of plural operator, we observe that they often can be used to denote a plurality of events with the singular.[10] For instance, the singular noun *fanfaronnerie* (from *fanfaron* 'boastful') can be used to describe several acts performed in a boastfulway.[11] On the other hand, our claim that plurality helps to bring about the transient interpretation is supported by the fact that the plural morphology is sometimes required for the transient interpretation of DNs. This is for instance the case of *finesse, familiarité* or *impudeur.*

Finally, we observe that although *-isme* nouns cannot be pluralized, they are sometimes OK with the adjective *constant*, but not with *frequent*:

(14) Son infantilisme/cynisme constant les exaspère.
His constant childiness/cynism get on their nerves.
(15) ??Son infantilisme/cynisme fréquent les exaspère.
His frequent childiness/cynism get on their nerves.

If the latter adjective is sensitive to the presence of a ClassP in the structure of the noun, while the former signals the presence of NumberP, the above contrast suggests that these nouns, like mass nouns, can appear together with adjectives that do not require different types of eventualities.

[10] That *-erie* DNs prefer the habitual reading when used to denote permanent properties can also be seen as a reflection of this plural operator.

[11] See also the deverbal noun *tuerie* 'killings (pl.)'. *-ade* nouns derived from evaluative adjectives also systematically have an eventive reading. But they do not have the iterative value of *-erie*, and thus must be pluralised in order to denote a series of acts. We would say that a *fanfaronnerie* is made of several *fanfaronnades* rather than the reverse.

References

Alexiadou, A. 2001. *Functional structures in nominals: nominalization and ergativity*. Amsterdam: Benjamins.

Alexiadou, A. 2011. "The Aspectual Properties of Nominalizations Structures", in *Morphology and its interfaces*, A. Galani *et al.* (ed.), Amsterdam: John Benjamins.

Alexiadou, A., G. Iordăchioaia & F. Schäfer, 2011, "Scaling the Variation in Romance and Germanic Nominalizations", in H. Perridon & P. Sleeman (eds.), *Variation and Change in the Romance and Germanic Noun Phrase*. John Benjamins, 25-40.

Alexiadou, A., G. Iordachioaia & E. Soare 2010. „Number/Aspect interactions in the syntax of nominalizations: a distributed Morphology approach. *Journal of Linguistics* 46: 537-574.

Arche, M. and R. Marín, 2012, "Evaluative Adjectives at the Light of their Nominalizations", talk presented at the *Linguistic Symposium on Romance Languages*, University of Southern, April 20-22.

Arad, M., 2003. Locality constraints in the interpretation of roots. Natural language and linguistic theory, 21: 737-778.

Aronoff, M. 1976. *Word formation in generative grammar*. Cambridge, Mass: MIT Press.

Bauer, L., 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Borer, H., 2005. *Structuring Sense*. Oxford: Oxford University Press.

Beauseroy, D. 2009. *Syntaxe et sémantique des noms abstraits statifs : des propriétés verbales ou adjectivales aux propriétés nominales*. PhD thesis, Université de Nancy.

Booij, G., 1977. *Dutch Morphology. A Study of Word Formation in Generative Grammar*, Dordrecht: Foris.

Booij, G., 2009. "Morphological analysis", in Heine, B. and H. Narrog (eds), *The Oxford Handbook of Grammatical Analysis*, Oxford, Oxford University Press, 563-589.

Bréal, M. 1892. "De l'irradiation grammaticale". *Mémoires de la Société de linguistique de Paris*, 7:20–23.

Bécherel, D. 1981. "Différenciation morpho-sémantique des suffixes nominalisateurs de l'adjectif". *Cahiers de lexicologie*, 38:45–59.

Corbin, D., 1984, «La forme et le sens: discussion », *Quaderni di Semantica*, 5, 288-302.

Dal, G. & F. Namer. 2010. "Les noms en *-ance/-ence* du français: quel(s) patron(s) constructionnel(s)? *"*, *in* Neveu, F. *et al.* (eds), *Actes en ligne du 2ème Congrès Mondial de Linguistique Française*, 893-907.

Daude, J. 2002. "Les substantifs abstraits de qualité", *in* Kircher-Durand, C. (ed), *Grammaire fondamentale du latin, t. IX. La création lexicale : la formation des noms par dérivation suffixale,* Pleuven/Paris: Peeters, 225-305.

Dubois, J. 1962. *Etude sur la dérivation suffixale en français moderne.* Paris: Nathan.

Drapeau, L. & A. Boulanger, 1982. "Les suffixes *-erie* en français populaire". *Revue québécoise de linguistique*, 11(2):73–91.

Embick, D. 2010. *Localism versus Globalism in Morphology and Phonology*, MIT Press.

Embick, D. & A. Marantz. 2008. "Architecture and Blocking". *Linguistic Inquiry*, 39/1:1-53.

Ferret, K., E. Soare & F. Villoing. 2011. "Rivalry between French *-age* and *-ée*: the role of grammatical aspect in nominalization", *in* Aloni, M., H. Bastiaanse, T. de Jager, K. Schulz, K. (eds), *Logic, language and meaning,* Amsterdam/New-York: Springer, 284-294.

Fernald, T. 1999. "Evidential coercion: Using individual-level predicates in stage-level environments". *Studies in the Linguistic Sciences*, 29/1:43–63.

Geuder, W. 2000. *Oriented adverbs*. PhD thesis, Universität Tuebingen.

Hathout, N. & L. Tanguy. 2002. "Webaffix: un outil d'acquisition morphologique dérivationnelle à partir du web", *in Proceedings of Taln 2002,* Nancy, 24-27 June 2002.

Horn, L. 1984. "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicatures", *in* Schiffrin (ed), *Meaning, form and use in context*, Washington: Georgetown University Press, 11-42.

Hüning, M. 1999. *Woordensmederij*. De geschiedenis van het suffix *-erij*. PhD Thesis, Universiteit Leiden.

Ippolito, M. 1999. On the Past Participle Morphology in Italian. In: MIT Working Papers in Linguistics 33, 111-137.

Koehl, A. 2009. "Are French *-ité* suffixed nouns property nouns? ", *in* Montermini, F., G. Boyé & J. Tseng (eds), *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*. Somerville: Cascadilla Proceedings Project; 95-110.

Koehl, A. 2012a. "The second life of French *-itude* suffixation", talk presented to the *15th International Morphology Meeting* (IMM15), Vienna.

Koehl, A. 2012b. "*Altitude, négritude, bravitude* ou la résurgence d'une suffixation", *in Actes du Congrès Mondial de Linguistique Française* (CMLF2012).

Kiparsky, P. 1982, "Lexical Morphology and Phonology", *in* Yang, I. (ed), *Linguistics in the morning calm*, Seoul: Hanshin, 3-91.

Krifka, M, F. Pelletier, G. Carlson, A. ter Meulen, G. Chierchia, and G. Link. 1995. "Introduction", *in* Carlson, G. & G. Pelletier (eds), *The Generic Book*. Chicago: Chicago University Press, 1-24.

Marantz, A. 2001. Words and Things. Handout, MIT.

Martin, F. 2008. *Les prédicats statifs. Etude sémantique et pragmatique.* Bruxelles: Duculot/de Boeck.

Martin, F. 2010. "The Semantics of Eventive Suffixes in French", in Rathert, M. & A. Alexiadou (eds), *The Semantics of Nominalizations across Languages and Frameworks*, Berlin, Mouton de Gruyter, 109-141.

Martin F. 2012. "Stage-level vs individual-level readings of dispositional nouns", in Hathout, N., F. Montermini & J. Tseng (eds), *Selected Proceedings of the 7th Décembrettes*, Munich: Lincom Europa

Moldenhauer, G. 1934. "Suffixstudien". *Z. fr. Spr.*, 58:296-306.

Meunier, A. 1981. *Nominalisations d'adjectifs par verbes supports.* PhD Thesis, Univ. Paris 7.

Rainer, F. 1988. "Towards a theory of blocking: the case of Italian and German quality nouns" *in* van Maarle, J. G. Booij (eds), *Yearbook of Morphology*. Dordrecht: Foris Publications

Rainer, F. 1989. *I nomi di qualità nell'italiano contemporaneo.* Vienna: Braumüller.

Rainer, F. 2005. "Semantic change in word formation". *Linguistics*, 43/2:415–441.

Scalise, S. 1984, *Generative Morphology*, Dordrecht: Foris.

Spitzer, L. 1931. "Warum ersetz frz. *-erie* (dtsch *-erei*) das alte *-ie* (-ei), *Zeitschrift für romanische Philologie*, 51, 70-75.

Tanguy, L., N. Hathout. 2007. *Perl pour les linguistes. Programmes en Perl pour exploiter les données langagières*, Paris/Londres: Hermès Lavoisier.

Uth, M. 2011. *Abstrakte Bedeutung und regelhafte Wortbildung. Französische Ereignisnominalisierungen auf* -age *und* -ment. Berlin/New-York: Mouton de Gruyter.

van Marle, J. 1985. *On the paradigmatic Dimension of Morphological Creativity*, PLS 18, Dordrecht: Foris.

van Marle, J. 1986. " The domain hypothesis: the study of rival morphological processes", *Linguistics*, 24, 601-627.

Appendix

Below are listed all dictionary words (DWs) and neologisms (Ns) taken into account for the study.
-erie DWs: *balourderie, bizarrerie, bouderie, bouffonnerie, brusquerie, chicanerie, cocasserie, connerie, coquetterie, coquinerie,crânerie, crapulerie, crétinerie, dégueulasserie, drôlerie, effronterie, espièglerie, faroucherie, finauderie,flagornerie, flatterie, folâtrerie, fourberie, gaminerie, grognonerie, hautainnerie, ignarerie, imbécilerie*

-erie Ns: *bestialerie, bonhommerie, candiderie, cinglerie, cuculerie, cupiderie, férocerie, fouinerie, fragilerie, génialerie, grand-guignolerie, grandioserie, honnêterie, honterie, ignoblerie, immonderie indignerie, infamerie, innocenterie,insolenterie, ironerie, jovialerie*

-isme DWs: *angélisme arrivisme, artisme, barbarisme, coquettisme, crétinisme, cynisme, despotisme, diabolisme, égoïsme,érotisme, extatisme, frénétisme, hirsutisme, humanisme, idiotisme, immoralisme, infantilisme, intégrisme, ironisme*

-isme Ns: *amicalisme, asocialisme, attentivisme, avarisme, balourdisme, bestialisme, bizarrisme, bonhommisme, bouffonisme,candi(di)sme, colérisme, coquettisme, coquinisme, crapulisme, cuculisme, drôlisme fanfaronisme,farouchisme, faussisme, fébrilisme, férocisme, fidélisme, fragilisme, grand-guignolisme, grognonisme, honnêtisme, hontisme, horribilisme, hostilisme, ignarisme, imbécilisme, immondisme, impatientisme, impudisme,impulsivisme, incongruisme, indignisme, inhumanisme, innocentisme, insolentisme, intelligentisme, jovialisme*

-ité DWs: *acerbité; affabilité, agilité, amabilité, amicalité, anxiété, asocialité, atrocité, avidité, banalité, bestialité, bonté, brutalité, causticité, convivialité, cordialité, coriacité, cruauté, crédulité, crudité, culpabilité, cupidité, curiosité, déloyauté, docilité, dureté, émotivité, étrangeté, excentricité, explosivité, expressivité, exquisité, fausseté, fébrilité, férocité, fidélité, fierté, fragilité, frivolité, grandiosité, honnêteté, hostilité, humanité, ignobilité, immoralité, impériosité, impétuosité, importunité, impulsivité, incapacité, incongruité, incrédulité, indignité, indocilité, inexorabilité, infidélité, inflexibilité, ingéniosité, ingénuité, inhumanité, insensibilité, intégrité, intensité, intrépidité, invincibilité, invulnérabilité, irascibilité, irrationalité, irritabilité, jovialité*

-ité Ns: *angélicité, astuciosité, attentivité, avarité, bizarrité, candidité, capriciosité, costaudité, cocassité, comicité,conscienciosité, coquinité, craintivité, crétinité, déliciosité, désagréabilité, fantasticité, fêtardité, formidabilité,froidité, furiosité, gaminité, génialité, grognonité, hidosité, horribilité, ignarité, immondité, infamité,*

-itude DWs: *certitude, habitude, incertitude, ingratitude, rectitude*
-itude Ns: *abjectitude, acariâtritude, acerbitude, artistitude, asocialitude, atrocitude, banalitude, bestialitude, bétitude, bouffonitude, cocassitude, comicitude coquinitude, coriacitude, cruelitude, cuculitude, cupiditude, faiblitude,fainéantitude, farouchitude, faussitude, férocitude, fidélitude, fiertitude, formidabilitude, fragilitude, froiditude, génialitude, grandiositude, granditude, hautainitude, hilaritude, hirsutitude, horribilitude, humanitude, humilitude, fidiotitude, ignarditude, ignobilitude, imbécilitude, immoralitude, incongruitude, indignitude, infamitude, inhumanitude, intégritude, ironitude, jovialitude*

other Ns: *abominance, admirance, affirmance, hésitance, fascinance, agilance, furibardise, fêtardise, froussardise*
-ance DWs: *ambivalence, bienveillance, clairvoyance, clémence, complaisance, condescendance, confiance, constance, déliquescence, désobligeance, diligence, distance, effervescence, élégance, extravagance, exubérance, impertinence, impuissance, inclémence, incohérence, incompétence, inconscience, inconstance, inconvenance, indécence, indépendance, indifférence, indolence, inélégance, innocence, intelligence, intempérance, intolérance, malveillance, pertinence, puissance*

-esse DWs: *grandesse, humblesse, faiblesse, indélicatesse, politesse, impolitesse, finesse, faiblesse, délicatesse, allégresse, adresse*

-eur DWs: *impudeur, candeur, froideur, fureur, hideur, laideur; impudeur, ferveur, douceur, chaleur, candeur*
-ice/ -ise DWs: *avarice, fainéantise, bêtise, justice, franchise, faiblardise, couardise*
-ie DWs: *ironie, barbarie, idiotie, infâmie, bonhommie, jalousie, folie, félonie, euphorie, discourtoisie, courtoisie*
-ion DWs: *attention, abjection, introversion, indécision, précision, imprécision, indiscrétion, discrétion, dévôtion, correction, confusion, compréhension, circonspection, ambition, affection.*

Simplification, complexification, and microvariation: Towards a quantification of inflectional complexity in closely related varieties

Raffaela Baechler
Albert-Ludwigs-University Freiburg
raffaela.baechler@germanistik.uni-freiburg.de

Guido Seiler
Albert-Ludwigs-University Freiburg
guido.seiler@germanistik.uni-freiburg.de

1. Introduction

In recent typological work the structural complexity of languages has become a centre of interest (cf. e.g. Miestamo *et al.* (eds.) 2008 or Sampson *et al.* (eds.) 2009). This is somewhat surprising given the fact that throughout the 20th century it has been more or less explicitly assumed that overall structural complexity is constant across languages. That is, greater complexity in one area of grammar (e.g. morphology) has been expected to be compensated by a lower degree of complexity in another (e.g. syntax): "[...] impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same" (Hockett 1958: 180). So, whereas the existence of complexity differences between languages has been at least doubted (if not denied) by structural linguistics and linguistic typology, another line of research, variationist linguistics, has talked about complexity differences quite unscrupulously from its very beginnings. In his seminal paper on diglossia Ferguson says, with regard to structural differences between High and Low varieties, that "[o]ne of the most striking differences between H[igh] and L[ow] variety [...] is in the grammatical structure: H has grammatical categories not present in L and has an inflectional system of nouns and verbs which is much reduced or totally absent in L" (Ferguson 1959: 241). Statements about the reduced, simplified structural characteristics of vernacular dialects as opposed to codified standard languages are abundant in the dialectological literature. However, to our knowledge (most of) these statements are purely intuitive, for they have never been based on solid measurements of complexity.

In a more recent line of research at the intersection between linguistic typology and sociolinguistics attempts are being made to (i) uncover complexity differences between languages / varieties and (ii) to explain those differences by reference to the structure of the community where the language/variety is spoken. In particular, it is claimed that languages spoken by small, close-knit, isolated communities display a greater degree of structural complexity (Trudgill 2004, 2009, 2011, Nichols 1992, Braunmüller 1984, 2003). We will call this idea 'Isolation Hypothesis' (IH). If the IH is correct, it predicts something not only about large-scale typological comparison but also about sets of genetically closely related and similar languages or varieties: In isolated varieties lacking contact processes of simplification are less likely to occur than in non-isolated cognate varieties. Similarly, isolated varieties are more likely to display complexification than others.

Trudgill (2011) proposes three possible extralinguistic scenarios with different effects on simplification or complexification, respectively. First, traditional, remote dialects with no L2 learners are an ideal biotope for those types of complexification which cannot be attributed to structural borrowing: "[…] spontaneous (as opposed to additive) complexification will develop on a large scale mainly in low-contact communities, where 'low contact' refers in particular to an absence of a history of large-scale acquisition by non-native adult speakers" (Trudgill 2011: 89). Also, archaic features

seem to be more stable in isolated languages (Trudgill 2011: 13). Second, languages / varieties which are (or have been) acquired by many adult non-native speakers are expected to display simplification processes such as regularization of irregularities, increase in morphological transparency, reduction in syntagmatic redundancy, or loss of morphological categories (Trudgill 2011: 34, 40, 62). The third type is also due to language contact, but of a different kind. In contact-induced change grammatical features may spread from one language into another, which may lead to the addition of new features and thus to greater complexity of the influenced language (Trudgill 2011: 27). This kind of contact-related complexification "is most likely to occur in long-term co-territorial contact situations involving child bilingualism" Trudgill (2011: 34).

In this preliminary study we attempt to put to test the IH, using evidence from different varieties of German. We believe that a set of cognate varieties provides a marvellous piece of evidence since we can observe the results of diachronic processes of simplification and complexification in a very direct way, due to the close genetic affiliation and thus the common historical origin of the varieties. In order to test the predictions of the IH in a substantial way, complexity must be operationalised. This is why the paper focuses on the complexity of noun inflection only. We are not yet able at this point to make any substantial claim about the overall complexity of the grammars of our varieties. However, our preliminary findings on noun inflection give us at least a hint whether the IH is worth to be pursued any further (we will argue that this is indeed the case). To put it differently: If our findings even within a limited, relatively cross-linguistically easily comparable area such as noun inflection were totally incompatible with the IH already, it seems very unlikely to us that including other areas of grammar would lead to a less disparate picture.

The paper is structured as follows. We will first address the research questions (section 2). Section 3 presents the sample of the varieties studied (3.1), a definition of absolute complexity (3.2), some previous approaches to complexity and microvariation (3.3), and our proposed procedure to measure complexity of noun inflection in closely related, similar varieties (3.4). The results of our investigation are presented in section 4. In section 5 we will discuss the results in the light of our research questions, and we will give a short outlook.

2. Research questions and hypotheses

Question 1: Is there an overall diachronic tendency?
In the light of the references mentioned above, the expectations are unclear. There seems to be a certain consensus that, all other things being equal, languages tend to gradually simplify their grammars, in particular their morphologies: If isolated languages / varieties (i) show a slower rate of change (Trudgill 2011: 2-8) and (ii) a greater degree of complexity, one might easily conclude that this greater complexity is an archaic trait which just survives longer if the language changes at a slow rate. This view presupposes, of course, the idea that the 'normal thing' for a language is to simplify across time. Perhaps this intuition is particularly influenced by linguists' familiarity with the older Indo-European languages and their intricate inflectional systems. Thus, if there is an overall diachronic tendency at all we might hypothesize this tendency leads towards simplification.

Question 2: What are the effects of isolation?
With regard to question 2, our expectations are much clearer: If the IH is correct, we expect a greater degree of complexity in isolated varieties.

Question 3: What are the effects of contact?

As outlined in section 1, contact situations can lead to both complexification and simplification. Complexification is expected in pre-threshold bilingualism, i.e., in situatations of stable contact where both languages are acquired early. Simplification in post-threshold bilingualism, i.e. in situations where the language in question is acquired by adult non-native speakers (Trudgill 2009:101). As will be shown in Chapter 3.1, we are concerned with pre-threshold bilingualism in the case of Issime German and therefore we expect complexification rather than simplification here.

Question 4: Are there instances of complexification?
Complexification seems to be uncommon in larger, non-isolated languages. Genuine ("spontaneous", Trudgill 2011: 89) complexification, i.e., complexification which is not due to structural borrowing, is expected to occur only in isolated dialects.

Question 5: What is the role of codification?
Here the expectations are unclear. If we take seriously Ferguson's quote from section 1, High varieties are notorious in their greater structural complexity if compared with spoken vernaculars. Also, there might be conserving effects of codification. It therefore seems plausible to assume that codified standard varieties display a greater degree of complexity than spoken dialects.

3. Method

3.1. Sample

To answer these questions we selected five German varieties. Old High German (OHG) is the oldest attested German variety and New High German (NHG) the present-day standard language. The non-standard varieties are the Alemannic dialect of the Kaiserstuhl, an area near Freiburg in the South-West of Germany, the Alemannic dialect of Visperterminen in the Canton of Valais in Switzerland and the Alemannic dialect of Issime, a linguistic island in the Aosta Valley in Italy. The data are based on the following grammatical descriptions: Braune/Reiffenstein (2004) for OHG, Eisenberg (2006) for NHG, Noth (2003) for Kaiserstuhl Alemannic, Wipf (1911) for Visperterminen Alemannic, Zürrer (1999) for Issime Alemannic. Unfortunately there aren't any more recent exhaustive grammars for the dialects in the Valais. For the analysis, this has to be kept in mind.

We consider OHG, NHG and Kaiserstuhl Alemannic as not isolated, the two Walser dialects of Visperterminen and Issime as isolated. There are a few qualitative criteria for considering the Walser dialects as isolated which are introduced in the following.

Visperterminen is situated in the canton of Valais in Switzerland and has 1373 inhabitants (2010; wikipedia). It is located at 1378m above sea level and at the dead end of its only road access from Visp.

Issime is one of several Alemannic colonies in northern Italy. In the 13th century people migrated from the canton of Valais to the Aosta Valley in Italy. As for Issime, there hasn't been any contact with the German speaking language area since then. Many of the 400 inhabitants in Issime are quinquelingual: Alemannic, Franco-Provençal, Piemontese, Standard French (which is the official language of the région autonome Vallée d'Aoste) and Standard Italian (the official language of Italy). However, they do not speak Standard German.

Although this sample is small, it contains some interesting contrasts: historical (OHG) vs. recent, codified (NHG) vs. vernacular, isolated (Visperterminen, Issime) vs. non-isolated, contact (Issime) vs. monolingual environment.

3.2. Absolute complexity

In the literature a difference is made between relative and absolute complexity. In relative complexity one is interested in whether a linguistic phenomenon is complex to a speaker, a hearer, an L1 acquirer, an L2 learner, etc, i.e. "how difficult a phenomenon is to process (encode/decode) or learn" (Miestamo 2008: 25).

In absolute complexity one considers only the language system itself. Following Miestamo "the [absolute] complexity of a linguistic phenomenon may be measured in terms of the length of the description of that phenomenon [...] A less complex phenomenon can be compressed to a shorter description without losing information" (Miestamo 2008: 24). We can adapt this to the language system and assume that the longer the description of the language system is (the less it can be compressed), the more complex the language system will be.

Another important point is that we consider here only inflectional complexity, more precisely the inflectional complexity of nouns, which does of course not mean that phonological or syntactic complexity should be excluded. Rather, they must be included if one wants to calculate the overall complexity of the entire language system. Other nominal and verbal parts of speech will be measured in a later stage of our project.

3.3. Previous approaches to microvariation and complexity

In this chapter we will briefly discuss some central proposals for measuring complexity (especially in closely related varieties) and show why they are not appropriate for our purposes.

There are large-scale typological comparisons (Shosted 2006, McWorther 2001, Nichols et al. 2006) whose common ground is that they count the number of grammatically encoded features. This is clearly operationalising, but too coarse for the purposes of micro-comparison of closely related languages and varieties. Dammel & Kürschner (2008) compare the noun plural allomorphy in ten Germanic languages. They incorporate ideas of Natural Morphology (Wurzel 1984) such as uniformity and iconicity to account for aspects of relative complexity. Relevant factors are e.g. the number of plural allomorphs, stem involvement, multiple exponence, zero marking and fusion. As the weights of the factors are assigned merely intuitively, an operationalisation does not seem to be possible. Szmrecsanyi & Kortmann (2009) compare 42 varieties of English: traditional L1, high-contact L1, L2 and creoles. They analyse 31 features which are a selection of the 76 features covered by the *World Atlas of Morphosyntactic Variation in English*. Their method is clearly operationalising. However, the features are very English-specific and themselves treated in a binary way (presence or absence of the feature). Therefore it is of limited use for microvariation especially in highly inflecting languages.

Since there has not been any appropriate tool to measure complexity in inflecting and closely related varieties, we have tried to develop a simple method adapted to our sample, which will be presented in the following chapter.

3.4. Measuring inflectional complexity

In this subsection we propose a simple procedure to uncover complexity differences in inflectional systems even of genetically closely related, similar languages / varieties. We will first outline the concrete steps we have undertaken when analysing our sample in a cookbook-like fashion, before we briefly address some of the insights which naturally follow from the proposed procedure.

The main goal of our procedure is to make visible the raw data structure in the first place. We deliberately do that as much as possible in a pre-theoretic way. The immediate results of the procedure should be analysable in theoretical contexts of different flavours (we come back to the issue in the concluding section 5). Originally, we believed that structuring the data in such a rather mechanical way is a relatively easy task. However, it turned out that even our toolkit-style procedure requires a considerable amount of hand-made morphological analysis, for many decisions can be made only if the functioning of the respective inflectional systems is linguistically well understood.

We use the following method in four steps in order to measure inflectional complexity:

- Step 1: Collect the distinguishable inflectional paradigms of the respective language/variety.
- Step 2: Break each paradigm down into a list of inflectional markers.
- Step 3: Put the markers on a list and remove repeated occurrences of markers. Count the remaining markers.
- Step 4: Multiply the number of markers by the number of marker combinations (=inflectional classes).

We thus define complexity as the number of inflectional markers multiplied by the number of inflectional classes.

Step 1: Every grammatical description forms the paradigms in a different way, even if we are concerned with the same variety. For example, with regard to NHG, the Duden-Grammatik (1998: 223-224) distinguishes ten inflection types (Deklinationstypen), but Eisenberg (2006: 152-154) only four types with two subtypes. However, since we aim to compare the paradigms of different varieties we need comparable paradigms, i.e. paradigms which are identified in similar ways. Our paradigms are not organised in inflection types but in inflectional classes. Furthermore each paradigm must be maximally compressed to obtain the shortest description of the noun inflection. We are then able to compare the shortest description of variety A with the shortest description of variety B.

Step 2: We define a marker as a distinct pairing of exponent and grammatical feature. For example the paradigm of *Tag* consists of three markers (for the full paradigm see table 2):

m_1: -es $\begin{pmatrix} \text{NUM} & \text{SG} \\ \text{CASE} & \text{GEN} \end{pmatrix}$

m_2: -e $\begin{pmatrix} \text{NUM} & \text{PL} \end{pmatrix}$

m_3: -n $\begin{pmatrix} \text{NUM} & \text{PL} \\ \text{CASE} & \text{DAT} \end{pmatrix}$

For convenience we write the markers without attributes as follows: -es:sg.gen, -e:pl, -n:pl.dat. In cases of multiple exponence, each exponent is counted as a marker. Thus, *Hand-Hände* is made of umlaut and the suffix –e. Umlaut demonstrates that not only segmentable morphs but also (not phonologically conditioned) stem alternations can be a marker. We note them as a rewriting rule: V → [+front, -low] / [NUM PL]. Again for convenience, we write the marker as UL:pl.

Table 1 displays the paradigm of *Student* which has homophonous markers –n. They may occur whenever they cannot be assigned to a uniform function. Thus, –n in the paradigm of *Student* has four distinct functions: –n:acc.sg, –n:dat.sg, –n:gen.sg, –n:pl. Concerning syncretism, we distinguish between "good" and "bad" syncretism. For example, the paradigm of *Student* (Table 1) has the following markers in the plural: -n:nom.pl, –n:acc.pl, –n:dat.pl, –n:gen.pl. However, this –n can be attributed to consistent

function, namely plural. Therefore the plural has only one and not four markers. This type of "good" syncretism does not add to complexity. The singular of the paradigm of *Student* contains the following markers: –n:acc.sg, –n:dat.sg, –n:gen.sg. Since it is impossible to assign a consistent function to this –n (the nominative singular is not marked), each of these three suffixes has to be counted as a separate marker, so the paradigm has three markers in the singular. This syncretism adds to complexity and we therefore call it "bad" syncretism.

Table 1: Paradigm of *Student*

	SG	PL
nom	Student	Studenten
acc	Studenten	Studenten
dat	Studenten	Studenten
gen	Studenten	Studenten

In step 3 the markers are put on a list and the repeated occurrences of markers are being removed. This is a very important step because varieties of German notoriously re-use the same markers across different paradigms. For instance, if the dative plural is marked in NHG, the marker –n is suffixed across all inflectional classes (cf. Table 11).

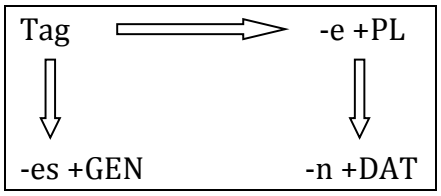
Step 4: Inflectional complexity is calculated by multiplying the number of markers by the number of inflectional classes. The inflectional class can be defined as a specific combination of markers. Therefore, both larger marker inventory and large numbers of inflectional classes add to complexity, but they do not automatically follow one from another. We multiply the number of markers by the number of inflectional classes because our intuition is that each marker combination is to be counted as one way of making use of the same marker inventory. For instance, if there are five inflectional classes, the morphology uses the marker inventory five times to create different paradigms.

In our method we assume Underspecification and the Elsewhere Condition (Anderson 1992, Kiparsky 1973; for German morphology cf. Eisenberg 2006, Thieroff&Vogel 2009). Traditionally paradigms of German nouns are represented by means of eight instructions, whereby each instruction contains a full specification of feature content and associated exponent. The paradigm of *Tag* (Table 2) contains the following eight instructions: nom.sg→Tag, acc.sg→Tag, dat.sg→Tag, gen.sg→Tages, nom.pl→Tage, acc.pl→Tage, dat.pl→Tagen, gen.pl→Tage. Assuming Underspecification, the paradigm of *Tag* contains only three instructions (cf. Table 3): Add –es in the genitive singular, -e in the plural and –n in the dative plural. However, how does the case-underspecified form *Tag* know that it may not be used as genitive? Why does the grammar not generate **wegen des Tag*, rather *wegen des Tages* (the preposition *wegen* governs a genitive). Here the Elsewhere Condition comes into play: If there is a more specific instruction you must not follow a less specific one. For example, if a genitive singular is required, the most specific available form must be used in the first place. Since *Tages* is more specific for genitive singular than *Tag*, *Tages* will be used first and blocks the insertion of *Tag* for the genitive singular.

Table 2: Paradigm of *Tag*

	SG	PL
nom	Tag	Tag-e
acc	Tag	Tag-e
dat	Tag	Tag-en
gen	Tag-es	Tag-e

Table 3: Paradigm of *Tag* assuming unterspecification



When the method as outlined above is applied to an inflectional system, it naturally falls out without any further assumptions or stipulations that the following factors add to the amount of inflectional complexity:

- Number of inflectionally distinguished grammatical features, e.g. the number of cases.
- Allomorphy created by a number of inflectional classes, e.g. the plural allomorphs (–e, –n, –er, etc.) in NHG.
- Multiple exponence, e.g. in *Wald-Wälder* the plural is expressed by the umlaut and the suffix –er.
- "bad" syncretism, e.g. the homophonous singular markers of the Paradigm *Student* (Table 1).

The following factors do not add to complexity:

- Re-use of markers across inflectional classes, e.g. the suffix –n (dative plural) in NHG.
- Absence of otherwise attested distinctions in particular inflectional or lexical classes, e.g. Kaiserstuhl Alemannic nouns do not distinguish cases, but determiners and pronouns do.
- Allomorphy which is predictable on phonological grounds.

4. Results

As already mentioned in section 3.4 above, we are faced with a great deal of decisions when analysing the inflectional systems of our varieties even if such a cookbook-like method is applied. In 4.1, we will briefly discuss a few of the analytical difficulties we encountered. We do that in a very exemplary way by choosing one or two typical problems for each variety in order to illustrate the reasoning which is behind the categorisations we have made. The complete paradigms of each variety are listed in the appendix. Subsection 4.2 presents the results of our investigation.

4.1. Paradigms

4.1.1. OHG

Traditionally so-called a-stems and wa-stems are analysed as two different inflectional classes of OHG. However, their sets of endings are identical. The difference between a-stems and wa-stems is the stem alternation in the wa-stems. For instance, the dative of *tag* (a-stem) is *tag-e* whereas the dative of *hleo* (wa-stem) is *hlew-e*. To form the dative of both a– and wa–stems, the ending –e is suffixed. The difference between the two paradigms is that *hleo-hlewe* does show a stem alternation, but *tag-tage* does not. However, as we cannot attribute any uniform meaning to this alternation, we consider these two stems (*hleo-*, *hlew-*) as stem allomorphs and therefore end up with only one inflectional class. What is now the shortest description of this alternation? The linguistic

generalisation is that there is a stem allomorph for unsuffixed forms (*hleo*) and another stem allomorph for affixed forms (*hlew*-). We note this contextual conditioning as a rewriting rule, which is counted as an additional marker: …eo→ew/_suffix, …o→aw/_suffix. In stems which do not end in ...eo/...ew- the rule simply runs vacuously, i.e., it is not applicable.

Table 4: a-stems and wa-stems in OHG

		SG					PL			
	IC[12]	nom	acc	dat	gen	instr	nom	acc	dat	gen
a-stem	1	tag	tag	tag-e	tag-es	tag-o	tag-a	tag-a	tag-on	tag-o
wa-stem		hleo	hleo	hlew-e	hlew-es		hlew-a	hlew-a	hlew-on	hlew-o
wa-stem		horo	horo	horaw-e	horaw-es		horo	horo	horaw-on	horaw-o
a-stem	8	wort	wort	wort-e	wort-es	wort-o	wort	wort	wort-on	wort-o

4.1.2. NHG

We did not take into account the –en/–n variation in the dative plural (e.g. *Staat-en*, *Wäld-er-n*). We assume that this variation is purely phonologically conditioned. More precisely, there is a preference for words to end in a trochee. In *Staaten* the ending is therefore syllabic, but not in *Wäldern*.

We ignored also the –es/–s variation in the genitive singular (e.g. *Gast-es*, *Schaden-s*) because the use of –es and –s depends on the final sound, the stress and the number of syllable the word has (Eisenberg et al. 1998: 224-225).

In inflectional class 7 (sg. *Wald* - pl. *Wälder*), the plural is formed by –er and umlaut. This inflectional class includes also words like *Bild-Bilder* without an umlautable vowel. However, words like *Bild* do not form their own inflectional class because words which form the plural with –er always umlaut the stem vowel if possible.

As in OHG, there is a stem alternation in the inflectional class 10 (*Blume-Blumen* and *Pizza-Pizzen*) (table 5). The endings of *Blume* and *Pizza* are identical. The only difference is the stem alternation of *Pizza* in the plural. For this contextual conditioning we have a rewriting rule which says: stem-final vowel is deleted in plural environment (… V#→ -ø[NUM PL]). This rule is like in OHG counted as a marker. For the same reasons *Konto-Konten* does not have its own inflectional class but makes part of the inflectional class 9.

Table 5: stem alternation in NHG

	SG				PL			
IC	nom	acc	dat	gen	nom	acc	dat	gen
10	blume	blume	blume	blume	blume-n	blume-n	blume-n	blume-n
	pizza	pizza	pizza	pizza	pizz-en	pizz-en	pizz-en	pizz-en
9	staat	staat	staat	staat-es	staat-en	staat-en	staat-en	staat-en
	konto	konto	konto	konto-s	kont-en	kont-en	kont-en	kont-en

4.1.3. Kaiserstuhl Alemannic

In Kaiserstuhl Alemannic we consider the –n– in *schdainer* ('stones') as purely phonological, because it is introduced only if otherwise suffixation would create a hiatus (table 6). We observe similar patterns also in other contexts. For example in *wu-n-er*: *wu* means 'as' or 'when', *er* means 'he'and *n* is a glide.

In inflectional class 3 the plural is formed by adding the suffix –er, the plural of *Wald* additionally by umlaut (table 6). However, for the same reasons as in Standard

12 IC=inflectional class

German, we have only one inflectional class for the plural on –er with or without an umlaut, i.e., if the plural is formed with –er, the stem vowel always takes an umlaut if it is possible.

Table 6: inflectional class 3 in Kaiserstuhl Alemannic

	SG			PL		
IC	nom	acc	dat	nom	acc	dat
3	schdai	schdai	schdai	schdai-n-er	schdai-n-er	schdai-n-er
	wald	wald	wald	wäld-er	wäld-er	wäld-er

4.1.4. Visperterminen Alemannic

As opposed to NHG and Kaiserstuhl Alemannic, we need two inflectional classes for the plurals ending on –er in Visperterminen Alemannic (IC 10 and 11) (table 7) because there are some words with an umlautable vowel and –er in the plural which do not umlaut the vowel (e.g. *lamm-lammer*) whereas others do (e.g. *chrut-chriter*).

We consider the -n- in *redlini* (IC 12) not as a plural marker but as phonologically conditioned for the same reasons as in the dialect of the Kaiserstuhl, i.e. to prevent a hiatus.

Table 7: plural on -er and the glide -n- in Visperterminen Alemannic

	SG				PL			
IC	nom	acc	dat	gen	nom	acc	dat	gen
10	chrut	chrut	chrut	chrut-sch	chrit-er	chrit-er	chrit-er-u	chrit-er-o
11	lamm	lamm	lamm	lamm-sch	lamm-er	lamm-er	lamm-er-u	lamm-er-o
12	redli	redli	redli	redli-sch	redli-n-i	redli-n-i	redli-n-u	redli-n-o

4.1.5. Issime Alemannic

Concerning the plural of the inflectional class 10 (table 8) we must first define the morphemes. We think it is uncontroversial that we can segment –i and –u. –I is the marker for nominative and accusative plural and –u the marker for dative and genitive plural. But how to deal with the –n– between the stem and the case endings? If we compare the paradigm of *berri* (IC 10) with the paradigm of *bet* (IC 9) we see that the endings are identical and the only difference between these two inflectional classes is this –n–. Therefore, we could have considered the –n– as phonologically conditioned, to prevent a hiatus as has been demonstrated for Kaiserstuhl and Visperterminen Alemannic.

However, a closer look at the data reveals that this -n- is not purely phonological. With *sia-siawa* (IC 8) we have a similar case. The endings are the same as in the inflectional class 1 (*weg-wega*) and –w– could be a glide. We would thus have two glides, –n– and –w–. However, the choice of –n– and –w– is unpredictable on purely phonological grounds. Therefore, we analyse –n– and –w– as two distinct plural markers.

We find further evidence for –n– as a plural marker in the paradigm of *uave* (IC 2). Here, the plural also shows an –n– between the stem and the case endings but we cannot find any phonological explanation: -n-insertion does not prevent a hiatus anyway.

To sum up, the –n– and –w– in the paradigm of Issime are plural markers.

Table 8: -n- and -w- as plural marker in Issime Alemannic

	SG				PL			
IC	nom	acc	dat	gen	nom	acc	dat	gen
10	berri	berri	berri	berri-sch	berri-n-i	berri-n-i	berri-n-u	berri-n-u
9	bet	bet	bet	bet-sch	bet-i	bet-i	bet-u	bet-u

8	sia	sia	sia	sia-sch	sia-w-a	sia-w-a	sia-w-e	sia-w-u
1	weg	weg	weg	weg-sch	weg-a	weg-a	weg-e	weg-u
2	uav-e	uav-e	uav-e	uav-endsch	uav-n-a	uav-n-a	uav-n-e	uav-n-u

4.2. Inflectional complexity of nouns

In this section we present and discuss the main results of our investigation, i.e. the complexity of noun inflection of our five varieties, which we calculated by multiplying the number of markers by the number of inflectional classes (figure 1). Subsequently we will compare the number of markers with the number of inflectional classes (figure 2). Table 9 shows the number of markers and inflectional classes as well as the complexity of noun inflection.

Table 9: markers-inflectional classes-complexity

varieties	markers	inflectional classes	complexity (markers * inflectional classes)
OHG	40	18	720
Issime	26	19	494
Visperterminen	24	18	432
NHG	11	14	154
Kaiserstuhl	7	7	49

First of all, we see in figure 1 that the five varieties are not equally complex, despite their close genetic affiliation. We can form three groups: The most complex is OHG, a second group with Issime and Visperterminen Alemannic, and a third group with NHG and Kaiserstuhl Alemannic.

Since OHG is the most complex variety, we observe an overall diachronic simplification tendency. Of course figure 1 is perhaps somewhat suggestive because we arranged the varieties from the most to the least complex. However, Issime, Visperterminen, Kaiserstuhl Alemannic and NHG are all present-day varieties.

To answer the question of whether codification leads to complexification or simplification, we compare NHG (codified) with the non-standard varieties. Between the Walser dialects (Issime, Visperterminen) and NHG there is a steep decrease in complexity. Thus, noun inflection in Issime and Visperterminen Alemannic is much more complex than the inflection in NHG. In contrast, Kaiserstuhl Alemannic is less complex than NHG. However, compared with the Walser dialects the decrease in complexity between NHG and Kaiserstuhl Alemannic is moderate. As NHG is neither more complex nor simpler than all the non-standard varieties (but between theses varieties) we can conclude that codification does not play a major role with regard to complexity.

We will now turn our attention to the nonstandard varieties and especially to the Walser dialects. Figure 1 displays a steep decrease in complexity between the Walser dialects (isolated) and the Kaiserstuhl Alemannic (non-isolated). This is in accordance with the IH: that isolated varieties are more complex than non-isolated varieties.

Between Issime and Visperterminen we can observe a moderate decrease in complexity. This is perhaps due to the double isolation of Issime or to language contact (which is in this case with Italian and French). First, Issime is not only topographically isolated but also linguistically (it does not make up part of the West-Germanic dialect continuum). Therefore, if it is correct that the more a language is isolated the more it is complex and if we consider Issime as doubly isolated, Issime's greater complexity is expected. A second possible explanation is that language contact has a complexifying effect, but (as discussed in sections 1-2) only in "long-term child contact situations" (Trudgill 2011: 120), which is indeed the case in Issime. However, this complexification

is a particular type of complexification called by Trudgill (2011) "additive complexification", i.e. morphological categories are borrowed from the contact language/s. As the noun inflection in French and Italian (and the respective dialects spoken in the Aosta Valley) is less complex than the one in Issime Alemannic, we would expect simplification rather than complexification as a result of contact. Therefore the higher complexity in Issime Alemannic is presumably due to the absence of contact with the West-Germanic dialect continuum and supports the IH.

Figure 1: complexity of noun inflection

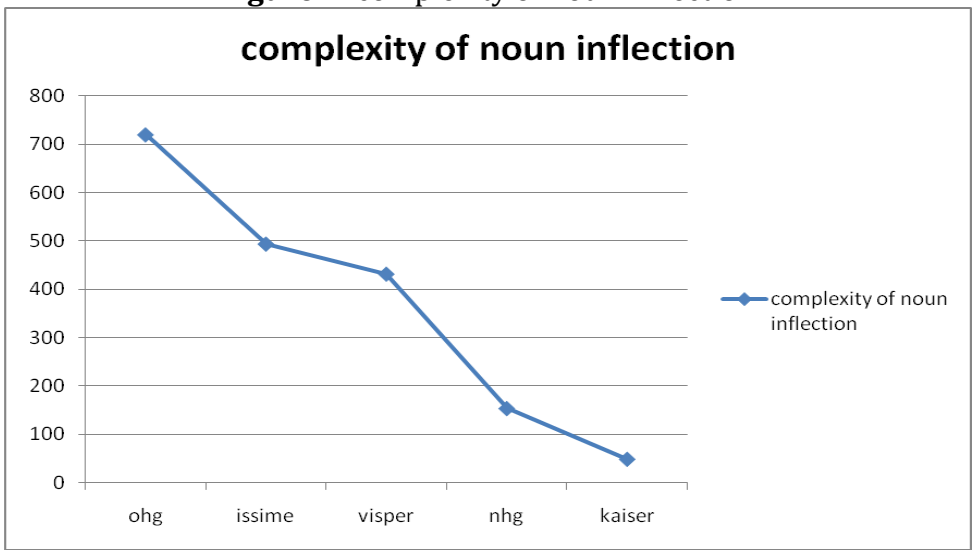
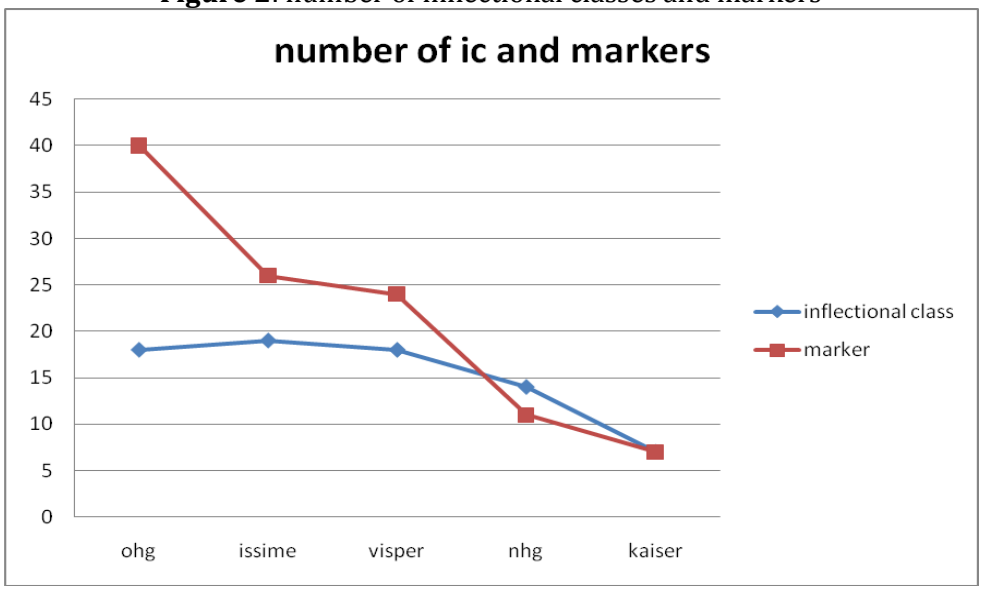


Figure 2 displays the number of inflectional classes and the number of markers. Compared to the overall complexity of noun inflection, the number of markers shows the same order: The variety with most markers (40) is OHG, the second group is constituted by Issime Alemannic (26 markers) and Visperterminen Alemannic (24 markers) and the third group by NHG (11 markers) and Kaiserstuhl Alemannic (7 markers).

The inflectional classes display a different pattern. They are relatively stable in the first three varieties: Issime Alemannic has 19 inflectional classes; OHG and Visperterminen Alemannic, 18. By contrast, we can observe a clear decrease in NHG (14 inflectional classes) and Kaiserstuhl Alemannic (7 inflectional classes). Issime Alemannic provides a very interesting case. Concerning the total complexity of noun inflection (markers*inflectional classes) and the number of markers, all the present-day varieties are less complex than OHG, which corresponds to an overall diachronic simplification tendency. However, Issime Alemannic has one inflectional class more than OHG, which we interpret as an instance of complexification. In the research questions (section 2) it was argued that instances of complexification could occur only in isolated dialects. This result is in accordance with the IH.

Figure 2: number of inflectional classes and markers



5. Discussion

Let us now discuss the findings in the light of the research questions and expected answers from section 2, repeated here:

Question 1: Is there an overall diachronic tendency? (Expected: simplification)
We have indeed found a general simplification tendency from OHG to all more recent varieties. The only exception is the number of inflectional classes in Issime which is greater than in OHG.

Question 2: What are the effects of isolation? (Expected: greater complexity)
With regard to question 2, our results are almost shockingly clear. The inflectional systems of our isolated varieties, Visperterminen and Issime, are clearly more complex than those of other recent varieties. Since their inflectional complexity is much closer to OHG than to the other varieties one might interpret this state of affairs as an instance of conservatism.

Question 3: What are the effects of contact? (Expected: Complexification is expected in pre-threshold bilingualism)
The high-contact dialect of Issime is more complex than the dialects without contact. However, it is not clear at this point whether this is due to the contact situation (Issime speakers are multilingual from childhood) or due to Issime's geographical and linguistic isolation from the West-Germanic dialect continuum since both factors are expected to have similar effects. Since the nominal inflection systems of the Romance contact varieties is much simpler we favour the second explanation.

Question 4: Are there instances of complexification? (Expected: only in isolated dialects)
There is one clear instance of complexification, namely the increase of the number of inflectional classes from OHG to the Issime dialect, which matches the expectations.

Question 5: What is the role of codification? (Expectation: greater complexity of codified varieties)
The only codified standard variety, NHG, displays a rather low degree of complexity, but it is more complex than the Kaiserstuhl dialect. On the basis of our data we can conclude

that codification is not a predictive factor with regard to inflectional complexity. Its possible relevance is outranked by other factors such as isolation. However, if the IH is applied consistently, NHG is expected to have the lowest degree of complexity since NHG is the variety with the farthest reach and the greatest number of speakers. Under these assumptions it is unexpected that the dialect of Kaiserstuhl (with much fewer speakers) is less complex. The greater complexity of NHG as compared to Kaiserstuhl must therefore be due to some other factor, and this factor might be codification.[13]

This paper may serve as basis for further analysis on more varieties and parts of speech. Since our sample is still relatively small, future research will include more (Alemannic) varieties to obtain more comparable results. Furthermore, our ultimate goal is to measure overall inflectional complexity. To do this, we will extend the analysis to the paradigms of other parts of speech, e.g. determiners, pronouns, adjectives, verbs. Since there is no obvious counterpart of inflectional classes in the inflectional systems of determiners, pronouns and adjectives, it will possibly be necessary to reconsider the influence of the number of inflectional classes on overall complexity.

It seems to us that our preliminary study feeds well into very recent approaches to morphological theory where principal parts play a crucial role. Principal parts could be used to measure the inflectional complexity especially of nouns and verbs where we are faced with a considerable number of inflectional classes. Principal parts are those morphosyntactic properties and their exponents which are necessary to predict the other cells of a paradigm. Finkel&Stump (2007) distinguish three kinds of principal parts: static, adaptive and dynamic. For illustration, table 10 shows a hypothetical conjugation system. The system contains seven conjugation classes (I-VI) and four morphosyntactic properties (W-Z). The different inflectional exponents are represented by a-o and the dynamic principal parts shaded:

Table 10: Dynamic principal parts (Finkel & Stump 2007: 44)

	morphosyntactic property			
conjugation	W	X	Y	Z
I	a	e	i	m
II	b	e	i	m
III	c	f	j	n
IV	c	g	j	n
V	d	h	k	o
VI	d	h	l	o

In a static system of principal parts the morphosyntactic property set which identifies the principal parts is the same for every conjugation class. For instance, for the paradigms in table 10 the static principal parts are the morphosyntactic properties W, X, Y and their exponents. By contrast, the dynamic principal parts "are neither linearly ordered nor necessarily parallel from one conjugation to another" (Finkel/Stump 2007: 44). If a lexeme has the exponent c for the morphosyntactic property W, we do not know to which conjugation class the lexeme belongs. However, if this lexeme shows the exponent f for the morphosyntactic property X, we can deduce that it belongs to conjugation class III. Therefore in this paradigm we need to know only one dynamic principal part for each conjugation. In a static conception we needed three principal parts. Finkel and Stump

[13] Many thanks to Helen Christen, Fribourg, for making this point.

summarise that a dynamic scheme "allows us to assume a much smaller inventory of principal parts than is possible under the static or adaptive conception" (Finkel/Stump 2007: 44). Underspecification and absolute complexity can be nicely implemented in this dynamic conception. If we assume underspecification, we have to specify only a minimum of forms and the rest can be underspecified. To measure the absolute complexity we assume that the longer the description of the language system is, the more complex the language system will be (cf. Miestamo 2008 and section 3.2). A system of dynamic principal parts shows the minimum of principal parts necessary to deduce all the other forms of the paradigm. Therefore the language system is maximally compressed, which is a necessary prerequisite for the comparison of different languages' complexities.

References

Anderson, Stephen R. 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Braune, Wilhelm; Reiffenstein, Ingo 2004. *Althochdeutsche Grammatik I. Laut- und Formenlehre*. Tübingen: Niemeyer.

Braunmüller, Kurt 1984. Morphologische Undurchsichtigkeit – ein Charakteristikum kleiner Sprachen. *Kopenhagener Beiträge zur Germanistischen Linguistik* 22: p. 48–68.

Braunmüller, Kurt 2003. Language, typology and society: possible correlations. In: Sherzer, Joel; Stolz, Thomas (ed.) *Minor Languages. Approaches, Definitions, Controversies*, p. 89–101. Bochum: Brockmeyer.

Dammel, Antje; Kürschner, Sebastian 2008. Complexity in nominal plural allomorphy. A contrastive survey of ten Germanic languages. In: Miestamo, Matti; Sinnemäki, Kaius; Karlsson, Fred (ed.) *Language Complexity. Typology, contact, change*, p. 243–262. Amsterdam: Benjamins.

Eisenberg, Peter 2006. *Grundriß der deutschen Grammatik. Das Wort*. Weimar: Metzler.

Eisenberg, Peter; Gelhaus, Hermann; Henne, Helmut; Sitta, Horst; Wellmann, Hans (ed.) 1998. *Duden. Grammatik der deutschen Gegenwartssprache*. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Finkel, Raphael; Stump, Gregory 2007. Principal parts and morphological typology. *Morphology* 17: p. 39-75.

Ferguson, Charles 1959. Diglossia. *Word* 15: p. 325–340.

Hockett, Charles 1958. *A course in modern linguistics*. New York: Macmillan.

Kiparsky, Paul 1973. 'Elsewhere' in phonology. In: Anderson, Stephen R.; Kiparsky, Paul (eds.) *A Festschrift for Morris Halle*, p. 93-106. New York: Holt.

McWhorter, John 2001. The world´s simplest grammars are creole grammars. *Linguistic Typology* 5: p. 125–166.

Miestamo, Matti 2008. Grammatical complexity in a cross-linguistic perspective. In: Miestamo, Matti; Sinnemäki, Kaius; Karlsson, Fred (ed.) *Language Complexity. Typology, contact, change*, p. 23–41. Amsterdam: Benjamins.

Miestamo, Matti; Sinnemäki, Kaius; Karlsson, Fred (Hg.) 2008. *Language Complexity. Typology, contact, change*. Amsterdam: Benjamins.

Nichols, Johanna 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nichols, Johanna; Barnes, Jonathan; Peterson, David A. 2006. The robust bell curve of morphological complexity. *Linguistic Typology* 10: p. 98–108.

Noth, Harald 1993. *Alemannisches Dialekthandbuch vom Kaiserstuhl und seiner Umgebung*. Freiburg i.Br.: Schillinger.

Sampson, Geoffrey; Gil, David; Trudgill, Peter (ed.) 2009. *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Shosted, Ryan K. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10: p. 1–40.

Szmrecsanyi, Benedikt; Kortmann, Bernd 2009. Between simplification and complexification: non-standard varieties of English around the world. In: Sampson, Geoffrey, Gil, David; Trudgill, Peter (ed.) *Language Complexity as an Evolving Variable*, p. 64–79. Oxford: Oxford University Press.

Thieroff, Rolf; Vogel, Petra M. 2009. *Flexion*. Heidelberg: Winter.

Trudgill, Peter 2004. The impact of language contact and social structure on linguistic structure. Focus on the dialects of modern Greek. In: Kortmann, Bernd (ed.) *Dialectology meets typology. Dialect grammar from a cross-linguistic perspective*, p. 435–451. Berlin/New York: de Gruyter.

Trudgill, Peter 2009. Sociolinguistic typology and complexification. In: Sampson, Geoffrey, Gil, David; Trudgill, Peter (ed.) *Language Complexity as an Evolving Variable*, p. 98–109. Oxford: Oxford University Press.

Trudgill, Peter 2011. *Sociolinguistic typology. Social determinants of linguistic complexity*. Oxford: Oxford Univ. Press.

Wipf, Elisa 1911. *Die Mundart von Visperterminen im Wallis*. Frauenfeld: Huber.

Wurzel, Wolfgang U. 1984. *Flexionsmorphologie und Natürlichkeit: ein Beitrag zur morphologischen Theoriebildung*. Berlin: Akademie-Verlag.

Zürrer, Peter 1999. *Sprachinseldialekte. Walserdeutsch im Aosta-Tal (Italien)*. Aarau; Frankfurt am Main; Salzburg: Sauerländer.

Appendix

Table 10: noun inflection in OHG

		SG					PL			
	IC	nom	acc	dat	gen	instr	nom	acc	dat	gen
	1	tag	tag	tag-e	tag-es	tag-o	tag-a	tag-a	tag-on	tag-o
	2	hirt-i	hirt-i	hirt-e	hirt-es	hirt-u	hirt-a	hirt-a	hirt-on	hirt-o
	3	gast	gast	gaste-e	gaste-es	gaste-u	gest-i	gest-i	gest-in	gest-o
	4	win-i	win-i	win-e	win-es		win-i	win-i	win-in	win-o
	5	sit-u	sit-u	sit-e	sit-es	sit-u	sit-o	sit-i	sit-in	sit-i
	6	han-o	han-un	han-in	han-in		han-un	han-un	han-on	han-ono
	7	fater	fater	fater-e	fater-es		fater-a	fater-a	fater-un	fater-o
	8	wort	wort	wort-e	wort-es	wort-o	wort	wort	wort-on	wort-o
	9	lamb	lamb	lamb-e	lamb-es	lamb-o	lemb-ir	lemb-ir	lemb-ir-on	lemb-ir-o
	10	kunn-i	kunn-i	kunn-e	kunn-es	kunn-o	kunn-i	kunn-i	kunn-in	kunn-o
	11	herz-a	herz-a	herz-in	herz-in		herz-un	herz-un	herz-on	herz-ono
	12	geb-a	geb-a	geb-u	geb-a		geb-a	geb-a	geb-on	geb-ono
	13	kuningin	-a	-u	-a		-a	-a	-on	-ono
	14	anst	anst	enst-i	enst-i		enst-i	enst-i	enst-in	enst-o
	15	zung-a	zung-un	zung-un	zung-un		zung-un	zung-un	zung-on	zung-ono
	16	hoh-i	hoh-i	hoh-i	hoh-i		hoh-i	hoh-i	hoh-in	hoh-ino
	17	muoter	muoter	muoter	muoter		muoter	muoter	muoter-un	muoter-o
	18	naht	naht	naht	naht		naht	naht	naht-on	naht-o
wa-stem	?	hleo	hleo	hlew-e	hlew-es		hlew-a	hlew-a	hlew-on	hlew-o
wa-stem	?	horo	horo	horaw-e	horaw-es		horo	horo	horaw-on	horaw-o

Table 11: noun inflection in NHG

	SG				PL				
IC	nom	acc	dat	gen	nom	acc	dat	gen	
1	gast	gast	gast	gast-es	gäst-e	gäst-e	gäst-en	gäst-e	
2	tag	tag	tag	tag-es	tag-e	tag-e	tag-en	tag-e	
3	schaden	schaden	schaden	schaden-s	schäden	schäden	schäden	schäden	
4	brunnen	brunnen	brunnen	brunnen-s	brunnen	brunnen	brunnen	brunnen	
5	vater	vater	vater	vater-s	väter	väter	väter-n	väter	
6	lehrer	lehrer	lehrer	lehrer-s	lehrer	lehrer	lehrer-n	lehrer	

7	wald	wald	wald	wald-es	wäld-er	wäld-er	wäld-er-n	wäld-er	bild-er
8	matrose	matrose-n	matrose-n	matrose-n	matrose-n	matrose-n	matrose-n	matrose-n	
9	staat	staat	staat	staat-s	staat-en	staat-en	staat-en	staat-en	konto
10	blume	blume	blume	blume	blume-n	blume-n	blume-n	blume-n	pizza
11	stadt	stadt	stadt	stadt	städt-e	städt-e	städt-e-n	städt-e	
12	mutter	mutter	mutter	mutter	mütter	mütter	mütter-n	mütter	
13	zoo	zoo	zoo	zoo-s	zoo-s	zoo-s	zoo-s	zoo-s	
14	pizza	pizza	pizza	pizza	pizza-s	pizza-s	pizza-s	pizza-s	

Table 12: noun inflection in Kaiserstuhl Alemannic

	SG			PL			
IC	nom	acc	dat	nom	acc	dat	
1	braif	=nom	=nom	briaf	=nom	=nom	
2	gumb	=nom	=nom	gimb	=nom	=nom	
3	schdai	=nom	=nom	schdai-n-er	=nom	=nom	wäld-er
4	grab	=nom	=nom	grab-a	=nom	=nom	
5	ghuch-i	=nom	=nom	ghuch-ana	=nom	=nom	
6	dand-a	=nom	=nom	dand-ana	=nom	=nom	
7	baziand-i	=nom	=nom	baziand-inna	=nom	=nom	

Table 13: noun inflection in Visperterminen Alemannic

	SG				PL			
IC	nom	acc	dat	gen	nom	acc	dat	gen
1	tag	tag	tag	tag-sch	tag-a	tag-a	tag-u	tag-o
2	chopf	chopf	chopf	chopf-sch	chepf	chepf	chepf-u	chepf-o
3	ar-o	ar-o	ar-u	ar-u	arm-a	arm-a	arm-u	arm-o
4	santim	santim	santim	santim-sch	santim	santim	santim	santim
5	han-o	han-o	han-u	han-u	han-e	han-e	han-u	han-o
6	bog-o	bog-o	bog-u	bog-u	beg-e	beg-e	beg-u	beg-o
7	senn-o	senn-o	senn-u	senn-u	senn-u	senn-u	senn-u	senn-o
8	jar	jar	jar	jar-sch	jar	jar	jar-u	jar-o
9	hor-u	hor-u	hor	hor-sch	hor-u	hor-u	horn-u	hor-o
10	chrut	chrut	chrut	chrut-sch	chrit-er	chrit-er	chrit-er-u	chrit-er-o
11	lamm	lamm	lamm	lamm-sch	lamm-er	lamm-er	lamm-er-u	lamm-er-o
12	redli	redli	redli	redli-sch	redli-n-i	redli-n-i	redli-n-u	redli-n-o

13	öig	öig	öig	öig-sch	öig-u	öig-u	öig-u	öig-o
14	farb	farb	farb	farb	farb-e	farb-e	farb-u	farb-o
15	bon	bon	bon	bon	bon-a	bon-a	bon-u	bon-o
16	sach	sach	sach	sach	sach-u	sach-u	sach-u	sach-o
17	mus	mus	mus	mus	mis	mis	mis-u	mis-o
18	tsun-a	tsun-a	tsun-u	tsun-u	tsun-e	tsun-e	tsun-u	tsun-o

Table 14: noun inflection in Issime Alemannic

	SG				PL			
IC	nom	acc	dat	gen	nom	acc	dat	gen
1	weg	weg	weg	weg-sch	weg-a	weg-a	weg-e	weg-u
2	uav-e	uav-e	uav-e	uav-endsch	uav-n-a	uav-n-a	uav-n-e	uav-n-u
3	noam-e	noam-e	noam-e	noam-endsch	noam-i	noam-i	noam-e	noam-u
4	hoan-u	hoan-u	hoan-e	hoan-ensch	hoan-i	hoan-i	hoan-u	hoan-u
5	vus	vus	vus	vus-sch	vüs	vüs	vüs-e	vüs-u
6	att-u	att-u	att-e	att-e	att-i	att-i	att-e	att-e
7	schu	schu	schu	schu-sch	schu	schu	schun-e	schun-u
8	sia	sia	sia	sia-sch	sia-w-a	sia-w-a	sia-w-e	sia-w-u
9	bet	bet	bet	bet-sch	bet-i	bet-i	bet-u	bet-u
10	berri	berri	berri	berri-sch	berri-n-i	berri-n-i	berri-n-u	berri-n-u
11	lam	lam	lam	lam-sch	lamm-er	lamm-er	lamm-er-e	lamm-er-u
12	lan	lan	lan	lan-sch	lenn-er	lenn-er	lenn-er-e	lenn-er-u
13	matt-u	matt-u	matt-u	matt-u	matt-i	matt-i	matt-u	matt-u
14	mum-a	mum-a	mum-u	mum-u	mum-i	mum-i	mum-u	mum-u
15	chötti	chötti	chötti	chötti	chötti-n-i	chötti-n-i	chötti-n-u	chötti-n-u
16	schuld	schuld	schuld	schuld	schuld-in-i	schuld-in-i	schuld-in-u	schuld-in-u
17	nacht	nacht	nacht	nacht	necht-in-i	necht-in-i	necht-in-u	necht-in-u
18	han	han	han	han	hen	hen	hen-e	hen-u
19	geiss	geiss	geiss	geiss	geiss	geiss	geiss-e	geiss-u

On the classification of compound verbs[1]

Alexandra Bagasheva
Sofia University "St. Kliment Ohridski"
abagasheva@gmail.com

1. Introduction

The classificatory scheme one uses and the framework of analysis one applies often skew the identification and interpretation of compounds. Traditionally compounds have been divided into *synthetic* (also called deverbal) such as *horse-riding*, *house-trained* and *root* (also called *primary*) compounds such as *apple pie*, *snow ball* (ten Hacken 2010; Scalise and Bisetto 2009). This classificatory scheme has influenced the understanding and analysis of compounds. The traditional classification described above has been significantly improved by the scheme recently proposed by Scalise and Bisetto (2009), which has led to looking more carefully at the data of compounding (in English), and indeed to finding a much wider range of compound types. Even this classificatory scheme fails to provide an adequate space for the complex nature of compound verbs (henceforth CVs), which seem to disrupt neat classificatory schemes for compounds. Assuming acategorial status of the compound-internal constituents of a CV, postulating a dedicated constructional idiom within an hierarchically organized lexicon and allowing for a dissociation between a word formation process and its products creates a more coherent context for discussing the nature of CVs (in English and Bulgarian) and their classification.

2. Classifications of compounds

The few existing specific CV/VV classifying systems are not consistently (if at all) utilized in the mainstream word-formation literature. Instead, the general classifying systems of compounds are directly applied to CVs in English in the belief that they can well be accommodated within them. Thus, if we uncritically apply the familiar categorization of compounds into root and synthetic compounds to CVs we would have to recognize *sleep-walk* as coordinate and by implication root compound, while *head-hunt* would be analysed as a subordinate, synthetic one. The coordinate (and implicationally derived root) status of *sleep-walk* runs into contradiction with the properties which the CV is presumed to acquire via the word-formation process applied in its creation, namely back-formation. According to Scalise and Bisetto (2009), the root/synthetic parameter is based on language-specific criteria (suited specifically to the reality of compound types in English) and for this reason not widely applicable. The distinction, in our view, is problematic even for English as it involves the recognition of a verbal base in the second group (e.g. *book-keeping*, *truck driver*). Naturally, this would suggest that all CVs are synthetic compounds because they contain a verbal base. Such a generalization is counterintuitive as among CVs we can recognize VVs (e.g. *stir-fry*, *crash-land*) which resemble nominal root compounds in terms of a direct concatenative pattern. Scalise and Bisetto's (2009) classificatory system, which recognizes coordinate, attributive and subordinate compound types with exo- and endocentric variants in each group, avoids

[1] The evolution of the ideas and the consequtive focusing of the argumentation presented here can be traced in two previous publications, which in expounding on the nature of CVs in the context of different research questions also discuss the advancement of a possible classificatory system for compound verbs (Bagasheva 2011a and Bagasheva 2011b).

the inadequacy of the root/synthetic opposition. However, as argued and illustrated below, even this classificatory system does not accommodate all significant properties of CVs in English and Bulgarian.

Sharing Bauer's dissatisfaction with all proposed schemes for the classification of compounds (Bauer 2006: 724), we believe that none of the available classifying nomenclatures of compounds captures the most relevant variable properties of CVs which should ideally be reflected in the criteria employed for their classification. The ample literature on compound classification (Scalise and Bisetto 2009; Booij 2005; Haspelmath 2002; Bauer 2001; Fabb 2001, to name but a few) provides diverse and sometimes contradictory specific criteria for the classification of CVs, and this generally leads to a heterogeneous set of classificatory systems and terminological confusion. The basic criteria[2] traditionally adopted for the classification of compounds include headedness, the nature of the relationship between the constituents, internal semantics, categorial labels of the head constituent, etc.

Despite their scarcity in comparison to general compound classifying systems, specific classifications of CVs exist. In his explicit classification of CVs Bauer (1983: 207-209) suggests that CVs can be classified by "form class", which leads to the identification of the following classes: "Noun + Verb (e.g. *carbon-date*); Verb + Noun (e.g. *shunpike*); Verb + Verb (e.g. *freeze-dry*); Adjective + Verb (e.g. *free-associate*); Preposition + Verb (e.g. *overachieve*); Adjective + Noun (e.g. *bad-mouth*) and Noun + Noun (e.g. *breath-test*)" (ibid.). Thus the heterogeneous class of CVs is ordered into sets on the basis of presumed well-specified part-of-speech categorial marking of the constituents of a CV. The exact "method of formation" of a CV in each case is not taken into account. Applying this classificatory scheme leads to the recognition of the following CV classes based on lexical categoriality of the constituents:

Table 1: Formal types of CVs

N + V	carbon-copy, babysit, blockbust, главозамайвам се
A + V	soft-land, fine-tune, whitewash, благоустройвам
Prprep + V	outnumber, overachieve, underrate, задминавам, прескачам
V+V	drink-drive, crash-land, dry-clean, stir-fry, Ø
A + N	brown-bag, bad-mouth, blacklist, Ø
N + N	breath-test, Ø
Num+V	double-cross, double-check, двоумя се

This form-based structural classification is supplemented by in-group specifications based on the method/process of formation (which is usually taken to predetermine the meaning properties of CVs, cf. Guevara and Scalise 2004, Lieber 2004, Nagano 2007). Thus the following three classes of CVs are identified:

Table 2: Types of CVs according to derivation pattern

back-derived CVs	converted CVs	compounded CVs
head-hunt ← head-hunting	sandbag → sandbag	drip-dry
breast-feed ← breast-feeding	blacklist → blacklist	sweet-talk
stage-manage ← stage-managing	railroad → railroad	fast-talk

[2] See Scalise and Bisetto 2009 for a detailed presentation and analysis of available classificatory systems and the criteria they adopt.

In another explicit classification of CVs, narrowed to one of Bauer's classes (VV), Shibatani (1990) suggests the following classifying scheme:

a) modifier-V – where the modifier names the manner of the activity named by the second verb
b) V-modifier – where the second constituent identifies the manner or direction of the verb
c) V-V – where both verbs have equal semantic contribution to the semantics of the whole, naming a complex event (based on Shibatani 1990: 246).

As the classification is proposed in the context of discussing the nature of Japanese VVs, it is not supposed to naturally apply to CVs in English and Bulgarian. The second type V-modifier is not characteristic of English, but the first and third types are attested (e.g. *deep-fry, stir-fry*). In Bulgarian only the first type can be recognized with certain qualifications (e.g. *злословя* [zloslovya, ill-speak, *bad-mouth*]). CVs of the third type (V-V) are not attested in Bulgarian. The distinction between the two types (modifier-V and V-V) resembles the distinction between *stir-fry* and *tap-dance*. The former is classified as a coordinate simultaneous compound (Lieber 2009), the latter is interpreted in varying ways depending on the recognition of *tap* as the activity of tapping or as *a tap* attached to shoes (Wald and Besserman 2002). Admittedly, the first class of VV (modifier-V) is recognizable in cases in which the nature of the first constituent is undecided between a Noun interpretation and a Verb interpretation. In Bauer's commentary of VVs (Bauer 1983: 208) the basic property of the type is the indeterminacy of the first constituent – the four VVs he discusses all display this property. The example he quotes from Marchand "*type-write* might but probably does not belong" to the type. Adams' example *test-market* is dubbed "dubious", "*freeze-dry* does not unambiguously belong in this class either" and the fourth VV *trickle-irrigate* "could be noun + verb or back formation from trickle-irrigation". Wald and Besserman succinctly summarise the problem indicating that

> [c]oncerns about NV are most intimately related to concerns about VV in the very frequent apparent ambiguity of category of the first constituent of the compound, e.g., *sleep* – noun (N1) or verb (V1)? – in *sleep*-walk (Wald and Besserman 2002: 417).

The authors do not specifically address problems of classifications of compounds but devote much of their discussion to the ambiguous category problem, which is among the central problems in the analysis of VVs. They achieve uniformity in the treatment of VVs by settling for the recognition of *possible* VVs coming from various diachronic sources and suggesting that a uniform synchronic analysis is possible if we take into account the activity constraint[3] (for the details of their argumentation see Wald and Besserman 2002). Instead of opting for a uniform possible V interpretation of the first constituent, to avoid the first constituent status controversy, we suggest that the constituents in a CV are categorially indeterminate (an argument we take up in the next section).

Lieber, adopting Scalise and Bisetto's (2009) classification scheme and broadening the subordinate class to include subject-oriented compounds, notes the following about CVs in English, "V+V endocentric compounds can be found, but the type is unproductive: MORBO contains *trickle-irrigate*, and a few others come to mind (*slam-dunk*, *blow-dry*), but these are not freely formed" (Lieber 2009: 359). In two subsequent tables summarizing the types of compounds characteristic of English as an IE, Germanic language, the author classifies *stir-fry* as a simultaneous endocentric coordinate

[3] "The activity constraint imposes a formal and semantic economy on the internal structure of a compound verb by limiting it to what is necessary to the semantic interpretation of that compound verb, and suppressing what is unnecessary in that context. Thus, when nominal or adjectival marking alters the syntactic properties of the first constituent of a compound, but does not otherwise alter its semantic properties, it is suppressed in favour of the root verb alone in the verb form of that compound" (Wald and Besserman 2002: 423).

compound and *headhunt*, *machine-wash* and *spoon-feed* as "endocentric verb-containing subordinate compounds of the output category V" (ibid.: 360-361) and dubs these "a marginal class" (ibid. 361).

Table 3: A summary of Lieber's (2009: 359-361) classification of CVs in English

	subordinate	coordinate
endocentric	(object) *head-hunt*, [*гласоподавам*][4] (subject) *machine-wash*, [Ø] (adjunct) *spoon-feed, [зловиди ми се]* (manner) *deep-fry,* [Ø]	*trickle-irrigate*, *slam-dunk*, *blow-dry*, *stir-fry;* [Ø]
exocentric	non-attested	non-attested

Counter Lieber, Bauer believes that "[t]o trickle-irrigate is to irrigate in a particular way" (Bauer n.d: 8). In a like manner,

> *stir-fry* and *freeze-dry* are both headed compounds. The hyponymy test works well here: *stir-frying* is a kind of *frying* (not a kind of *stirring*) and *freeze-drying* is a kind of *drying* (achieved by *freezing*). Again they seem to be excluded from the set of dvandvas (Bauer 2008: 4).

Alongside these criteria, internal (thematic) relations are often exploited as a basis for classifying CVs. Basing the lexico-semantic analysis of CVs on derivational properties has led to the establishment of the following verb-internal relations, which are defined on the basis of the source or parent compound (root or synthetic) giving rise to the CV (Clark and Clark 1979, Nagano 2007) and in accordance with postulated argument relations or semantic roles (Lieber 2004, 2009).

Table 4: Classification of CVs according to compound-internal relations

Object	babysit, head-hunt, прахосмуча
Instrument	spoon-feed, tumble-dry, ръкомахам
Manner	free-associate, soft-land, славословя
Place	quarter-deck, house-train, Ø
Time	day-dream, moon-light, Ø

There is yet another kind of classification provided for converted CVs specifically. Working on Clark and Clark's (1979) thesis of contextuals in conversion, Nagano (2007) comes up with the following classificatory scheme for CVs (both converted and back-formed[5]) based on their semantics,

> "BF from compound nouns or adjectives is semantically parallel to conversion:
> (a) **Locatum**: air-condition (<air-conditioner$_N$), face-lift (<face-lifting$_N$), ill-treat (<ill-treatment$_N$), etc[6].
> (b) **Location**:
> (c) **Goal**: hard-boil (<hard-boiled$_A$), horrorstrike (<horror struck$_A$), jam-pack (<jam-packed$_A$), tongue-tie (<tongue-tied$_A$)

[4] The Bulgarian examples have been introduced by the author, while the English ones belong to Lieber as indicated by the quotation.

[5] The main hypothesis on which this classificatory scheme is based is that back-formation should be interpreted as conversion. The author concludes that "BF [back-formation] consists of conversion, a rule-based word-formation process, and clipping, a non-rule-based speech-level process, and the various properties of BF have been proved to be deducible form the properties of these two processes" (Nagano 2007: 68).

[6] The author provides numerous examples for each type only a subset of which are directly quoted here for brevity purposes. The use of "etc." indicates that many of the examples for a particular class have been left out in the quotation.

> (d) **Manner**: art-edit (<art-editor$_N$), baby-sit (<baby-sitter$_N$), match-make (<match-maker$_N$), etc. tub-thump (<tub-thumper$_N$), etc.
> (e) **Instrument**: hang glide (<hang glider$_N$), knuckle-dust (<knuckle-duster$_N$), loud-hail (<loud-hailer$_N$), etc.
> (f) **Duration**:
> (g) **Source**:
> (h) **Crop**: bird's-nest (<bird's-nesting$_N$) (Jespersen 1942, 101)
> (i) **Action**: (i) book-hunt (<book-hunting$_N$), hand-write (<hand-writing$_N$), house-clean (<house-cleaning$_N$), house-keep (<house-keeping$_N$), job-hunt (<job-hunting$_N$), etc. (ii) affix-hop (<affix-hopping$_N$), brainstorm (<brainstorming$_N$), etc.
> (j) **Sound symbolism**: prize-fight (<prize-fighter$_N$)
> (k) **Unclassifiable**: cliff-hang (<cliff-hanger$_N$), frostbite (<frostbiting$_N$) (Matsuda 1999), logroll (<logrolling$_N$), show-jump (<show-jumping$_N$), shadow-cast (<shadow-casting$_N$), skywrite (<skywriting$_N$)" (Nagano 2007: 62-63; bold face added for clarity and ease of reading).

The belief that the parent noun is necessarily involved in the meaning generation mechanism of a denominal CV leads to implausible lexical semantic interpretations and classifications of attested CVs: see above *air-condition, face-lift, ill-treat, ill-use, pressure-treat, triple-tongue, turbocharge, valet-park* classified as Locatum verbs while *tailor-make*, *jam-pack* and *hard-boil* as Goal and *prize-fight* as Sound-symbolism.

Another semantics-based classification has been proposed, which tries to locally classify the VV structural subtype of CVs. Renner (2008: 611) elaborates the semantic classification of VVs thus

> V.V coordinate compounds belong to three semantic categories: asynchronous compounds, synchronous compounds, and disjunctive compounds. The classification is based on paraphrases, which reveal simultaneity or consecutiveness of events. The disjunctive type contains verbal constituents but its members are nouns and adjectives (e.g. *lend-lease* and *pass-fail*).

This classificatory scheme is doubly restricted: first it takes into account only coordinate verbs (where the coordinate status of the internal relations directly ensues from the lexical categorial status of the first constituent) and second it focuses on VVs exclusively, which is preconditioned by the first criterion in the classification – a coordinate relationship which can only obtain between syntactic elements with the same status (or functional uniformity).

When applying general compound classificatory schemes (criteria) to the classification of CVs, intra-family classificatory dissociations arise, e.g. *stir-fry* is classified as coordinate endocentric as opposed to another member of the word-formation niche *deep-fry* which is categorized as subordinate exocentric (for the definition and discussion of the nature and analytical utility of word-formation niches see Hüning 2009).

3. The problem

The adoption of Scalise and Bisetto's (2009) classifying system seems promising. It predicts the division of CVs into coordinate and subordinate, as an attributive relation is precluded between a verb and an element in its frame (modification is admissible but it is of a different nature from the attributive type of relation). Within each class there is room for distinguishing between endocentric and exocentric CVs, though exocentricity is not recognized as operative in the CV lexicon by Lieber (2009: 360-361, see specifically tables 18.1. and 18.2). As becomes obvious from Lieber's classification of CVs in English (see above), finer subdivisions in the specified endocentric and exocentric subgroups can be established, which relate to the simultanous or consecutive ordering of subevents in a complex event (e.g. *stir-fry*). These finer distinctions presume a classification based on a definite recognition of verbal vs. nominal/adjectival categorial status of the first

constituent in a CV. The simultaneity/consecutiveness distinction is applicable only in cases in which we recognize a VV compound, which, according to the higher distinguishing property, would all be classified as coordinate.

The scholars' wide-ranging disagreement (cf. Lieber 2009 vs. Bauer n.d., 2008) and the ensuing intra-family classificatory dissociations (cf. *stir-fry* vs. *deep-fry* or *drip-dry* vs. *sun-dry*) hinge crucially on two factors:

a) the assumption that compound constituents have specified lexical (categorial) status (V vs. N, etc.);
b) the premise that compound-internal relations fully subscribe to syntax-determined relations – subordination, coordination and attribution (which are also ultimately dependent on categoriality considerations as the chosen types of relations require categoriality specifications).

The presumably opposed CVs in intra-family dissociations seem to occupy a single semantic space and to develop an identical frame with different values assigned to the relevant dimension of the frame activated in the CV with the MANNER / TO A CERTAIN EFFECT conceptual space activated and symbolically represented. They belong to well-established word-formation niches, which in our view, have unified semantics. We assume that suspending the categoriality of constituents might lead to interesting results concerning the classification of CVs. The question is whether we have good reasons to allow for acategoriality of CV internal constituents?

4. Categorial indeterminacy of CV constituents

Part-of-speech classes are assumed to correlate with experiential complexes (when notionally defined). For many speakers the semantic, syntactic and formal distinctions between nouns and verbs correlate unequivocally with the way they experience the world. As Laudanna observes,

> [f]irst and foremost for speakers of Indo-European languages, language is arranged in such a manner that on the one side it compels to think of the world in terms of nouns as names for objects and verbs as names for actions. On the other side, the phenomenological experience of the world – made up of entities and processes – favours and/or strengthens the characterization of nouns and verbs as labels for the former and the latter, respectively. The *naive* way of thinking, but sometimes *even the scientific reasoning*, is based on this approach to a *supposedly meaningful partition* of the world (Laudanna 2002: 3, emphasis added).

But ongoing debates concerning the cross- and intra-linguistic realities of part-of-speech distinctions and the principles and criteria for their recognition reveal "growing evidence to suggest that the verb-noun distinction is scalar rather than discrete" (Rijkhoff 2002: 115).

This general noun/verb indeterminacy relates directly to the categotiality of constituents controversy. The status of compound constituents as lexemes or root/stems has not been unambiguously settled. Bauer's (2001) idea of formal isolation as a basic criterion for compoundhood is open to interpretations and permist acategorial treatment of the constituents.

> Compound is a lexical unit made up of two or more elements, each of which *can function* as a lexeme independent of the other(s) in other contexts, and which shows *some phonological and/or grammatical isolation from normal syntactic usage* (Bauer 2001: 695, emphasis added).

The fact that it is possible, but not necessary, for a compound constituent to have independent lexemic status, i.e. the optionality of lexemehood and the stipulation for grammatical isolation from normal syntactic usage open up the possibility for postulating categorial indeterminacy of CV constituents. Without explicitly or totally dismissing the relevance of lexical categoriality of CV constituents, Bauer opens the way for relaxing the N/V debate in relation to CV internal constituency.

Another implicit prerequisite for such an approach can be traced back to Anderson's contention about the uniqueness of compounds as structural units.

> Compounding ... involves the combining of stems from the lexicon into a *quasi-syntactic structure*. This word-internal structure seems to be *unique* to compounds, in fact.... (Anderson 1992:292; emphasis added)

Indeed, one is tempted to claim that the N/V indeterminacy is among the properties that make compounds unique among linguistic elements, but such a conclusion is premature. Hopper and Thompson (2004/1984) put forward the hypothesis about the general categorial indeterminacy of traditional parts-of-speech classifications. They claim that the lexical and semantic properties of verbhood and nounhood are secondary and are primed and ultimately determined by their discourse roles, i.e. the determinants of nounhood and verbhood are predominantly pragmatic (Hopper and Thompson 2004: 287) and coerced by syntagmatic relations. The actual proposal the authors make is that linguistic entities set out as acategorial elements, i.e.

> the continua which in principle begin with acategoriality, and which end with fully implemented nounhood or fully implemented verbhood, are already partly traversed for most forms. In other words, most forms begin with a propensity or predisposition to become Ns or Vs; and often this momentum can be reversed by only special morphology. It nonetheless remains true that this predisposition is only a latent one, which will not be manifested unless there is pressure from the discourse for this to occur (Hopper and Thompson 2004: 287).

In parallel to their suggestions it is plausible to assume that linguistic elements making up a CV set out as acategorial elements. When they are coerced by the dedicated constructional idiom it ascribes the whole a verbal categorial marking. The first constituent ambiguity is easily avoided if we accept the acategorial status of constituents.

Findings in psycho- and neuro-linguistic research gave Laudanna grounds to conclude that

> [l]inguistically based concepts articulated in terms of categories like "noun" and "verb" are supposed to be the epiphenomena of correlated clusters of elementary features. They are not thought to correspond to distinct cognitive representations; rather, they just mark different values of continuous variables like, for instance, perceptual features (Laudanna 2002: 6).

From a purely linguistic point of view, Rijkhoff argues that even "in languages that do have a more or less rigid distinction between verbs and nouns, members of both word classes can be analyzed in a similar fashion semantically" (2002: 141). Such arguments point to the plausibility of ascribing acategoriality to CV constituents and adopting semantic criteria for analyzing and classifying CVs. As far as English and Bulgarian CVs are concerned, first constituents never bear explicit morphological marking and have predominantly semantic contribution.

Furthermore, in support and even as an extension of Rijkhoff's contentions, Vogel (2000: 263) claims that Modern English has undergone a "degrammaticalization shift from a 'specialized' noun-verb language (with a grammaticalized part-of-speech system) towards a 'flexible' type-token language (without a grammaticalized part-of-speech system)." For the more conservative, Vogel suggests that English might be thought of as having two parallel part-of-speech systems: "Thus, there are now two overlapping systems: a specialized noun-verb-adjective-adverb-system and a flexible noun/verb/adjective-adverb-system" (ibid. 277). We claim that it is the flexible system that is utilized in compounding.

Table 5: Vogel's summary of the two part-of-speech systems in English

Specialised	V	N	Adj	Adv
Flexible	V/N/Adj			Adv

(Vogel 2000: 277).

Further evidence for the acategorial status of CV-internal constituents can be found in Farrell's (2001) contention that nominal/verbal construal is a matter of alternative profiling of underspecified symbolic units which are related via functional shifts. The lexical semantic representations of such words include event schemas that are compatible with either noun or verb meanings. The verb vs. noun aspect of the meanings is supplied by the morphosyntactic contexts in which they appear (Farrell 2001: 109). Thus it appears that in "thinking for speaking" (Slobin 2003: 158) a speaker has at their disposal alternative scenarios whose employment in a particular communicative event will depend exclusively on immediate situational variables and will be exceptionally pragmatically conditioned. Farrell's argument is couched in his analysis of conversion as a word-formation pattern. Within this model conversion from compound nouns no longer necessitates the functioning of the noun as an argument or semantic determinant of the newly formed verbs.

The acategoriality postulate may well capture the fluidity of conceptualization in the sense that on hearing a linguistic element a listener builds interpretative hypotheses which need not necessarily involve categorially marked treatment of constituents, even though there are marked tendencies as evidenced by the processing of garden-path sentences. But the fact that contradictions raised by garden-path sentences are resolved without much effort as they unfold indicates that categoriality marking is pragmatically superseded. Consequently, we might hypothesize that the constituents of CVs have phonetic shape, conceptual frame activation but no categorial marking. The acceptance of categorially undetermined constituents is beneficial not only for analyzing CVs in a unified manner, but seems like a probable line of research concerning the bracketing paradoxes of synthetic nominal compounds and provides for a functionally and pragmatically informed classification of CVs. From a methodological perspective, the acategorial treatment of constituents is fully justifiable in a constructionist theory because the constructions themselves have a significant contribution to specifying the properties of the linguistic items that realize them in particular instantiations.

5. Headedness and CVs

To add substance to our arguments we also need to consider how lexical categoriality of compound internal constituents in general and CV ones in particular interact with the headedness properties of compounds. Headedness remains a controversial issue in compounding even today. Scalise and Fábregas (2010) admit to the possibility of multiple heads in a single compound which is fully congruent with Scalise, Fábregas and Forza's (n.d.) parameterized treatment of exocentricity based on a three-fold understanding of head and headedness. The three types of exocentricity identified are categorial, morphological and semantic. The authors define categorial exocentricity as the case in which "the constituent in the head position does not impose its categorial features to the whole construction" (Scalise, Fábregas, Forza n.d.: 61).

A special case of categorial exocentricity is ACE (Absolute Categorial Exocentricity, cf. *pass-fail*) (for a summary of treatments of exocentricity and its role in CVs see Bagasheva 2011c). The authors define ACE as the phenomenon of the output being completely different from the input categories (ibid.: 55). Morphological exocentricity which is defined as the case in which "morphological features of the compound are not identical to the morphological features of any of its internal constituents" (ibid.: 62). In the view of the authors, this type of exocentricity is highly sensitive to type of language and the general theoretical framework adopted as regards the concept of morphological features. When operationalized as an analytical concept, semantic exocentricity is identified when

"the semantic type of the compound cannot be derived from the semantic type of any of its constituents" (ibid.).

In a similar vein, Scalise and Fábregas (2010: 124) contend that "[i]t could well be the case that inside a compound different elements can be identified as heads, depending on which features we are considering." We fully subscribe to the views of the authors expounded on above.

Thus for inflectional (i.e. categorial purposes) CVs in both English and Bulgarian are right-headed and inflections are marked compound-externally (e.g. *гласоподава<u>м</u>, гласоподава<u>ш</u>, гласоподава<u>хме</u>*, etc.; *has been pink-slipp<u>ed</u>*, *pink-slip<u>s</u>*, etc.). This uniformity is not a chance coincidence (despite the different morphological systems of the two languages), but the result of the regular process of constructional coercion which operates in CV creation.

The morphological understanding of headedness is not pertinent for CVs because it is difficult to trace the percolation of morphological features in CVs which arise from conversion or back-formation, not compounding proper, but still share all the properties of compound lexical objects.

The last type of headedness, semantic headedness, is particularly pertinent in the analysis of CVs. Scalise and Fábregas (2010: 121) propose to define a semantic head as "the constituent whose semantic contribution allows us to determine the class of objects denoted by the compound." In our view it is not any of the compound-internal constituents that determines the denotation of the CV, but the dedicated constructional idiom which determines the verbal profiling. Depending on the contribution of the input semantic frames, we can distinguish two general types of CVs – some in which the constituents contribute comparably by functioning as inputs to the CV frame configuring, and some in which the lexical meaning of the resulting CV is not directly dependent of the input frames as semantic contributors. Rather, the constructional idiom reinforces a conceptual reinterpretation congruent with the immediate context, later subject to a subsequent process of semantic drift or lexicalization via various linguistic and cognitive mechanisms.

To recap, the head in English and Bulgarian CVs is determined neither positionally nor morphologically, in keeping with Štekauer's onomasiological understanding of headedness. Rather, it is identified with an onomasiological base defined "as that constituent of the onomasiological structure which stands for the whole group or class of objects" (Štekauer 2005: 225). By implication, the onomasiological base is the most general constituent of the onomasiological structure. "The criterion of headedness is thus shifted to the *conceptual level* of the WF process" (ibid.). This is the only type of headedness whose analysis can help establish relevant distinctions in types of CVs. We assume that, morphologically and categorially, the dedicated CV constructional idiom, which coerces the verbal construal, functions as a categorial and morphological head, while the typology of CVs hinges on the nature of the semantic configuring executed.

6. An Alternative Classification of CVs

To replace the attributive, coordinate and subordinate classificatory model (which leads to unnatural disruptions of intra-niche unity), a new unified model based on the specific semantic configuring in CV subschemas is proposed. After all, "[t]he primary purpose of a good classification is to enable the linguist to make the best generalizations possible about linguistic phenomena" (Booij 2005: 110). The classification operates with scalar criteria and may be attacked for being vague. The strongest argument against such criticism is that the classificatory principle adopted (i.e. unity of linguistically relevant schemata and their dedicated constructional idioms which display hierarchical inheritance relations) tallies with intuitive users' knowledge and use of constructions and is not neatly tailored by and for the purposes of the analysts' theoretical constructs.

Most classifying systems are defined with a particular purpose in mind and work within an overall rationale. From the point of view of word-formation objects with a specific onomasiological function – to collapse the relation/conceptual core distinction and to both name and describe an event – CVs represent a unified class with numerous shared properties which make them distinct from all other compounds. The classification scheme proposed here tries to capture the lexico-semantic properties of CVs as word-formation products actualized as subschemas of a constructional idiom, directly utilized by speakers in their generation and by listeners in their interpretation.

The hypothesis put forward hinges on the application of two basic and closely interlinked criteria, both of which represent clines rather than discrete sets, in order to exhaustively and revealingly classify CVs in English. Both are semantic in nature, but while the first concerns the mechanism of internally configuring of their semantics, the second concerns the semantics of the lexically specified construction, i.e. the external semantics of CVs.

The first criterion employed relates to the internal constituency of CVs, where by constituency is understood the nature of the conceptual relation or configuring between the acategorial constituents – within the following two extremes: a) a relational property embedded within a relation (e.g. *force-feed, злословя*) and b) a thing embedded within a relation (e.g. *boyfriend-drop, гласоподавам*). Thus CVs subdivide into two subschemas which inherit the categorial properties of the constructional idiom and develop specific distinct properties associated with a different underlying conceptual operation of classification – superclassification and subclassification. Before we proceed with the specific suggestion, we need to make it clear that the type of classification referred to here is epistemological, with no implications intended whatsoever in relation to grammatical classification. In his *Verb Classification in Australian languages* McGregor (2002) draws the following distinction in operations of classification:

- grammatical classification: systems of overt or covert classification of lexemes; and
- epistemological classification: systems of linguistic units that categorise a domain of (conceptual) referents (McGregor 2002: 22).

So our metaclassification (the establishment of types of CVs) is based on the two distinct kinds of epistemological classification (i.e. categorization of referents) which CVs linguistically encode. The author himself argues for a distinction between superclassification and subclassification in the verbal lexicon.

> Certain noun-verb compounds in English (e.g. *hand-pick, pistol-whip, horse-whip, test-drive*, etc.) also represent a type of verbal subclassification: they specify subtypes of the event denoted by the verb. Gooniyandi, by contrast, shows a system of verb superclassification (McGregor 2002: 5).

As can be surmised from the suggestions of the author, certain CVs in English are instances of verbal subclassification, which might be taken to resemble the endocentric modifier type. The same applies to subdivisions in the Bulgarian CV lexicon. By implication it can be concluded that other CVs in English do not belong to the subclassifying type. The question as to what other class they might belong to remains to be discussed. Our working hypothesis is that the second class is an example of a superclassifying system creating new epistemological types of activities. Such a distinction is not paralleled by the simplex verbal lexicon where epistemological classification is uniform and is based on situation types (Rappaport, Doron, and Sihel 2010).

No parallel is intended here in any way between the phenomenon of verb classification (an object language phenomenon) and CV classification in English (a meta-language phenomenon), nor any implication of essential similarities between CV in English or Bulgarian and CVs in Australian languages. There are no distinct verb classes

either in English or in Bulgarian[7]. McGregor's distinctions are used for the formulation of a hypothesis according to which we can draw an informative distinction between two groups of CVs – A and B. Group A members fit the definition of epistemological superclassification in which the CV does not specify a subtype of the event named by an associated simple verb (if there is such), but names a new semantic type of event (e.g. *dipsix*). Such verbs resemble McGregor's (2002: 5) description of superclassification in which verb classifiers indicate "to which [new] category the event belongs." Group B CVs closely resemble the verbs identified by McGregor (2002) as instances of subclassification.

Although McGregor "proposes that certain types of noun incorporation – specifically, Mithun's Type I lexical compounding and a subset of Type II manipulation of case (Mithun 1984) – involve verb subclassification" (McGregor 2002: 4), we would argue that a particular class of CVs in English represents an instance of epistemological superclassification. In analyzing the lexical semantics of *shoulder-surf* (which is defined by word spy as *to steal a computer password or access code by peeking over a person's shoulder while they type in the characters*) and *kitchen-sink* with its two meanings (according to word spy: a) *to announce all of a company's bad financial news at one time* and b) *when arguing or fighting with a partner, to complain not only about a recent problem, but also about numerous past problems*), it transpires that *shoulder-surfing* and *kitchen-sinking* are not subtypes of *surfing* and *sinking* respectively. Both verbs are undeniable instances of what has been identified as noun incorporation Type I lexical compounding, but neither is epistemologically a name for a subtype of the event named by the associated simplex verb. The analysis works for all CVs in Group A (even though not all of them are instances of Type I lexical compounding). These verbs are chosen as illustrative examples of the properties of Class A verbs for two basic reasons – they are recent creations, name socially significant activities, instantiate lexical compounding and no doubt involve complex metaphtonymic[8] processes of semantic change. Sticking for argument's sake to the Type I incorporation claim, and McGregor's supposition that these should be instances of subclassification, we would expect *kitchen-sinking* to name a subtype of *sinking*. *Sink* being associated with both transitive (causative) and intransitive uses, we would expect syntactic blocking to occur for the appearance of Type I incorporation. Blocking (if there is such) is superceded by naming needs that can be satisfied by the dedicated constructional idiom whose lexical specification in this instance leads to the lexicalization of a novel conceptual configuring, not specialization of the meaning of the "head" (which is presumed as a basic semantic operation in compounds, cf. Huddleston and Pullum 2002). Word spy defines the process of conceptual configuring of the verb thus,

> [t]his verb is based on the idiom *everything but the kitchen sink*, which hails from World War II. (Back then it referred to a heavy bombardment in which it appeared the enemy was firing *everything but the kitchen sink*.) The *verb is based on a sensible strategy*: If a company must divulge some bad news in its financial results, then it might as well bring all of its fiscal skeletons out of the accounting closet. The reasoning is that although the company's share price may drop a bit more than it otherwise would, it will drop far less than if the company announced each bit of bad news separately (word spy at http://www.wordspy.com/words/kitchen-sink.asp.; emphasis added).

As is obvious from the proposed semantic and cognitive motivation of the CV, what has lead the coiner to produce and use the lexical item are not morphosyntactic rules but a

[7] The aspectual distinction and conjugation classes in Bulgarian are disregarded here as they apply equally to simplex, derived and compound verbs. The inticate mutual determinacy (if there is one) between derivation and conjugation classes fall outside the focus of the present argument.

[8] This term is used as defined by Goossens (2003). It is intended to indicate that metaphor and metonymy often work together in a symbiosis to back up human creativity in language use and understanding.

naming need to satisfy an instance of complex conceptual configuring based on metaphtonymic elaborations. There is metonymic mapping between *kitchen* and *destroying everything* (present in the initial idiomatic creation) and a set of metaphoric extensions tying up (the divulging of) *bad financial results* and *physical destruction*.

CVs in Group A create new individuated types of activities, i.e. names of socio-culturally significant activities (usually pragmatically primed); those in Group B receive such readings only on the basis of metonymic and metaphoric extensions,which leads to enhanced semantic exocentricity (e.g. the development of *spoon-feed* from a manner of feeding CV into one with an extended negatively marked sense of giving too much information or help to someone). The latter start off as more explicit descriptions of already named activities and end up as lexical items that have undergone semantic change.

Epistemologically speaking, Group B CVs can be interpreted as instances of subclassification manner verbs naming subtypes of the event named by a conceptually associated simplex verb, while Group A CVs name newly conceptualized events for which a conceptually associated simplex verb may not even exist and which represent emergent conceptual configuring.

The second criterion, semantic exocentricity, is understood to constitute a cline. The choice of the criterion is motivated by the recognition that "…exocentricity, even though it constitutes a sort of « anomaly » in language design, is nevertheless one of the defining properties of compounding phenomena" (Scalise and Guevara 2006: 185). Scalise and Guevara recognize the centrality of exocentricity in compounding and provide a definition not restricted to the "type of" requirement traditionally associated with endocentric compounds.

> Exocentricity is an « anomaly » in language design in the following sense: describing a construction as exocentric means acknowledging that we cannot account for all the information conveyed by it (ibid.)

To further specify our use of exocentricity as a classificatory criterion, we need to emphasize that we adopt Scalise et al.'s third dimension of exocentricity – semantic exocentricity, "in which the semantic class denoted by the compound cannot be predicted from the semantic class of their constituents" (Scalise et al. n.d.: 59-60).

According to this criterion CVs can be classified into a type whose semantics preserves the semantic predictability of the whole on the basis of the frames of the constituents (e.g. *bottle-feed*, *kick-start*, *ръкомахам, ръкопляскам, водоснабдявам*, etc.), while the second necessarily involve some kind of metaphtonymic transfer (e.g. *fast-talk*, *piggyback*, *ръкополагам, главоблъскам се*, etc.). Both Group A and Group B have the general potential to have semantically exocentric members. There are no restrictions concerning the metaphtonymy susceptibility of CVs. Only very general pragmatic constraints regulate the metaphtonymic elaborations of CVs. Furthermore exocentricity might be associated only with particular senses of a CV. It might even be the case that exocentricity is directly dependent of lexicalization, but such a claim is in need of further corroboration, which is beyond the scope of the present argument. The cline of semantic exocentricity supplements the two basic classes A and B.

semantically endocentric		semantically exocentric
bottle-feed		*spoon-feed*
tumble-dry		*freeze-dry*
fox-hunt	*head-hunt*	*job-hunt*
ръкомахам	*ръкопляскам*	*ръкополагам*

Figure 3: Niche-internal exocentric cline

The classification better captures the specificity of CVs in English and Bulgarian because it is able to accommodate the conceptual and lexico-semantic similarities among verbs, whose classification into the standard subordinate, attributive and coordinate types will lead to their classification in different categories (e.g. *stir-fry* and *deep-fry*, *drip-dry* and *rough-dry*). Considering the fact that most novel CVs arise as analogical constructions based on established exemplars, it is plausible to surmise that users rely more on lexico-semantic criteria than on morphosyntactic ones as every user relates to meaning but few are able to carry out the detailed linguistic analysis which underlies most classificatory schemes.

To replace classifications based on the lexical categoriality of compound constituents, or ones based on meaning generalizations stemming from the particular derivational processes, we can postulate construction-specific subschemas with unified semantics shaped by frame-based conceptual configuring:

Table 4: Three types of CVs

value-foregrounding (manner) (sub-classification)	*deep-fry, sun-dry, soft-land; злословя, благославям, ръкопляскам*
spatial scenario (sub-classification)	*overrate, undertake, outmaneuver; *задминавам*
metaphtonymic (super-classification)	*ear-mark, name-ambush, sandbag; самозабравям се, главоболя се, зловиди ми се*

Thus the CV lexicon of English and Bulgarian can be classified in a more comprehensive system which includes the general construction schema and its three specifications.

1. General schema of the constructional idiom: [X Y]V, where X stands for a compound-internal acategorial constituent and Y also stands for a compound-internal acategorial constituent, which are coerced into a relational concept by the constructional schema that sanctions them.

2. Group A: Super-classification CVs [X Y]V where the overall meaning of the CV involves the use of an initial situational interpretation which provides the onomasiological motivation for the CV: *to sandbag, to deadpan, to background; ръкополагам, словоблудствам* (slovobludstvam, 'word-abuse', "speak nonsense"), *боготворя* (bogotvorya, 'god-create', "worship").

3. Group B: Sub-classification CVs: [X V]V, which is subdivided into two groups on the basis of the lower-level schemas:

B_1. One level removed schema: [x v]v where v is not categorially specified within the construction, but is homonymous with a simplex verb, e.g. *to spray-paint, to spoon-feed, to headhunt*; *гласоподавам, водоснабдявам*

B_2. [SPATIAL SPECIFIER V]V – Marchand's genuine CVs: *to outnumber, to undergo, to oversee; подминавам*

Correlated with the 3 lower-level schemas, 3 different patterns of configuring can be postulated (which, for lack of space, will not be discussed in detail here, but see Bagasheva forthcoming):

I. A [X Y]V – configuring where the generic space is a newly emergent one in which the attribute values to be projected from the two input frames are selected in keeping with the graded salience hypothesis (Giora 1997, 2002; Huang 2009) and following pragmatically driven mapping principles (e.g. *railroad, piggyback, moonlight; ръкополагам, боготворя*).

II. B_1. [X F,D,P]V – configuring of the two frames where the first frame fills an available slot in the second one and foregrounds it, creating a new perspectivized profile of the second frame (e.g. *deep-fry, водоснабдявам*).

III. B_2. [SPATIAL SPECIFIER V]V – frame configuring where the two frames merge and the spatial specifier frame augments the second frame by embedding it in a spatial scenario

via the Location and Event branches of the Event Structure Metaphoric System. It has been argued that the prefixes in Bulgarian realizing this onomasiological need have core spatial meanings (Radeva 2007) and extended comparative meanings (Grozdanova 2005). They presuppose the existence of an implicit conceptual norm against which their specific lexical semantics can be appropriately interpreted. In view of grammaticalization theory (Heine et al. 1991) and the overall model of graded schematization (Langacker 2008) it might be claimed that the only difference between English spatial scenario "genuine" CVs and Bulgarian prefixed verbs is their degree of schematization (e.g. *overindulge, downgrade, преодолявам, преобръщам*). This leads to the following classificatory scheme:

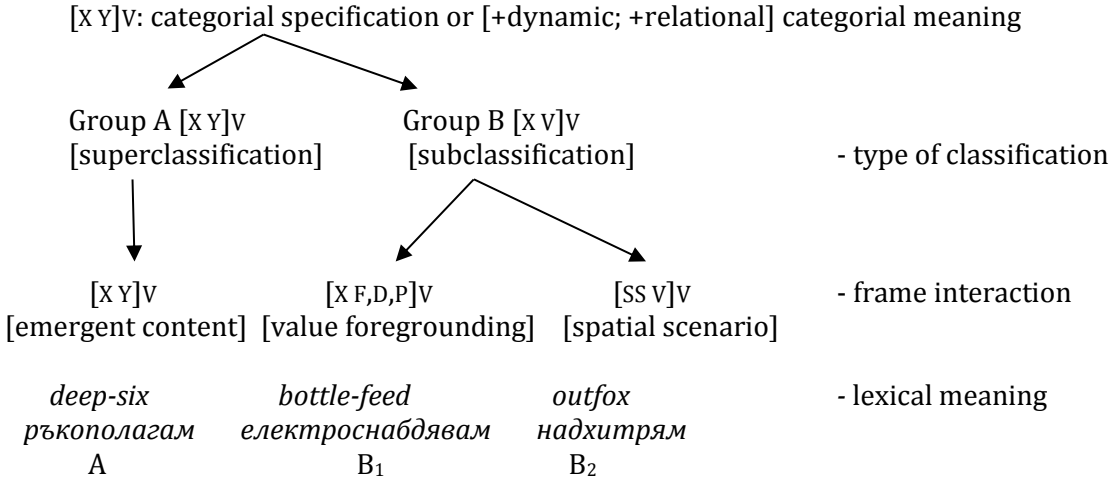


Figure 4: A diagrammatic representation of the semantic space of the CV lexicon

The proposed classificatory system correlates with three identifiable mechanisms of frame interaction which uniformly apply to the members of the respective three subclasses. The elaboration and illustration of these claims will be the next step in this ongoing piece of research.

6. Concluding remarks

It appears from the potentially unlimited permissibility of the constructional idiom[9] that there are no morphosyntactic constraints that might preclude the lexical specification of the subschemas in any significant way. This is in keeping with the rising analytical tendencies in English and the severely undermined rigidity of its part-of-speech system. In Bulgarian the necessary aspectual marking of a V and the inflecting-fusional obligatory marking of part-of-speech membership (Manova 2005, Nitsolova 2008) seem to impose stricter constraints and to reduce the analogical potential of single CVs. In English, by contrast, the only constraint to be satisfied by CVs is the Conventional Frame constraint, as defined by Goldberg (2010: 50), "[f]or a situation to be labeled by a [*compound*] verb, the situation or experience may be hypothetical or historical and need not be directly experienced, but it is necessary that the situation or experience evoke a cultural unit that is familiar and relevant to those who use the word." There are no grammatical constraints in English for the appearance of novel CVs, arising via approximation (Rainer 2005: 23) to established local schemas.

[9] The claim is based on the analysis of a self-compiled corpus of 460 CVs in English and 66 CVs in Bulgarian. The data for the corpus have been extracted from from CoCA, BNC, OALD 7th edition, OED on CD-Rom 2nd edition, dicitionary.com, urbandictionary.com, word spy.com, BulNC, the DBL and various research articles, works of fiction, occasional movies and TV series. For details see Bagasheva (forthcoming).

Studying compounds as usage events instantiating a constructional idiom helps better explain in a unified manner the properties of CVs. In both English and Bulgarian, CVs are consistently categorially and morphologically right-headed, but display wide variability in terms of semantic exocentrcity. The natural further step in this line of research will be to see how and if the proposed classificatory scheme for CVs can accommodate the properties of CVs in other (preferably typologically distinct) languages. It should also be supplemented by detailed analyses of the semantic mechanisms and patterns involved in the three types of configuring.

Primary sources

BNC: British National Corpus at: http://www.natcorp.ox.ac.uk/. BNC Webmaster *(bnc-queries@rt.oucs.ox.ac.uk)* 2009-01-26. OUCS. *© 2010, University of Oxford*.
BulNC: Bulgarian National Corpus at: http://search.dcl.bas.bg/. Department of Computational Linguistics, Bulgarian Academy of Sciences.
CoCA: Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 425 million words, 1990-present. Available online at http://www.americancorpus.org.
DBL: Dictionary of the Bulgarian Language [Речник на българския език. Кристалина Чолакова (гл. ред.) (1997-2008). София: Издателство на Българската академия на науките.
Dictionary.com Unabridged (Based on the Random House Dictionary, © Random House, Inc. 2011) at http://dictionary.reference.com/ (last accessed 30 May 2011)
OALD: Oxford Advanced Learner's Dictionary, (2005) 7th edition; Oxford: Oxford University Press.
OED 2 Oxford English Dictionary (2009) 2nd edition, Version 4.0, Simpson, J. (ed.). Oxford: Oxford University Press.
urbandicitionary.com at http://www.urbandictionary.com/ (accessed on a regular basis August 2009 – February 2011)
word spy at: http://www.wordspy.com/ (accessed on a regular basis December 2009 – February 2011)

References

Adams, Valerie 2001. *Complex words in English*. Harlow: Longman.
Anderson, Stephen 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.
Bagasheva, Alexandra 2011a. Beyond Dichotomies: On the Nature and Classification of Compound Verbs in English. *Linguistica L1: Les Frontières Internes et Externes de la Morphologie*, 2011 (1): 333-357.
Bagasheva, Alexandra 2011b. Compound verbs in English revisited. *Bucharest Working Papers in Linguistics* XIII 2011 (1): 125-150.
Bagasheva, Alexandra 2011c. Why do we stone-wall about compound verbs? *Contrastive Linguistics* 2011/2: 49-71.
Bagasheva, Alexandra forthcoming. *Reflections on Compound Verbs and Compounding* Sofia University "St. Kliment Ohridski" Publishing House.
Bauer, Laurie 1983. *English word-formation.* Cambridge: Cambridge University Press.
Bauer, Laurie 2001. Compounding. In M. Haspelmath, E. König, W. Oesterreicher, W. Raible (eds.) *Language Typology and Language Universals.* 695-707. Berlin – New York: Walter de Gruyter.
Bauer, Laurie 2006. Compound. In Keith Brown (ed.) 2nd edition. *The encyclopedia of language and linguistics*. 719-726. Oxford: Elsevier, vol 2.
Bauer, Laurie 2008. Dvandva. *Word Structure* 1: 1-20.
Bauer, Laurie n.d. English Exocentric Compounds (the English version of the paper published in French 'Les composés exocentriques de l'anglais' in Dany Amiot (ed), *La composition dans uneperspective typologique*. 2008. Arras: Artois Presses Université, pp. 35-47). 20 September 2010. http://www.victoria.ac.nz/lals/staff/Publications/Bauer-EnglishExocentricCompounds.pdf.
Booij, Geert 2005. Compounding and derivation: evidence for construction morphology. In Dressler, Wolfgang, D., O. Pfeiffer and F. Rainer (eds), *Morphology and its demarcations*. 109–132. Amsterdam and Philadelphia: John Benjamins.
Clark, Eve and Herbert Clark 1979. When Nouns Surface as Verbs. *Language* 55 (4): 767-811.
Fabb, Nigel 2001. Compounding. In Spencer, A. and A. Zwicky (eds.) *The Handbook of Morphology*. 1-13. Blackwell Online Reference: DOI: 10.1111/b.9780631226949.2001.00006.x.
Farrell, Patrick 2001. Functional Shift as Category Underspecification. *English Language and Linguistics* 5: 109-130.
Giora, Rachel 1997. Understanding figurative and literal language: the Graded Salience Hypothesis. *Cognitive Linguistics* 7(1): 183-206.
Giora, Rachel 2002. Optimal innovation and pleasure. In Stock, O., Strapparva, C. and A. Nijholt (eds.) *Processing of The April Fools' Day Workshop on Computational Humour. April 2002*. 11-28. Trento, Itlay: ITC-itst, 2002.

Goldberg, Adele 2010. Verbs, constructions and semantic frames**.** In M. Rappaport Hovav, and I. Sichel, (eds.) *Syntax, Lexical Semantics, and Event Structure*. 39-58. Oxford: Oxford University Press.

Goossens, Louis 2003. Metaphtonymy: The interaction of metaphor and metonymy in expressions for linguistic action. In Dirven, R. and R. Pörings, (Eds.) *Metaphor and Metonymy in Comparison and Contrast*. 349-377. Berlin: Mouton de Gruyter.

Grozdanova, Lilyana 2005. In the Network of Language. Sofia University Publishing House "St. Kliment Ohridski".

Guevara, Emiliano and Sergio Scalise 2004. V-Compounding in Dutch and Italian. In Cuadernos de Lingüística del Instituto Universitario Ortega y Gasset Madrid – XI: 1-29.

Haspelmath, Martin 2002. *Understanding Morphology*. Hodder Education Publishers.

Heine, Bernd, Ulrike Claudi and Friederike Hünnemeyer 1991. *Grammaticalization: A conceptual framework.* Chicago: University of Chicago Press.

Hopper, Paul and Sandra Thompson 2004. The discourse basis for lexical categories in Universal Grammar. In Aarts, Bas, David Denison, Evelien Keizer and Gergana Popova (eds.). *Fuzzy Grammar: A Reader*. 247-291. Oxford: Oxford University Press. (originally published 1984 in Language 60: 703-52).

Huang, Mimi 2009. Solving the riddle of metaphor: A salience-based model for metaphorical interpretation in a discourse context. In Evans, V. and Pourcel S. (eds.) *New Directions in Cognitive Linguistics.* 107-126.
Amsterdam/Philadelphia: John Benjamins Publishing Company.

Huddleston, Rodney and Geoffrey Pullum (2002) *The Cambridge Grammar of the English Language*. Cambrdige: Cambridge University Press.

Hüning, Matthias 2009. Semantic niches and analogy in word formation: Evidence from contrastive linguistics. *Languages in contrast* 9/2: 183-201.

Langacker, Ronald 2008. *Cognitive Grammar: A basic introduction*. Oxford: Oxford University Press.

Laudanna, Alessandro 2002. Levels of processing for nouns and verbs: some issues and controversies. *Rivista di Linguistica* 14.1: 3-8.

Lieber, Rochelle 2004. *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Lieber, Rochelle 2009. IE, Germanic: English. In Lieber, R. and Štekauer, P. (eds.) *The Oxford Handbook of Compounding*. 357-370. Oxford: Oxford University Press.

Manova, Svetla 2005. Towards a Theory of Conversion in Slavic: Evidence from Bulgarian, Russian and Serbo-Croatian. *The Slavic and East European Language Resource Center, Issue 6*, at http://seelrc.org/glossos/

Marchand, Hans 1969. *The Categories and Types of Present-Day English Word-Formation*. 2nd edition. München, Beck.

McGregor, William 2002. *Verb Classification in Australian Languages*. New York: Mouton de Gruyter.

Mithun, Marianne 1984. The Evolution of Noun Incorporation. *Language* 60/4: 847-894.

Nagano, Akiko 2007. Marchand's analysis of back-formation revisited: back-formation as a type of conversion. *Acta Linguistica Hungarica* 54/1: 33–72.

Nitsolova, Rusalina 2008. *Българска граматика. Морфология*. [Balgarska gramatika. Morfologiya]. Bulgarian Grammar. Morphology. Sofia: Universitetsko izdatelstvo "Sv. Kliment Ohridski".

Radeva, Vasilka 2007. *В света на думите*. [V sveta na dumite]. *In the world of words*. Sofia: Universitetsko izdatelstvo "Sv. Kliment Ohridski".

Rainer, Franz 2005. Typology, Diachrony, and Universals of Semantic Change in Word-Formation: A Romanist's Look at the Polysemy of Agent Nouns. In Booij, Geert, Emiliano Guevara, Angela Ralli, Salvatore Sgroi and Sergio Scalise (eds.), *Morphology and Linguistic Typology, On-lineProceedings of the Fourth Mediterranean Morphology Meeting (MMM4) Catania 21-23 September 2003*, University of Bologna, 2005. http://morbo.lingue.unibo.it/mmm/

Rappaport Hovav, Malka, Edit Doron and Ivy Sichel 2010. Introduction. In Rappaport Hovav, Malka, Edit Doron and Ivy Sichel (eds.) *Syntax, Lexical Semantics, and Event Structure*. 1-18. Oxford: Oxford University Press.

Renner, Vincent 2008. On the Semantics of English coordinate compounds. *English Studies* 89/5: 606-613.

Rijkhoff, Jan 2002. Verbs and nouns from a cross-linguistic perspective. *Rivista di Linguistica* 14.1: 115-147.
Scalise, Sergio, Antonio Fábregas and Francesca Forza. (n.d.) Exocentricity in Compounding, at http://morbo.lingue.unibo.it/user-documents/scalise/exocentricityincompounding.pdf (accessed 17 October 2009).
Scalise, Sergio and Emiliano Guevara 2006. Exocentric Compounding in a Typological Framework. *Lingue and Lingaggio* 2: 185-206.
Scalise, Sergio and Antonietta Bisetto 2009. The classification of compounds. In Lieber, Rochelle and Pavol Štekauer (eds.) *The Oxford Handbook of Compounding*. 34-53. Oxford: Oxford University Press.
Scalise, Sergio and Antonio Fábregas. 2010. The head in compounding. In Scalise, Sergio and Irene Vogel (eds.) *Cross-disciplinary Issues in Compounding*. 109-125. Amsterdam/Philadelphia: John Benjamins Publishing House.
Shibatani, Masayoshi 1990. *The Languages of Japan*. Cambridge: Cambridge University Press.
Slobin, D. (2003) Language and Thought Online: Cognitive Consequences of Linguistic Relativity. In Gentner, Dedre and Susan Goldin-Meadow (eds) *Language in Mind: Advances in the Study of Language and Thought*. 157-192. Cambridge, Massachusetts: The MIT Press.
Štekauer, Pavol 2005. Onomasiological approach to word formation. In Štekauer, Pavol and Rochelle Lieber (eds.) *Handbook of Word-Formation*. 207-232. Springer.
Ten Hacken, Pius 2010. Synthetic and Exocentric Compounds in a Parallel Architecture. In Linguistische Berichte Sonderheft 17: 233-251.
Vogel, Petra 2000. Grammaticalization and Part of Speech Systems. In Vogel, Petra and Bernard Comrie, (eds.) *Approaches to the Typology of Word Classes*. 259-284. New York: Mouton de Gruyter.
Wald, Benji and Lawrence Besserman 2002. The emergence of the verb-verb compound in twentieth century English and twentieth century linguistics. In Minkova, Donka and Robert Stockwell (eds.) *Studies in the history of the English language: a millennial perspective*. 417-447. Berlin: Mouton de Gruyter.

The English noun-noun construct: a morphological and syntactic object

Melanie J. Bell
Anglia Ruskin University
melanie.bell@anglia.ac.uk

0. Abstract

In English, debates about the boundary between morphology and syntax have often focussed on combinations of two nouns (NNs) in which the first modifies the second, e.g. 'coffee cup' or 'silk shirt'. These are widely regarded as falling into two classes: morphological compound nouns on the one hand, and syntactic noun phrases on the other (e.g. Payne & Huddleston 2002: 449). But although various tests have been proposed by which the two types might potentially be distinguished from one another, the results of these tests do not converge, and their reliability has been questioned (e.g. Bauer 1998). Either the distinction between morphological and syntactic types is purely a matter of definition, and depends on the test that is chosen, or there is actually no clear boundary.

This paper investigates one of the most widely accepted tests for phrasal status, namely the possibility of independent modification: in cases where either noun can be adjectivally modified independently of the other, proponents of this test take the NN to be syntactic (e.g. Payne & Huddleston ibid., Lieber & Štekauer 2009: 11). But what properties of a particular NN determine whether such modification is possible? The present study attempts to answer this question by examining a large database of constructions of the form [AdjN]N or N[AdjN] randomly extracted from the British National Corpus. It is shown that, except for a small number of cases where the second noun is appositive, the possibility of modifying the first noun (N1) independently of the second (N2) depends on whether there is a combination of adjective plus N1 that is lexicalised, institutionalised, or at least more frequent than the NN itself. In the case of N2, the possibility of independent modification seems to depend largely on the nature of N1. In nearly all N[AdjN] constructions, the first noun is either a proper noun, a noun with an incorporated numeral such as 'one way', a material noun such as 'silk', or another noun that occurs very frequently in attributive position compared to its occurrence elsewhere.

Overall, the results suggest that, for a given NN, the probability of either noun being modified independently of the other depends largely on the relative frequencies with which the two nouns occur in various positions. If we accept that such modification distinguishes between objects usually viewed as compounds and those usually viewed as phrases, then a possible conclusion is that the distinction between morphological and syntactic objects is itself based on relative frequencies: as such, it is gradient and usage-based, and the lack of a clear boundary is expected.

1. Introduction

In considering the role of morphology in the grammars of natural languages, a basic question concerns the demarcation of morphological versus syntactic objects: which of the patterns found in a language should we regard as products of morphology, and which as products of syntax? If morphology deals with the structure of words, and syntax deals with the combination of words into larger linguistic units, then the proper criteria for the demarcation of morphological versus syntactic objects are those that distinguish words from phrases. Dixon and Aikhenvald (2002: 35) suggest that, cross-linguistically, the grammatical word can be defined as 'a number of grammatical elements which always occur together, in a fixed order and have a conventionalised coherence and meaning'. Other well-known criteria include the notion of the word as a 'minimal free form', in the sense of being the smallest unit than can constitute an utterance (Bloomfield 1935: 178), as well as the tendency for word formation to be non-recursive (Matthews 1991: 213). However, despite the fact that the issue has received considerable attention from generations of linguists, and despite a general recognition that words tend to have the

characteristics described above, no criterion has been found by which words can be categorically and universally identified (Matthews 1991: 215, 2002: 271). It seems that, if categorical criteria exist at all, they must be language specific.

In the case of English, attempts to find criteria for distinguishing words from phrases have often focussed on noun-noun combinations (henceforth NNs), such as *teacup* or *silk shirt*: these are widely regarded as falling into two categories, with some being analysed as morphological compound nouns while others are seen as syntactic nominals, i.e. as noun phrases without determiners. Various criteria have been proposed for distinguishing these two purported classes, including the tendencies of English compounds to have a single main stress on the first element, to be written as single or hyphenated words, to involve different semantic relations from those of phrases, and to obey principles of 'lexical integrity', which is effectively a way of saying that the constituents of compounds tend to have different distributional properties from those of their phrasal counterparts. However, as demonstrated by Bauer (1998), these criteria 'do not draw a clear and consistent distinction between a syntactic and a morphological object': not only is it debatable whether some of the criteria do distinguish words from phrases, but even the more robust criteria produce conflicting results. In other words, the categorisation of a given NN as phrase or compound depends on the test used, and choosing any test as criterial therefore amounts to defining the word or phrase as a construction that passes or fails that particular test. This has led some authors, e.g. Bauer (ibid.), Olsen (2000) and Bell (2005, 2011), to argue that the English NN in fact represents a single but variable class of construction. But this does not solve the problem of distinguishing morphological from syntactic objects: at least in the case of English, the difficulty of finding a reliable language-specific definition of the notion WORD is comparable to the difficulty of finding one that applies cross-linguistically.

A logical possibility is that the difficulties of finding a clear demarcation between syntax and morphology arise because no rigid demarcation actually exists. On this view, words and phrases can be regarded as prototypes rather than categories, and the lack of a clear boundary is therefore no longer a problem. The prototypical word is both a grammatical and a phonological unit, and has the characteristics that tend to be associated with words cross-linguistically. It consists of a string of sounds that can stand alone as an utterance but cannot be broken into smaller strings that can also stand as utterances. It does not include any recurring grammatical elements and stands in a paradigmatic implicational relationship to other word forms: in other words, the form-meaning correspondences of a known paradigm can be applied to a newly learnt or newly formed word (Matthews 1991: 187). The prototypical syntactic construction, on the other hand, not only can stand alone as an utterance but also includes smaller parts that can stand alone. Prototypically, these smaller parts have the same meaning whether they occur as free forms or as part of the construction, and the meaning of the construction itself is transparently composed of the meaning of the parts in conjunction with their arrangement relative to one another. Between these two extremes, however, are a range of possibilities: in complex words, for example, just one part of the string might be able to stand alone, e.g. *create* in *creative*. Furthermore, the possibility of a string functioning as an utterance is itself a gradient notion: some elements, for example, can occur as free forms only in a directly contrastive context, e.g. '*re*' in answer to '*Did you say revise or devise?*' (Matthews 1991: 210-11). On the other hand, substrings might occur elsewhere as free forms but not with the same form-meaning correspondence as they have within the construction. This is the case, for example, with idioms and with English complex tenses, where the auxiliary does not have the same sense as the corresponding lexical verb.

The English NN has some properties of the prototypical word and some properties of the prototypical phrase. In inflected languages, compounds pattern like complex words because, with the exception of the final one, the elements are uninflected, and could therefore not form utterances. In an uninflected language like English,

however, both elements of a compound have the same form as possible utterances (unless one is phonologically reduced or constitutes a combining form, as in neoclassical compounds). Furthermore, compounding is a recursive process, and may even involve repetition of the same constituents, as in *table tennis table*. In these ways, the English NN is syntactic. On the other hand, NNs have the same distribution as simplex nouns, and the possibility of higher level compounding is a reflection of this fact. Furthermore, they stand in paradigmatic relations to one another, such that each combination can be recognised as belonging to two 'morphological families', one consisting of all combinations that share a first constituent, the other consisting of all combinations with the same second constituent (de Jong 2002). The psychological reality of these families is demonstrated by their predictive significance in e.g. word naming and visual lexical decision studies (Baayen et al. 2010) as well as their involvement in the placement of prosodic prominence (e.g. Plag 2010). In these ways, the English NN is morphological. The inevitable conclusion is that the English noun-noun construct, rather than being in some cases syntactic and in other cases morphological, in most cases shares properties of both. This is similar to the conclusion reached by Giegerich (2005) about combinations of noun plus associative adjective in English. However, whereas Giegerich (ibid.) concluded that syntax and morphology represent two 'overlapping modules', a more radical but equally plausible conclusion would seem to be that they do not constitute discrete modules at all.

Both cross-linguistic and English-specific evidence, then, suggests that the distinction between morphological and syntactic objects is not categorical, but gradient. Nevertheless, tests have been proposed by which two such purported classes might be recognised, and in some cases these tests enjoy wide currency: it is therefore interesting to explore what the results of such tests might reflect. This is the purpose of the present paper; not to debate the proper criteria for the demarcation of morphological versus syntactic objects, but rather to investigate in more detail an already widely-accepted criterion, namely the supposed inseparability of the parts of a complex word. If there is no absolute distinction between words and phrases, then what do tests for this property actually measure?

In English, the inseparability of the word, or 'lexical integrity', has been operationalised in terms of several distributional tests. This paper investigates one such test as it applies to NNs, namely whether the constituent nouns can be modified independently of one another, to produce constructions of the form [AN]N or N[AN], where A is an adjective. In a two-class analysis of NNs, the assumption is that those where independent modification is possible are phrases, whereas those that do not permit such modification are compounds. With a gradient analysis, we might hypothesise that those NNs that allow independent modification have a relatively high degree of syntactic as opposed to morphological character. But if there is no categorical distinction between words and phrases, then what does it mean to say that, by this criterion, one NN is more or less phrase-like than another? The paper has two objectives. The first objective is to provide a detailed corpus-based description of the types of [AN]N and N[AN] constructions that occur, and hence of the circumstances under which independent modification of NN constituents arises. The second objective is to test a particular hypothesis, namely that the extent to which such modification is possible depends at least partly on the identity of the first noun. This is an extension of the suggestion by Plag (2003: 160) and Bell (2005) that certain classes of noun in first position tend to give a phrasal or phrase-like flavour to NNs in which they occur.

It will be helpful at the outset to state a number of assumptions on which the methodology and argumentation of this paper are based. Firstly, I assume that linguistic classes, in so far as they can be recognised, are based on distribution: that is to say, that strings with the same distribution in a language, relative to specific lexical items, can broadly be regarded as belonging to the same class. Secondly, I assume that nominal compounding in English is recursive. This means that compound nouns have the same

distribution as simplex nouns of comparable length and of the same type: singular, plural or mass. One of the implications of this is that any compound noun can itself function as the first or second constituent of a larger compound. It follows that, if some compound nouns have the form AN, then such AN strings can also occupy either the first or second position in a longer compound noun. Thirdly, I assume that lexicalised, institutionalised or locally lexicalised phrases can function as first elements in English compound nouns, giving rise to so-called 'phrasal compounds'. The implication of this is that any established or locally lexicalised AN combination, whether or not it constitutes a compound in itself, could function as the first element in a compound.

The rest of the paper is organised as follows: section 2 gives the background to the study and explains in more detail the reasons for choosing modification as the test-bed for this paper; section 3 describes the methodology of the corpus study; section 4 discusses the results regarding modification of the first noun in NN; section 5 discusses the results regarding modification of the second noun; and finally, section 6 is the conclusion.

2. Background

2.1. The morphosyntactic status of the English noun-noun

In all Germanic languages except Present-day English, compounds are distinguished from phrases on the basis of inflectional criteria: in a phrase, all constituents are inflected, whereas in a compound, only the final constituent is inflected (cf. Bell 2011: 138-143). By this criterion, all NN constructs in these languages are analysed as compounds, since the first noun is never inflected. If this criterion were applied to Present-day English, however, we would have to conclude that only gradable adjectives can occur as pre-head modifiers in English noun phrases, because the paucity of inflectional morphology in the language means that this is the only class that can be productively inflected in that position. The usual analysis, however, is that both gradable and non-gradable adjectives can syntactically pre-modify English nouns, and therefore that no inflectional criterion distinguishes English compounds from phrases. In other words, unless they are gradable adjectives, the pre-head modifiers in English noun phrases are not inflected, and are therefore morphologically indistinguishable from the first elements of compounds. This opens up the possibility for NNs to be analysed as phrases, both consciously by scholars of the language, and unconsciously by speakers: the fact that the first noun is not inflected no longer means that it cannot be a syntactic modifier.

If NNs are to be analysed as constituting two groups, syntactic nominals and morphological compounds, then the question arises as to how these two classes can be identified: given a particular English NN, how do we know whether it is a phrase or a compound? In the absence of any inflectional criterion, Anglicists have sought other methods by which to make this distinction.

It has sometimes been suggested, for example by Marchand (1969: 23), that phrasal and compound NNs can be distinguished in English on the basis of phonological stress: those with main stress on the first noun (N1) are taken to be compounds, whereas those perceived to have main stress on the second noun (N2) are analysed as phrases. However, stress is a notoriously unreliable criterion, not least because the stress assigned to a particular NN often varies between speakers and even for the same speaker on different occasions. In fact, a significant body of work conducted over the last six years, e.g. Plag *et al.* (2007, 2008), Bell (2012), Bell & Plag (2012), has shown that stress assignment in English NNs can be modelled probabilistically on the basis of semantic and frequency-based variables, and does not appear to reflect any underlying morphosyntactic difference.

Other authors, e.g. Biber *et al*. (1999: 590), have used an orthographic criterion to divide NNs into two groups: those written as two words are regarded as phrases,

whereas those written as single or hyphenated words are regarded as compounds. However, English orthography is notoriously variable in this respect, and it is not uncommon to find the same NN written, quite acceptably, in all three forms. Such a variable characteristic seems most unlikely to reflect any underlying structural difference: one would have to assume that the same NNs are for some speakers compounds, for other speakers phrases, and for some speakers, phrases on some occasions but at other times compounds. Nevertheless, it would be untrue to suggest that the orthography is completely random, and some tendencies can certainly be recognised. For example, combinations involving shorter constituents are on the whole more likely to be written as single words than those involving longer constituents (Bauer 1998). It has also been shown that orthography correlates with frequency (e.g. Plag et al. 2007, 2008): compounds usually written as one word tend to have higher frequencies than compounds usually written as two separate words. But neither of these correlations necessarily reflects any underlying morphosyntactic difference between the spaced and concatenated types.

Yet another criterion proposed in the literature is semantic: Jespersen (1942: 137), for example, suggests that 'we have a compound if the meaning of the whole cannot be logically deduced from the meaning of the elements separately'. But again, this is a poor basis for a categorical distinction, since semantic transparency is a gradient notion, and the degree to which the meaning of a particular NN can be deduced from the meaning of its parts will reflect the extent to which it has become semantically lexicalised. Furthermore, as argued by various authors, notably Di Sciullo & Williams (1987), semantic opacity indicates that a string needs to be listed in the lexicon but does not tell us anything about its status as a word or phrase: complex words can be fully transparent, e.g. *manageable*, *achievable* etc., and fully inflecting phrases can be semantically opaque, e.g. *kick/kicked/kicking the bucket*, meaning DIE.

In fact, as argued by Payne & Huddleston (2002: 451), if phrases and compounds cannot be distinguished on the basis of inflectional morphology, then it is appropriate to turn to syntactic criteria: considerations of semantics, phonology and orthography are secondary since the purported distinction is between morphological and syntactic constructions. Morphosyntactic arguments for the supposed phrasal status of NNs are usually based on the principle of lexical integrity, the notion that 'syntactic processes can manipulate members of lexical categories ('words') but not their morphological elements' (Giegerich 2009: 183). On this premise, data such as those in (1) and (2), from Payne & Huddleston (2002: 449), and (3), from Quirk et al. (1985: 1332), are taken to indicate that the NNs in italics are phrases, since their constituents can undergo, respectively, modification, coordination and substitution by the proform *one*, all of which are assumed to be purely syntactic operations.

(1) (a) *London colleges*
(b) [south *London*] *colleges*
(c) *London* [theological *colleges*]

(2) (a) various [*London* and Oxford] *colleges*
(b) various *London* [schools and *colleges*]
(c) [two *London* and four Oxford] *colleges*

(3) She wants an *oak table* but I'd prefer a teak one.

However, the assumptions that these operations constitute tests for syntactic constituency are by no means universally accepted, particularly in the cases of coordination and proform substitution.

The use of coordination as a test for compound status rests on the assumption that only whole words rather than parts of words can be coordinated. However, this assumption can easily be shown to be false. In English, neo-classical combining forms,

which are not found as independent words in the language, and some prefixes, which Spencer (2005: 82) describes as 'loosely bound', can be freely coordinated. Examples are given in (4). These examples are taken from the British National Corpus, version 3 (BNC XML Edition), and the references in brackets give the three-letter text identifier and sentence number in the corpus. Unless stated otherwise, all subsequent examples in this paper come from the same corpus.

(4) (a) …all dealing with a mixture of **over and underconstrained** problems. (FE6 1086)

(b) … one of the best known officers of the **pre and postwar** RAF… (J56 276)

(c) …the problems of **inter and intraobserver** variation… (HWS 4916)

The exact circumstances under which such coordination can occur are not well understood, although Plag (2003: 84) suggests that both sub-lexical coordination and gapping in English can be explained in terms of prosody. On the basis of data similar to (4), he concludes that English affixes and compound constituents can be coordinated provided they do not form a single prosodic word with the element that is omitted.

In some other languages, notably Turkish, there is a phenomenon known as suspended affixation (Lewis 1967: 35), in which two related words are coordinated but only the second is inflected, the inflection taking scope over both coordinated words. Kabak (2007) argues that the extent to which this is possible reflects the tightness of the morphological cohesion between the stem and potentially suspended affix: the tighter the bonding, the less likely is suspension to occur. Furthermore, Kabak (ibid.) shows that the degree of morphological cohesion is correlated with the degree of phonological cohesion. Suspension is less likely with more tightly phonologically bound affixes. This is reminiscent of Plag's (2003) analysis for English, and suggests that coordination may be at least partly phonologically conditioned. Booij (1985) reaches a similar conclusion for German and Dutch.

Another possibly relevant factor in the availability or otherwise of sub-lexical coordination may be the semantic relation between the potentially coordinated constituents: in particular, whether they exhibit 'natural coordination' or 'accidental coordination'. Natural coordination is the coordination of terms that 'express semantically closely associated concepts' (Wälchli 2005: 1), such as kinship terms, e.g. *brother and sister*, body parts, e.g. *fingers and toes*, and cutlery, e.g. *knife and fork*. However, the notion 'closely associated' may be culturally dependent, so that what constitutes natural coordination may vary from language to language, and may even be determined by the local context (Dalrymple & Nikolaeva 2006). In some languages, e.g. Finnish, Tundra Nenets, Russian and Kurdish, coordinated singular nouns fall into two categories: some such coordinate structures are modified by adjectives with plural inflection while others are modified by singular adjectives. The distinction between the two types depends on whether the coordinated nouns represent natural or accidental coordination. In cases of natural coordination, a plural adjective is required, but in cases of accidental coordination, the adjective must be singular. Dalrymple & Nikolaeva (ibid.) argue that the structure involving natural coordination is more like a compound or even a simple plural noun than it is like a phrase. If these constructions are word-like, then the coordinated units within them are sublexical, and this is further evidence that the possibility or otherwise of coordination may be a poor test by which to distinguish word level units from phrasal ones. In general, it seems that coordination as a test for syntactic constituency is at best unreliable, and therefore not a good basis on which to draw conclusions about the morphosyntactic status of English NNs.

The second morphosyntactic test that arises from the notion of lexical integrity concerns anaphora: according to the lexical integrity principle, sub-lexical constituents should not be available to participate in anaphoric operations. In the case of compound nouns, this means that the constituent nouns should neither be able to act as antecedents

for the pro-form *one* nor be individually replaceable by it. Accepting this assumption, Quirk, Greenbaum, Leech & Svartvik (1985: 1332) and Giegerich (2005, 2009) regard pro-*one* substitution as a purely syntactic operation, and therefore criterial for phrasehood. They have used this idea as a test for the status of English NNs: in cases where it seems possible for either the head or the modifying noun to act as the antecedent for *one*, they conclude that the structure is a syntactic phrase.

However, the idea that proforms cannot refer to parts of words is by no means uncontested. For example, Lieber (1992: 130) quotes the sentences in (5) from Postal (1969):

(5) (a) Harry was looking for a bookrack, but he only found racks for very small *ones*.

(b) Max's argument was pointless, but Pete's did have *one*.

Although Postal (ibid.) judges these sentences to be unacceptable and therefore uses them to argue that words are 'anaphoric islands', Lieber (ibid.) finds that they are acceptable for at least some speakers, whom she regards as having a 'permissive' dialect. She sees this as evidence that sublexical constituents can function as antecedents for anaphoric *one*, since in both cases the proform refers to just part of a previously mentioned word. In (5a) *ones* refers to *books*, and in (5b) *one* refers to *point*, but neither *books* nor *point* occur as freestanding words in the given contexts. In fact, contra Postal (ibid.), it is now generally recognised that there is no absolute constraint against outbound anaphora, that is to say against sublexical constituents functioning as anaphoric antecedents. Rather, as demonstrated by Ward et al. (1991), the extent to which it is felicitous depends on 'a number of morphosyntactic, semantic, and pragmatic factors that increase the accessibility of discourse entities' (ibid.: 468).

A number of authors, e.g. Culicover & Jackendoff (2005:137) and Keizer (2011), have also questioned the reliability of pro-form substitution as a test for constituency at phrase level. Keizer (ibid.) bases her argument on many attested examples from the BNC and Corpus of American English (COCA) (Davis 2008-). For example, she cites the sentence reproduced here as (6):

(6) So Paul had a **big** blue **felt marker** for days and a red **one** for nights. (HGU 451)

In this example, the pro-form *one* can refer either to *felt marker* or to *big felt marker.* The first case is to be expected if *one* substitutes for strings generally regarded as syntactic constituents*.* But if *one* is substituting for *big felt marker,* then it is representing a discontinuous string which would not, in most theories, be regarded as a structural unit. In the light of such examples, Culicover and Jackendoff (ibid.: 138) conclude that the interpretation of *one* is 'simply the interpretation of the antecedent NP less the material in contrast'. It seems that, just as coordination is at least partly governed by phonology, so anaphora falls largely within the domain of pragmatics, and is therefore likely to be an unreliable criterion by which to establish the morphosyntactic status of noun-noun constructions.

Generally speaking, there is a lack of consensus about the reliability of coordination and *one* substitution as criteria for distinguishing NN compounds from putative NN phrases. Giegerich (2009: 193), for example, regards coordination as unreliable but places more faith in the pro-form test. Payne & Huddleston (2002: 449), on the other hand, include the coordination test but not the pro-form one. Overall, however, most authors who discuss the issue agree that one of the most reliable criteria is the possibility or otherwise of independently modifying the constituent nouns. The argument is that, because of lexical integrity, the constituents of a compound cannot be modified independently of one another, whereas those of a phrase can be. Payne & Huddleston (ibid.) give the example in (1), reproduced for convenience in (7). They argue that, because each element of *London colleges* can be independently modified by an adjective, *London colleges* itself must be a syntactic phrase.

(7) (a) *London colleges*
(b) [south *London*] *colleges*
(c) *London* [theological *colleges*]

Even here, however, there is not complete agreement. For example, Lieber & Štekauer (2009: 11), regard independent modification of N2 as one of the most reliable criteria for phrasal status, because it involves separation of the two nouns. Payne & Huddleston (ibid.), on the other hand, regard separate modification of N2 as the least useful of the tests they list, on the grounds that such modification might be blocked by constraints on the ordering of pre-nominal constituents in the noun-phrase. I understand this to mean that they take the possibility of modification of N2 as a sufficient but not necessary criterion for phrasehood: if independent modification of N2 is possible, then NN is a phrase, but if modification is not possible, NN is not necessarily a compound. Despite this reservation, amongst those who analyse English NNs as falling into two classes, modification is the most widely agreed-upon criterion for distinguishing phrasal and compound types. And for this reason, it will be used as the basis for the empirical investigation reported in this paper.

2.2. Modification by adjectives

Constructions of the form [AN]N, such as *south London Colleges*, and N[AN], such as *London theological colleges*, occur quite frequently in Present-day English. However the existence of these constructions does not necessarily tell us anything about the status of the corresponding NN constructs. NN compounding is recursive in Present-day English, and any compound noun can therefore occupy either the N1 or the N2 slot in a larger compound. Furthermore, most accounts of English compounding agree that compound nouns can have the form AN, as in *blackbird*, for example. So in cases where the AN component of [AN]N or N[AN] can be analysed as a compound, then the whole construction can also be regarded as a compound, e.g. *blackbird nest* or *mother blackbird*. An alternative analysis of (7) is therefore that s*outh London* and *theological colleges* are themselves compounds, so that (7b) is simply a compound of s*outh London* and *colleges* and (7c) is a compound of *London* and *theological colleges*. In other words, if the AN constituent can be analysed as a compound, the existence of [AN]N and N[AN] says nothing about the status of the corresponding NN, and the existence of NN is not a necessary precondition for the formation of the larger constructions.

Spencer (2003) has argued that Present-day English does not in fact have productive AN compounding and that all apparent AN compounds are actually lexicalised phrases. However, even if we accept this view, it does not preclude the compound analysis of the larger constructions, at least in the case of [AN]N, since Present-day English has a well-recognised type of compound in which a noun is modified by a phrase. These so-called phrasal compounds have been discussed by a number of authors, including Lieber (1992, 2009: 363), Bresnan & Mchombo (1995), Lieber & Scalise (2006) and Giegerich (2009: 197). Examples are given in (8): in each case, a noun is pre-modified by a string that has the form of a phrase.

(8) (a) … spraying insecticide … is not feasible in hilly, **hard to reach areas**. (J2N 63)
(b) Where this wins … is in its upfront and **in your face approach** … (HWX 1375)
(c) … **'come to bed' plea** by girl, 15. (CS1 1542)

These constructions are usually regarded as compounds because stress can fall on the phrasal element rather than the head noun, the head noun is usually not amenable to further modification, and the construction overall does not conform to any of the syntactic patterns recognised for English phrases.

The exact circumstances under which such compounds can be formed are not well understood. Bresnan and Mchombo (ibid.) suggest that the modifying phrase has to be either lexicalised or have the status of a quotation: in other words, to have some degree of institutionalisation. Lieber (1992, 2009: 363), on the other hand, concludes that the modifying phrase need not be lexicalised. If this is correct, and fully syntactic phrases can occupy the modifier slot in English compound nouns, then all [AN]N constructions can be regarded as compounds, whether or not the AN constituent is lexicalised or institutionalised. However, Lieber's (1992, 2009) analysis is not universally accepted (cf. Giegerich 2009: 197), and indeed some of her examples do not appear to support her own argument. For example, Lieber (2009: 364) gives the example of the compound *out-of-context nature.* She argues that the phrasal constituent is not lexicalised, since it is completely semantically transparent, and that nor does it have the status of a quotation. However, *out-of-context* is listed in the OED online: indeed it is listed as an adjective, with *out-of-context summations* and *out-of-context bites* given as examples. This suggests that, while the phrase might not be lexicalised in the sense of being semantically opaque, it is nevertheless institutionalised, in the sense of being an established lexical item (Bauer 1983: 48).

It may be that the phrases in phrasal compounds are best understood as naming units. As defined by Lipka *et al*. (2004), these are lexemes, linguistic expressions or proper names that are used to name extralinguistic entities, as opposed to describing them. Naming units are candidates for lexicalisation and may be lexicalised to varying degrees. Other authors have expressed this idea in terms of the concept of 'nameworthiness' (Downing 1977). For example, Dahl (2004: 252) notes that, cross-linguistically, in most cases of constructions with incorporated nouns, the entity denoted must 'have a status that in principle makes it possible to invent a name for it.' In other words, these constructions denote 'unitary concepts' (ibid.). It seems likely that some such constraint might also apply to the modifying phrases in English phrasal compounds. If so, this would be consistent with the observation made by a number of authors, e.g. Booij (2009) and Spencer (2011), that compounds are essentially names. In other words, if the modifying phrases in phrasal compounds are naming units, then it is no surprise that they can combine with other naming units (i.e. nouns) to form larger naming units (i.e. compounds).

The classification of a phrase as a 'naming unit' does not presuppose that it has been diachronically lexicalised, or entered the lexicon of the population at large. Names can be coined, and phrases coined as names undergo a temporary, 'local lexicalisation'[1], perhaps for the duration of a single conversation or even a single utterance. However, because of the practical difficulty of determining whether items in a corpus are locally lexicalised, this study will focus on established items in the first instance.

What are the implications of phrasal compounds for an analysis of [AN]N constructions? If we accept Lieber's (2009) analysis that syntactic strings can indeed act as modifiers in compound nouns, then the existence of [AN]N says nothing about the status of the corresponding NN, since in all cases it will be possible to analyse [AN]N as a compound. If, on the other hand, only lexicalised or institutionalised phrases can occupy the modifier position in phrasal compounds, then it might be possible to distinguish a set of phrasal [AN]Ns from the compound class. In the compound type, there should be evidence that the AN component is itself a lexicalised or institutionalised expression, whereas in the phrasal type, the AN combination will have the characteristics of a productively formed syntactic phrase. In particular, we might expect that the adjective in

[1] A story that demonstrates the possibility of local lexicalisation concerns a passenger flight on which one person makes repeated trips to the toilet and therefore becomes known to the other passengers as the 'john man'. I was given this example by John Hawkins, but do not know the original source.

a phrasal AN constituent would be able freely to undergo further modification by adverbs, a point which is further developed in the next section.

For the purposes of this paper, I will adopt the more conservative assumption that only lexicalised or institutionalised phrases can function as modifiers in compound nouns. This means that constructions of the form [AN]N and N[AN], in which the AN constituent is not lexicalised or institutionalised, can be regarded as NN constructions in which respectively N1 or N2 has been modified independently of the other constituent. However, this leaves us with the problem of deciding which AN combinations should be regarded as lexicalised or institutionalised, and this will be discussed in the next section.

2.3. Modification by adverbs

In order to investigate the conditions under which either constituent of a NN can be adjectivally modified independently of the other, we want to find examples of such modification from a corpus. This entails finding constructions of the form [AN]N and N[AN], in which the AN constituent is not lexicalised or institutionalised, since these cases do not necessarily involve modification of a NN combination. To make this clearer, consider for example *right hand man*: this is clearly a direct combination of the AN constituent, *right hand*, with the second noun, *man*, rather than a NN, *hand man*, in which N1 has been independently modified. Assuming that we can extract a set of [AN]N and N[AN] types from a corpus, how can we subsequently eliminate those that do not represent modification of a NN?

As pointed out by Croft (2001: 13) the 'basic method of empirical grammatical analysis' is distributional analysis: the categorisation of linguistic forms on the basis of their distribution relative to other forms in a corpus of language. Distributional analysis leads to the recognition of 'substitution classes', sets of speech fragments that can occupy the same position in a longer string (cf. Harris 1946). Since, by definition, compounding words produces longer words, rather than structures of a different class, any AN or NN that is a compound noun will have the same distribution as a non-compound noun of comparable length and the same type (singular, plural or mass). On the other hand, if the AN or NN constitutes a different kind of construction, a NP or nominal, then we would expect a different distribution.

Croft (ibid.) expresses concern that the way in which distributional analysis is sometimes applied in linguistics can lead to a logical inconsistency, when '[c]onstructions are used to define categories ... then the categories are taken as primitive elements of syntactic representation and are used to define constructions' (ibid. p. 45). However, distributional analysis applied in the traditional way, avoids such circularity by defining substitution classes in terms of the possible occurrence of strings in specific positions in particular lexically defined utterances.

Because it can occur as a free-standing utterance, for example, in answer to a question such as *What are you looking for?* the English noun phrase (NP) is taken here to be a primitive unit and a suitable starting point for an analysis. We can then define the English nominal as a string that can fill the blank space in (9), where the square brackets enclose a NP:

(9) [the __]$_{NP}$

The English simple noun is taken to be the smallest unit that can occupy the nominal slot. However, the same space can clearly be filled by longer strings, including those with the form AN, irrespective of whether they are compound nouns, lexicalised phrases or syntactic nominals. Remember that, in order for adjectival modification to be viable as a test for the morphosyntactic status of NNs, it would first be necessary to distinguish between morphological and syntactic ANs. The question is whether there is any distributional criterion that might be used to distinguish two such classes.

In fact, there is a relevant distributional fact, identified by Jesperson (1914: 318-9) and restated by Payne, Huddleston & Pullum (2010). It is this: with the possible exception of *almost*, English adverbs do not act as modifiers of a following noun. Since they can premodify any other part of speech, this fact is sufficient to distinguish nouns from other classes. Thus, if an AN can be premodified by an adverb, it cannot be a compound noun; on the other hand, if adverbial premodification is inadmissible, then the AN does have the distribution of a noun, irrespective of whether it is analysed as a compound or a lexicalised phrase. Adverbs have the great advantage, for present purposes, of being one of the most morphologically distinct groups in English. We can therefore define two distributional patterns as shown in (10 a, b):

(10) (a) [the (+ly) __]$_{NP}$
(b) [the *+ly __]$_{NP}$

In (10a, b) the symbol '+' represents a string expressing a property concept, so that '+ly' is what we might designate an English morphological adverb, or more accurately, since not all morphological adverbs can occur in this position, an English prenominal morphological adverb. The brackets around '+ly' in (10a) indicate that the adverb element is optional, and the asterisk before '+ly' in (10b) indicates that, in this construction, an adverb is impossible. The space in (10a) can therefore be filled by AN strings that would normally be regarded as syntactic nominals, whereas the space in (10b) can be filled by AN strings that might broadly be classed as lexical, i.e. by compound nouns or lexicalised phrases. The reason for using a generic adverb, rather than *very*, is to allow for the possibility of non-gradable adjectives occurring as modifiers in syntactic nominals. The frame in (11) would select only a sub-class of syntactic types, namely those in which the adjective is gradable:

(11) [the very __]$_{NP}$

In looking for examples in which one element of a NN has been independently modified, we therefore want to find [AN]N and N[AN] strings in which the AN constituent fits the pattern in (10a) rather than the pattern in (10b). One way to do this is to start by eliminating those types in which, for various reasons, the adjective is clearly not amenable to adverbial modification.

There are at least three classes usually labelled AN which are well-known not to accept adverbial modification and therefore to have the distribution of nouns. The first, exemplified in (12a), consists of expressions that are semantically lexicalised as defined by Bauer (2001: 45). This is to say that the meaning of the whole cannot be compositionally derived from the meanings of the constituents: a *hard disk* is not simply a disk that is hard. As a result of this loss of semantic transparency, the adjective cannot be adverbially modified without a change in meaning. Thus, although (12b) is fine, (12c) would be infelicitous:

(12) (a) A **hard disk** is required with about two Mb free space (HAC 499)...
(b) drilling holes into **extremely hard masonry** (A16 1050)
(c) *An **extremely hard disk** is required with 2Mb free space

Another group of AN strings in which the first element resists modification are proper names, exemplified by (13a). Because these are 'expressions which have been conventionally adopted as the name of a particular entity' (Payne & Huddleston 2002: 515), they have a semantic unity similar to that described in the preceding paragraph for lexicalised types. Thus (13b) occurs, but (13c) could not, except perhaps in some ironic or humorous sense:

(13) (a) It's Mark ... from the **Daily Telegraph**. (HYE 161)

(b) ... these **apparently daily murders** ... (HHV 2133)
(c) *It's Mark from the **apparently Daily Telegraph**

According to Lipka *et al*. (2004: 11), proper names 'prototypically demonstrate the naming function of words'. Thus, both because of their status as naming units, and because of their unavailability to adverbial modification, [AN]N and N[AN] strings in which the AN constituent is a proper name can be analysed as compounds of N and AN.

A third class that we will find labelled AN, but which is well known to resist adverbial modification, consists of those types in which the first element belongs to the set of words variously called nominal (Levi 1978, Sadler & Arnold 1994: 210), relational (Beard 1991: 195–229) or associative (Giegerich 2005, Payne & Huddleston 2002) adjectives. In these cases:

'*the property expressed by the adjective does not apply literally to the denotation of the head nominal, but rather to some entity associated with it*' (Payne & Huddleston 2002: 556)

For example, in *medical bag*, the adjective *medical* does not describe the bag in the way that *big* or *old* might; rather it describes activity associated with items the bag is intended to hold. Combinations of 'associative adjective' plus noun are therefore to some extent semantically opaque: the exact nature of the 'associated with' relation usually depends upon encyclopaedic knowledge, so that the meaning of the whole is not simply compositional (cf. Levi 1978: 52). Other notable semantic features of associative ANs are that the associative adjectives usually have fairly restricted distributions in terms of the nouns they can modify (Giegerich 2005: 576) and, in some cases, associative adjectives have virtually synonymous nouns with which they are interchangeable. Levi (1978: 38), for example, gives the examples shown in (14).

(14)	(a)		(b)	
		atom bomb		atomic bomb
		mother role		maternal role
		industry output		industrial output
		ocean life		marine life
		language skills		linguistic skills
		city parks		urban parks

In each case, the NN combination in (14a) is virtually synonymous with the corresponding AN combination in (14b). Overall, the semantic properties of associative adjectives lead Giegerich (2005: 576) to conclude that associative ANs and certain NN compounds 'are virtually identical in many aspects of their behaviour'.

In terms of distribution, associative adjectives only occur in attributive position: they are therefore effectively bound forms, since they can only occur with a following noun. And because associative adjectives are not amenable to adverbial modification, combinations of associative adjective plus noun have the distribution of nouns, a fact well-recognised across a range of theoretical approaches, e.g. Levi 1978: 66–74, Alexiadou et al. 2007: 219. However, the same strings that function as 'associative adjectives' can in many cases be modified by adverbs when they occur in different contexts. So (15b) is possible, even though (15c) is not:

(15) (a) Shelley ran to the jeep for the **medical bag**. (JYA 1850)
(b) ... a decision for the doctor to make, based upon **wholly medical criteria**. (ASK 1232)
(c) *Shelley ran to the jeep for the **wholly medical bag**.

In (15b), the adjective *medical* has a slightly different meaning, representing a property of the concept expressed by the noun, rather than something associated with it. With this

type of meaning, adjectives are classed as ascriptive (e.g. Pullum & Huddleston 2002: 557) or qualitative (e.g. Beard 1991). A particular adjectival string may have both associative and ascriptive uses or, to put it another way, associative adjectives can have ascriptive homophones.

Overall then, associative adjectives represent a so-called 'mismatched category': while they have the semantics and distribution of nouns, they have the morphological form of adjectives (e.g. Giegerich 2005: 576). Because associative adjectives cannot be adverbially modified, and because they are also semantically similar to nouns, combinations of associative adjective plus noun fit the pattern in (10b) rather than (10a). This means that, in cases where the adjective is associative, [AN]N and N[AN] constructions can be analysed as compounds, and such constructions therefore provide no evidence about the morphosyntactic status of the corresponding NN.

2.4. Summary

In the absence of inflectional or other reliable criteria for compoundhood, scholars have used the existence of [AN]N and N[AN] constructions to argue for the phrasal status of some NNs in Present-day English. This argument rests on the assumption that these constructions are themselves phrasal, but in fact they can also be analysed as compounds in which one constituent is itself a compound or lexicalised phrase. If they are compounds, they provide no information about the corresponding NN, which may not even have been coined.

Assuming for the moment that two classes might exist, I have argued that two types of evidence can help to distinguish [AN]N and N[AN] compounds from putative syntactic strings with the same surface form. Firstly, if the AN constituent is lexicalised or institutionalised, then a compound analysis cannot be ruled out. Secondly, if the AN constituent is not lexicalised or institutionalised, then the possibility arises than the larger construction is phrasal, or at least phrase-like. If such phrasal constructions exist, we would expect that the adjectives within them are amenable to adverbial modification. In this case, it ought to be possible to find constructions of the form [AdvAN]N and N[AdvAN] in which the AdvAN constituents are not themselves lexicalised or institutionalised.

If such phrase-like types are found, then a further question arises as to the circumstances under which they can be formed. Plag (2003: 160) suggests it could be argued 'that there are only some restricted classes of nouns whose members are allowed to act as syntactic modifiers of nouns'. In constructions that satisfy the modification criterion for phrasal status, it will therefore be instructive to look at the nouns that occur in N1 position, to see whether they do indeed fall into particular categories. However, if categorisation does not fully explain the patterns found, then other, more gradient explanations will need to be sought.

In the corpus study that follows, a large number of constructions with the form [AN]N, [AdvAN]N, N[AN] or N[AdvAN] are extracted from the British National Corpus and tested against the criteria described above. It is shown that, while in the great majority of cases these constructions have the distribution of compound nouns, there are some that have properties associated with phrases. In cases that seem to satisfy the criteria for phrasal status, a further analysis is made of the N1 constituents. As predicted by Plag (ibid.), certain classes of N1 are particularly frequent in these constructions. But over and above this, it is shown that, in the phrase-like constructions, even those N1s that do not fall into any easily-recognisable category in fact have the distribution of frequent modifiers, and that this appears to be a gradient rather than categorical property.

3. Method

3.1. Creating a database

The British National Corpus (BNC) was chosen because it is a large and well-balanced corpus, consisting of approximately 90 million words of written and 10 million words of spoken English, across a wide range of text types. Furthermore, because the corpus is grammatically annotated, it can be searched for strings matching particular parts of speech. For this study, the corpus was queried using BNCweb (Hoffmann & Evert 2006), a web-based interface that allows searches by part of speech and will return up to 5000 hits for any query. Most of the searches in this study yielded more than 5000 hits, and so the random selection option included with the interface was used to select 5000 tokens at random from the total number found.

Four initial searches were conducted: firstly for strings labelled ANN, secondly for NAN, thirdly for AdvAN(A)N, where (A) indicates an optional adjective, and finally for NAdvAN. The first two searches were conducted twice, giving a total of 10,000 tokens of each type, from which duplicate hits were removed before further processing. The third and fourth queries were run once each, yielding 5000 and 2622 tokens respectively. All tokens were then inspected in their corpus context to find those in which the A and Adv constituents selectively modified either N1 or N2, in other words, those with the following semantic structures: [AN]N, N[AN], [AdvAN]N and N[AdvAN]. The tokens with these structures formed the database for the study.

3.2. Correlates of lexicalisation and institutionalisation

Each item in the database was tested to find out whether the construction as a whole, and/or the AN constituent within it, could be regarded as lexicalised or institutionalised, i.e. as an established lexical item. This is not the same as establishing morphosyntactic status. Remember that both words and phrases can have opaque semantics and may therefore need to be listed, and that strings with the form of phrases can function as first constituents in English compound nouns, especially (though not exclusively) when those strings represent established lexemes. Various measures can be used to operationalise the notions of lexicalisation and institutionalisation, and these measures fall into the broad categories of listedness, orthography and frequency. The study presented in this paper uses each of these types, as described in the following paragraphs.

To operationalise semantic opacity and institutionalisation one can use dictionaries. In general it can be assumed that dictionaries, for economic and practical reasons, tend to list those complex words that are in some sense idiosyncratic; for example, have a meaning that is not inferable from the constituent parts, or a particular meaning amongst several theoretically possible ones. Hence, 'listedness', that is to say, having an entry in a dictionary, can be taken as an indication that a NN is likely to be institutionalised or semantically opaque. Of course, dictionaries also list some fully transparent complex words, but one can assume that among those NNs listed in a dictionary there is a large proportion of non-transparent ones. In any case, what is at issue in this study is not simply whether a particular AN pair is semantically lexicalised, but rather the broader question as to whether it is an established combination.

OED Online, the online version of the Oxford English Dictionary, was checked for each type in the database, as well as their AN components. There is considerable variation in how NNs are listed in the dictionary, sometimes as full entries and sometimes under one of their constituents, usually the modifier. Because of this inconsistency, any hit from the main electronic search page (i.e. not including the full text) was counted as an entry. Nevertheless, there were marked discrepancies in the results: for example, *general hospital* is listed, whereas *depressed fracture* is not, even though it is non-compositional, meaning a fracture of the skull. To compensate for this, all items that did not have an entry in OED Online were then checked in the on-line encyclopaedia, Wikipedia. If the search term was found to be the title of a page in

Wikipedia, even if that page redirected the search elsewhere, the term was counted as listed. The only exception was made for entries referring to proper names. For example, a search for *younger brother* brings up a page in Wikipedia, but the page is about a pop group with that name: such results were not counted as a listing.

A second correlate of lexicalisation is orthography. It is generally assumed that lexicalisation strongly correlates with frequency (e.g. Lipka 1994: 2165) and it has also been shown, for NN constructs, that frequency correlates with orthography (e.g. Plag et al. 2007, 2008). NNs written as one word tend to have higher frequencies than those written as two separate words, which is a strong indication that orthographically concatenated NNs are more lexicalised on average than non-concatenated ones. The assumption is made here that the same is true for AN combinations. The related assumption, that concatenated or hyphenated orthography is used by speakers when they perceive the constituent parts as constituting a single conceptual unit, seems equally true for ANs as for NNs.

The query syntax used in this study returned only strings written with spaces between all the words, and so all AN constituents of items in the database were known to occur in the BNC with spaced orthography. However, many strings that occur spaced can also be found hyphenated or even concatenated. In this study, I therefore used two frequency-based variables as measures of lexicalisation. These were AN frequency and 'spelling ratio', which is the number of non-spaced tokens of a string found in a corpus divided by the number of spaced tokens, i.e. the ratio of non-spaced frequency to spaced frequency (Bell & Plag 2013). For all non-listed AN types in the database, which were also not names, lemmatised frequencies were taken from the whole 100 million words of the BNC using the BYU-BNC interface (Davies 2004-). Separate frequencies were obtained for AN written as two words (spaced) and one word (non-spaced), with hyphenated tokens included in the non-spaced count. AN frequency was then defined as the sum of the two different spelling frequencies, while spelling ratio was the non-spaced frequency divided by the spaced frequency.

Finally, all items and their constituents were checked to see whether they were proper names. These included not only prototypical personal and place names, but also names of companies, products, other organisations, events and so on. Occasionally it was unclear whether a writer/speaker intended a particular string as a name. In such cases, capitalisation was taken as an indication of intended name status and the wider context was also taken into consideration.

3.3. Morphological family sizes

In order to test the hypothesis that certain nouns in N1 position are more likely than others to occur in phrase-like NNs, I calculated the family size ratio for a subset of N1 constituents in the database. The family size ratio is the positional family size of a constituent divided by its reverse family size, that is to say the number of NN types in which it occurs in the same position, N1 or N2, divided by the number of NN types in which it occurs in the other position. Each NN has a left constituent family and a right constituent family. The group of NNs in which the same constituent occurs in the same position constitute the positional family for that constituent. For example, the left positional constituent family of *country house* would include NNs such as *country club, country music, countryside*, while the right positional constituent family would feature NNs like *town house, jailhouse,* and *summer house*. The reverse family of house would include, for example, *house mate, house mice, house coat, house boat* and so on. Likewise, the reverse family of *country* would consist of *mother country, gulf country, farm country, donor country* and so on. Positional and reverse family sizes can be extracted from the corpus by searching for NN strings in which particular constituents occur in one position

or the other: Bell (2012) demonstrates that these raw measures are highly correlated with accurate family sizes.

Family size ratio was calculated for those nouns that occurred as N1 in potentially phrasal N[AN] constructions, where the noun did not fall into any category proposed in the literature to favour a syntactic analysis. The hypothesis to be tested is that the first nouns in phrase-like NNs are likely to be those that typically occur as modifiers, and therefore have some adjective-like properties in terms of distribution. This leads to the prediction that these nouns will have higher family size ratios in N1 position than a random selection of nouns in that position, i.e. they will modify a wide range of nouns, but will themselves be modified by relatively few.

3.4. Procedure

Each construction type, [AN]N, N[AN], [AdvAN]N and N[AdvAN], was analysed separately. In each case, every example of the construction in the database, as well as the AN constituents within them, were checked for listedness using OED online and Wikipedia, as described above. Secondly, all items and their constituents were checked to see whether they were proper names. Thirdly, for those types where neither the whole construction nor AN was listed or a name, a check was made to ascertain whether the adjective could be classed as associative. As described in section 2.3, constructions with any of these three patterns can be regarded as compounds, and therefore do not constitute evidence that the corresponding NN is phrase-like.

Finally, for each construction type, the remaining tokens were inspected for obvious patterns, such as those suggested by Plag (2003: 160). Residual tokens that did not fall into any easily-recognisable category were then analysed using various quantitative measures. The details of these analyses vary slightly for each construction type, and for clarity of exposition they are therefore described together with the presentation of the results in the following sections.

4. Modification of N1: results and discussion

4.1. [Adjective Noun] Noun

The search for strings labelled adjective noun noun yielded 555,122 hits in 3932 different texts, a frequency of 5646 instances per million words. Of these, 8002 randomly selected tokens were inspected in context. In 1260 cases, about 16% of the total, the adjective selectively modified the first noun, so that the string had the structure [AN]N. This suggests that such constructions occur about 0.16 x 5646 = 903 times per million words, or approximately once in every thousand words. The 1260 tokens represented 1190 types of [AN]N and 831 types of AN. The distribution of various patterns within the [AN]N types is shown in Table 1.

Table 1: Distribution of patterns in [AN]N

AN and/or ANN listed	992	83.4%
not listed, but AN and/or ANN is proper name	64	5.4%
other evidence that AN forms a unit (e.g. NN not possible with same meaning)	59	5.0%
none of the above, but A is associative	7	0.6%
sub-total	1122	94.3%
none of the above, but N2 is appositive	8	0.7%
none of the above, but AN forms a 'compound adjective'	41	3.4%
residual types	19	1.6%
total	1190	100.0%

4.1.1. AN has the distribution of a noun

Perhaps the most striking result is that in the great majority of types (83.4%) the AN constituent and/or the construction as a whole is listed. Examples are given in (16), where *right hand* has an entry in OED Online while *floating rate* and *cold weather payments* have entries in Wikipedia:

(16) (a) ... Jason's trusted **right hand man** ... (ADR 1529)
(b) ... in the case of **floating rate loans** ... (CBU 4668)
(c) The hon. Member ... referred to **cold weather payments**. (HHX 10274)

In about a quarter of cases (26%), either [AN]N, AN, or both, were names. These largely overlapped with the listed types. Examples of these three types are given in (17a-c) respectively:

(17) (a) ... proceedings on the **Criminal Justice Bill** ... (EEC 689)
(b) ... Gallacher applied for the **Labour Party whip** ... (JXM 1099)
(c) ... a 27–7 victory over the **Green Bay Packers** ... (CEP 3163)

In a further 59 cases (5%), there was other evidence that the AN constituent formed a lexical unit even though it was not a proper name and was not listed. In some cases, it was clear that the whole construction was a compound of AN plus N, rather than a modified NN, because the corresponding NN would not have had the same meaning as in the overall construction. For example, in (18a), *adjustable back rucksacks* is clearly a compound of *adjustable back* and *rucksacks*, because *back rucksacks* could not be taken to mean 'rucksacks with backs', whereas *adjustable back rucksacks* means 'rucksacks with adjustable backs'. Similarly, in (18b), *angled mouth pipette* has to be a compound of *angled mouth* and *pipette*, since it means 'a pipette with an angled mouth', and *mouth pipette* would not be taken to mean 'a pipette with a mouth'. Both these cases arise because the first noun represents an integral part of the entity represented by the second noun, *back rucksacks* is not possible with the same meaning as *adjustable back rucksacks*

because all rucksacks have backs, and similarly *mouth pipette* is not possible with the same meaning as *angled mouth pipette* because all pipettes have mouths.
Other types of evidence that the AN forms a lexical unit are exemplified in (18c) and (18d). In several cases, exemplified in (18c), the AN or the whole ANN were found to have institutionalised meanings in particular fields, as evidenced by their frequent reduction to acronyms. For example, *slow transit constipation* has an institutionalised meaning in medicine and is often abbreviated to *STC*. In other cases, including (18d), the AN constituent had a locally lexicalised meaning, defined in the context. For example, *subterranean passage view* occurs in a text about the Loch Ness Monster, where the possibility has been discussed that monsters might enter the lake through a subterranean passage.

(18) (a) **Adjustable back rucksacks** (G2S 1703)
(b) Use an **angled mouth pipette** to localize a few embryos … (EV6 690)
(c) ... patients complaining of **slow transit constipation** ... (HU4 782)
(d) The **subterranean passage view** offers a plausible account ... (AMT 714)

Another seven items, which did not conform to any of the patterns discussed so far in this section, were nevertheless found to involve associative adjectives, and examples of these are given in (19).

(19) (a) ... the **environmental labelling issue** ... (ALV 82)
(b) ... his yard ran an efficient **... marine supplies business** ... (CCW 204)
(c) ... the AL1-BL is a compact **dual arm loader** ... (HST 87)

Altogether, the aforementioned types constituted 94.3% of the [AN]N types found. In other words, in the overwhelming majority of cases of [AN]N, the AN constituent is a lexical unit and the whole construction is therefore best understood as a compound of AN and N. Only 68 [AN]N types in the data did not show any obvious evidence that the AN constituent had the distribution of a noun: these 68 types can therefore be regarded as potentially phrasal.

4.1.2. Appositive modifiers

Amongst the potentially phrasal [AN]N types, eight had an appositional structure exemplified in (20):

(20) (a) … guerrillas rained rockets ... on the **Afghan capital Kabul.** (CH6 264)
(b) Malcolm was followed by his **red-haired brother William** … (EF2 342)
(c) … please drop that **stupid name Aotearoa** … (HH3 9030)

It seems that this appositional construction may provide evidence that the corresponding NN is phrasal, and we might therefore expect to find constructions of the form [AdvAN]N with this same kind of appositional relation between the constituents.

4.1.3. AN has the distribution of an adjective

Amongst the remaining types, there was a striking dichotomy according to whether or not the AN constituent occurred with non-spaced orthography in the corpus. In the 41 cases where the AN constituent was found concatenated, hyphenated or both, the constituent seemed to represent a lexicalised unit of the kind regarded by Bauer (1983: 211) as compound adjectives. These are exemplified in (21).

(21) (a) … I took master and mistress their **early morning tea** … (A0D 2397)

(b) Return the coupon today for a free **full colour brochure** … (CFS 2270)
(c) In general, only **high priority cases** are able to gain a place. (G1C 1369)

As Bauer notes, these same combinations of adjective + noun, when used in non attributive position, are straightforward noun phrases. However, when used in attributive position, they assume the characteristics of adjectives. In the present study, evidence that they are lexicalised items comes from the frequency data.

If these AN pairs are indeed established units, despite not being listed and not being names, then we would predict that their spelling ratio (the proportion of times they occur in the corpus with hyphenated or concatenated orthography) would be significantly higher than the equivalent measures for AN combinations in general. Furthermore, if they have the distribution of adjectives, we might expect them to occur as attributive modifiers more often that the average AN combination and also to modify a larger number of nouns, i.e. to have a significantly larger positional family size.

To test these hypotheses, 200 AN combinations were selected at random from the BNC, using the BNCweb (CQP-Edition) interface (Hoffmann & Evert 2006). This interface allows searches based exclusively on part of speech, so it was possible to search for strings of the form AN. Starting at the beginning of the list, the random selection of AN strings produced by the interface was inspected to find hits in which the AN pair constituted a premodified noun. Sampling ended when 200 such types had been found. Total frequency, spelling ratio, frequency in attributive position and positional family size were then calculated both for the 41 potential 'compound adjectives' and for the 200 randomly selected AN combinations. The proportion of times each AN occurred as a modifier (ATTRIBUTIVE PROPORTION) was calculated as its frequency in attributive position divided by its total frequency. Spelling ratio, attributive proportion and positional family size were all logarithmatised in order to guard against the effects of extreme values and produce sufficiently normal distributions to use parametric tests of significance. In the case of family size, 1.0 was added to the raw values before taking logs, since some of the randomly selected AN combinations did not occur in attributive position and it was necessary to avoid taking the logarithm of zero.

All three hypotheses were shown to be correct. Compared to the random sample of ANs, the 'compound adjectives' had significantly higher attributive proportions (t=7.6849, p=3.089e-12), significantly higher spelling ratios (t=5.1099, p=2.293e-06) and significantly higher positional family sizes (t=10.5929, p= 8.296e-15). The frequency data therefore strongly support the view that there is a group of AN collocations that function as compound adjectives, as suggested by Bauer (1983: 211), Jespersen (1914: 320) and Arnaud (2008). Just as with the appositional types, we would therefore predict that [AdvAN]N constructions will be found in which the AN constituent forms one of these compound adjectives. From a qualitative point of view it is striking that certain adjectives seem to occur particularly frequently in this compound adjective construction. Of the 41 [AN]N types in which the AN can be regarded as a compound adjective, 16 of them involve the adjective *high*, five times in the context of *high quality* and three times in the context of *high risk*. The combination *early morning* occurs in six of the 41 types.

4.1.3. Relative frequencies of AN and NN

The remaining 19 types of [AN]N are shown in table (2):

Table 2: 'Residual' types in [AN]N

	[[AN]N]
1	bare brick Kitchen (CJT 786)
2	dependent employee status (FEW 1272)
3	dilute solution data (HRG 730)

4	Fatal crash trial (K5D 5875)
5	green code business (JS7 266)
6	green strategy document (JP7 1052)
7	high debt country (K59 1334)
8	high sulphate period (HU4 4034)
9	Lateral adjustment lever (KA3 26)
10	local office monitoring (HCL 656)
11	major offence categories (G1J 60)
12	minimum competencies legislation (FAM 106)
13	minimum fill mark (HWF 3418)
14	Multiple licence packs (CR0 51)
15	natural leather couches (C8S 1237)
16	online catalogue terminals (GXE 280)
17	personal questionnaire approach (HJ0 10088)
18	special protection service (JS9 32)
19	Western democracy influence (EFA 514)

Although the AN combinations in these types are not listed, are not names and do not occur with unspaced orthography in the corpus, it is nevertheless striking that many of them represent common collocations such as *dilute solution*, *fatal crash* and *online catalogue*. Two of them, *high debt* and *high sulphate* are reminiscent of the compound adjectives discussed in previous paragraphs. In other cases, the adjective could arguably have been tagged as a noun e.g. *green* and *minimum*, in which case the construction could simply be regarded as a tri-constituent compound noun. In other words, this residual group are far from being convincingly phrasal, and the hypothesis arises that they in fact belong to the group of compound nouns in which the first constituent is a lexicalised or institutionalised AN. If this hypothesis is correct, we might expect the ANs in this group to have higher frequency that an average AN.

To test the hypothesis that, in the residual group of [AN]N constructions, the AN constituents are institutionalised, their frequency was compared against the frequencies of a large number of AN combinations selected at random from the BNC. Because, for this test, it was not necessary to calculate family sizes, which is a time consuming procedure, it was possible to use a larger random sample than in the previous section. The BNCweb (CQP-Edition) interface (Hoffmann & Evert 2006) was used to search for strings with the following form: article, adjective, common noun, punctuation. This ensured that the AN combinations retrieved were units in which the adjective modified the noun. 5000 such hits were extracted at random, together with their type frequencies in the corpus, and these were compared with the type frequencies of the AN combinations in the residual group shown in Table 2. The average frequency for the random group was mean=1.11, median=1, and in the group from Table 2 the mean was 25.47 and the median 6. Even after log transformation, these frequencies were not even approximately normally distributed, so a non-parametric test, the Wilcoxon rank sum test, was used to assess the significance of this difference: it was indeed found to be highly significant ($w=2769.5$ $p < 2.2e\text{-}16$). In other words, the AN combinations in these residual types are significantly more frequent than the average AN.

This difference in frequency is so large that it suggests the possibility that it might be due to an artefact in the data. If the AN constituents of our residual types contain particularly frequent adjectives and nouns, the AN frequencies of this small set might be artificially elevated. To check this possibility, a second test was run. This time, the frequencies of the AN combinations in Table 2 were compared with the frequencies of all other AN combinations in the BNC composed from the same set of constituents, in other

words all combinations in which the adjective was one of *bare, dependent, dilute, fatal* etc and all combinations in which the noun was one of *kitchen, employee, solution, crash* etc. The mean frequency of the AN constituents in this group was 4.55 and median frequency was again 1. Using the Wilcoxon rank sum test to compare these values with the values for the AN constituents in Table 2 again showed a very significant difference (w=14775, p=3.962e-10). The fact that the group in Table 2 has much higher frequency both than AN combinations in general, and than other combinations with those particular adjectives and nouns, suggests that the combinations found in [AN]N constructions are relatively lexicalised. If this is so, then these constructions can simply be interpreted as compounds of N and AN, and they say nothing about the status of any putative corresponding NN construction.

If these combinations should indeed be interpreted as compounds of N and AN, one might expect there to be a closer bond between the adjective and first noun that between the two nouns. To test this hypothesis, a paired Wilcoxon test was conducted comparing the frequency of the AN in each of these combinations with the corresponding NN frequency. As stated previously, the mean AN frequency was 25.47 and the median AN frequency was 6; this compared with a mean of 3.74 for NN frequency and a median of 2. After adding some jitter to the data in order to avoid having tied values (cf. Baayen 2008: 74), the paired Wilcoxon test showed a highly significant difference between AN frequency and NN frequency (v=24, p=0.002838). Overall, for the [AN]Ns in Table 2, the AN combination occurs significantly more frequently than the corresponding NN.

The result described in the previous paragraph might be irrelevant to the current discussion if AN constructions are in general more frequent than NN constructions. To check this, 5000 NN combinations were selected at random from the BNCweb (CQP-Edition) interface (Hoffmann & Evert 2006) in the same way as described above for AN combinations. The frequencies of the random NN pairs were then compared with the frequencies of the random ANs. For random AN the mean frequency is 1.11 and the median is 1; for random NNs the mean is 1.14 and the median is also 1. Overall, the NN combinations are marginally more frequent than the AN combinations, and, surprisingly, this difference turns out to be highly significant (w=9619326, p=1.211e-0.6). This highly significant difference, despite a relatively small difference in the means and no difference in the medians, is presumably due to the fact that the data sets are so large.

Overall then, there is evidence that the AN constituents in these residual types are significantly more frequent than AN combinations in general, significantly more frequent than other AN combinations with the same adjective or noun, significantly more frequent than the corresponding NN combinations, and that these differences are not due to differences in the language at large. This suggests that in order for an ANN sequence to be interpreted as having the structure [AN]N, the AN combination has to be more strongly bound than the corresponding NN combination would be. If this is not the case, in other words if NN is more strongly bound than AN, the natural interpretation is that the adjective modifies N2, or perhaps the NN as a whole.

In summary, all the examples of [AN]N in the database, with the exception of eight appositive constructions, show evidence that AN is lexicalised, or at least more tightly bound than the corresponding NN. This is perhaps not surprising, since in cases where NN is more tightly bound than AN, the natural interpretation is that the adjective modifies N2 or the NN as a whole. In most cases where AN is a lexical unit, the AN combination has the distribution of a noun, although in some cases it has the distribution of an adjective. It may be that even in the appositive types, it would be possible to demonstrate a tighter connection between the adjective and first noun than between the two nouns, although this remains a question for future research. What these results indicate is that, given a particular NN, the possibility of forming a corresponding [AN]N depends more on the availability of a lexicalised or institutionalised AN constituent than it does on the morphosyntactic status of the NN. Having said that, the more frequent and/or semantically tightly bound NN is, the more difficult it will be to find an AN

constituent that is even more frequent and/or tightly bound. To this extent, the availability of N1 for independent modification can be seen as a reflex of the frequency and degree of lexicalisation of NN.

4.2. [AdverbAdjectiveNoun]Noun

A search for strings labelled 'adverb adjective noun (adjective) noun' returned 16624 hits in 2894 different texts. These were thinned, using the random selection method provided by the corpus interface, to 5000 hits, and these 5000 were inspected in context to establish their structure. In the majority of cases (3772), the structure was [AdvA][NN]: in other words, a prenominal adjective phrase modifying (the head of) a NN. In *very unfair power battle* (KRL 5239), for example, it is the battle that is very unfair. In a further 1171 cases, the AdvANN string did not constitute a constituent, for example: *however, by then feelings were* ***so high Mr Pennell*** *resisted arrest* (HJ3 7205). This left only 71 hits with the structure [AdvAN]N. These included 63 different types, which are shown in Tables 3 and 4.

Table 3: Institutionalised expressions, names and apposition in [AdvAN]N

	institutionalised
1	too fast ascent warnings (ARE 390)
2	massively parallel systems builders (CNF 19)
3	massively parallel applications gap (CPL 2)
4	massively parallel processing pioneers (CTN 277)
5	very small aperture terminal (CBU 1920)
6	very low birthweight infants (EA2 632)
	names
7	Less Favoured Areas Directive (B02 14)
8	Most Favoured Nation status (K5D 5435)
9	Less Favoured Area supplement (K5H 456)
	apposition
10	widely used text-book Elementary Chemical (A1W 141)
11	normally tedious rogue Autolycus (AJN 297)
12	very dear friend Alexander (CKC 996)
13	pretty blonde tourist Julie (HAE 3022)
14	then Soviet counterpart Eduard (HLD 2950)
15	twice champion driver Graham (K4C 280)
16	internationally famous hypnotist Andrew (K4N 22)

In Table 3, items 1-6 involve institutionalised expressions similar to those discussed in the previous sections: *massively parallel* is a conventionalised expression in computer science and *massively parallel processing* is often abbreviated to *MPP*. *Too fast ascent* is an institutionalised expression in the field of diving, *very small aperture terminal* is a frequent expression in the field of satellite communication, often abbreviated to *VSAT*, and *very low birth weight* is a lexical expression in the field of medicine, abbreviated to *VLBW*. In items 7-9, either the AdvAN constituent or the whole construction are names. These various types do not therefore constitute evidence about the status or even existence of the corresponding NN. Items 10-16, however, are appositional. These are the types we expected to find if appositional structures of the form [AN]N are phrasal. It therefore seems that constructions of this type may be best analysed as the apposition of two noun phrases.

The remaining 47 tokens are shown in Table 4. It is immediately striking that many of the AN combinations resemble those classed as compound adjectives in section 4.1.3, both in terms of their familiarity as collocations and the prevalence of *high* in adjective position.

Table 4: 'Compound adjectives' in [AdvAN]N

an almost **short scale** element (C9J 874)
comparatively **low salt** diets (ABB 360)
the completely **free market** approach (CE8 69)
distinctively **inner city** problems (BN8 34)
the essentially **old hat** rock opera theatrics(CHB 2230)
extremely **good value** banking service (F9D 688)
extremely **low temperature** regions (KRH 2905)
formerly **Eastern Bloc** countries (ACR 3411)
a generally **low key** display. (HJ3 4463)
increasingly **higher order** objectives (EVV 301)
these largely **working class** conservatives (EAY 866)
the more **common sense** view (CS2 675)
much **better quality** possession (CB3 735)
a much **longer term** thing (AKU 270)
much **lower level** functions (CSK 444)
predominantly **good class** housing (FBJ 136)
a predominantly **working class** area (FR4 225)
the previously **low wage** areas (HXP 193)
purely **private sector** companies (EX2 903)
a rather **bad taste** way (G1W 2802)
really **good quality** typesetting (G00 2622)
this relatively **low budget** film (A0E 53)
relatively **low cost** partner production (HXJ 40)
a relatively **short term** thing (JA9 231)
somewhat **better quality** Other Ranks (BNB 470)
substantially **free market** economies (H9F 835)
ultra **high quality** Josephson junction devices (BMK 893)
ultra **high speed** serial processors (BMC 3278)
ultra **long range** aircraft (CAU 54)
the very **good fitting** garments (KRJ 38)
a very **good quality** bitch (AR5 1196)
very **good quality** Fender Strat derivatives(C9K 2549)
very **high energy** particles (KRH 3021)
very **high energy** protons (KRH 3017)
the very **high grade** Norlands nanny training (KC0 5234)
a very **high quality** synthetic range (CC0 1008)
a very **high quality** tool (G00 3049)
a very **high speed** backbone (KA4 308)
very **high value** crops (APN 460)
very **high yield** synthesis (ALW 331)
very **large capacity** disk drives (CPY 11)
very **large scale** unemployment (CAN 117)
a very **long term** problem (BN4 1642)
a very **long term** solution (HRK 582)
a very **low calorie** diet (B3G 1361)
a very **low profile** game (FUK 604)
very **real time** intelligence (ADL 863)

A query to the BNC revealed that all of the AN types in Table 4 do occur hyphenated or concatenated in the corpus, sometimes with very high frequencies. In order to test the hypothesis that these ANs belong the 'compound adjective' group, the following variables were calculated for each AN combination in Table 4: spelling ratio, attributive proportion and positional family size. These were compared with the same variables for the random sample of ANs described in section (4.1). In all cases, the values for the types in Table 4 were significantly higher than the values for the random selection. In other words, the AN combinations in Table 4 are significantly more likely to be spelt with unspaced

orthography (t = 7.6458, p = 1.043e-11), occur in attributive position for a significantly higher proportion of their total occurrences (t = 7.1895, p = 1.626e-11) and modify a significantly larger number of nouns (t = 13.8494, p< 2.2e-16). Of course, these factors are not unrelated: AN types that modify a large number of head nouns are likely to occur in attributive position relatively often, so that attributive proportion and positional family size will tend to be correlated. Furthermore, there is a tendency for AN combinations to be written hyphenated when they occur in attributive position, so that a high attributive proportion is likely to be associated with a high spelling ratio. Nevertheless, the fact that these 'compound adjective' types differ so significantly in these respects from AN combinations in general, provides strong evidence that they are atypically prone to behave as modifiers.

The values of these variables for the items in Table 4were then compared against the values found for our compound adjective group in [AN]N constructions. In all cases there was no significant difference at a 5% level, suggesting that these AN types do indeed constitute a recognisable cluster with similar distributional properties. The [AdvAN]N constructions listed in Table 4 are those we predicted would occur if these AN types have the distribution of adjectives, and they therefore constitute further evidence for this analysis. In other words, although these strings are 'syntactic' in the sense that they seem to have, or to be derived from, expressions with the internal structure and semantics of phrases, they are lexicalised in the sense that they are very frequent collocations with the distribution of single words.

Despite the evidence that the AN combinations in Table 4 are institutionalised and have the distribution of adjectives, the question arises as to whether the adverbs in the larger constructions modify the AN as a unit or modify the adjective alone. For example, is *completely free market approach* best analysed as *[completely [free market]] approach*, i.e. an approach which is completely 'free market' in nature, or as *[[completely free] market] approach*, i.e. an approach in which the market is completely free? In some cases, one interpretation may seem more likely than the other, while in other cases, both interpretations seem equally plausible. What is striking, however, is that with the exception of the appositional constructions and highly institutionalised expressions listed in Table 3, all structures of the form [AdvAN]N found in the corpus involve highly institutionalised AN pairs, as indicated by the high spelling ratios. If the correct analysis is that the adverb modifies the adjective alone, it is surprising that the strings with the most apparently phrase-like internal consistency of any in our database seem, with few exceptions, to involve such frequent and highly collocated combinations. In fact, if the adverb modifies the adjective alone, then the AN string is not a constituent of the larger construction and there would therefore be no way of explaining the fact that this construction only seems to arise where the AN combination forms a relatively tightly-bound unit. It therefore seems that the analysis which best corresponds with the empirical evidence is that the adverb modifies the AN as a unit, although it should be conceded that there is some ambiguity in terms of possible interpretation of this structure. Jespersen (1914: 32) reaches a similar conclusion.

The frequencies of the various different types of [AdvAN]N are shown in Table 5. These results serve to confirm the results found for [AN]N types: when N1 appears to be modified independently of N2, the AN or AdvAN constituent forms a lexicalised or institutionalised unit, relative to NN, except where N2 constitutes an appositive modifier.

Table 5: Distribution of patterns in [AdvAN]N

AAN and/or AANN is proper name	3	4.5%
not proper name, but AAN and/or AANN is lexicalised	6	9.1%
sub-total	9	13.6%
neither of the above, but AN forms a 'compound adjective'	50	75.8%
none of the above, but N2 is appositive	7	10.6%
total	66	100.0%

5. Modification of N2: results and discussion

5.1. Noun [Adjective Noun]

The search for strings labelled 'noun adjective noun' returned 105,248 hits in 3628 different texts. A random selection of 8629 of these tokens were manually checked in context, and those with the structure N[AN] were extracted. These represented about 14% of the total, suggesting that this construction occurs about 150 times in every million words. In other words, it is about six times less frequent than the [AN]N construction. In all, 1233 N[AN] tokens were found, corresponding to 1070 N[AN] types and 878 AN types. The most striking thing about this data is that in 701 cases, i.e. about 66% of the N[AN] types, N1 is a proper noun. Out of a total of 719 types of N1, 464 (65%) were names, and of these, 88 (19%) were acronyms. A further breakdown of the results is shown in Table 6.

Table 6: Distribution of patterns in N[AN]

N1 is proper noun and NAN is proper name	356	33.3%
N1 not proper noun, but NAN is proper name	47	4.4%
neither of above, but AN and/or ANN listed	365	34.1%
none of the above, but A is associative	19	1.8%
sub-total	787	73.6%
none of the above, but N1 is a 'proper noun'	156	14.6%
none of the above, but N1 is a 'material noun'	16	1.5%
none of the above, but N1 has an incorporated number	38	3.6%
residual types	73	6.8%
total	1070	100.0%

In 33.3% of the examples, both N1 and the whole construction constitute names, e.g. (22a), and in a further 4.4% of cases, the whole construction is a name, even though N1 is not, e.g. (22b). The various types of name found are shown in Table 7: by far the most common type is one where N1 is a place name and N[AN] is the name of an organisation based in that place (22a).

Table 7: Name types in N[AN]

Name type	N1 name types		NAN name types	
place	269	58%	58	14%
company	80	17%	40	10%
group/organisation	60	13%	250	61%
personal	35	8%	11	3%
product	0	0%	8	2%
publication	1	0%	14	3%
other	19	4%	31	8%
	464	100%	412	100%

In a further 34.1% of cases, the AN constituent and/or the whole construction was listed in OED Online and/or Wikipedia. An example is given in (22c), where *inner tube* is listed in the OED Online. The cases where AN was listed also included a large proportion with a proper noun as N1. A further 19 types involved associative adjectives, and an example of this pattern is shown in (22d).

(22) (a) ... was acquired by the **York Archaeological Trust** ... (JTE 47)
(b) ... **the Gas Advisory Service** ... will check all appliances ... (FTY 260)
(c) … has to rely on hand tools … and the odd **bicycle inner tube** … (BMD 1116)

(d) A controller is serviced in the **Depot Electrical Compound**. (B09 1316)

The remaining types are better candidates for a phrasal analysis. It is immediately striking that, as predicted by Plag (2003: 160), many of the N1s in this group fall into particular classes. In most of these cases, N1 is a proper noun, even though the construction as a whole is not a name. Examples are shown in (23).

(23) (a) … Holywood and Instonians use the **Olympia synthetic pitch** … (HJ3 7958)
(b) … consultations on options for the **Ipswich northern bypass** … (KN3 652)
(c) … Edward Lucente, once an **IBM bright light** … (CMX 475)

In addition, there are two other clearly recognisable groups: firstly, those where N1 is a 'material noun', as exemplified in (24), and secondly, those where N1 is a combination of integer plus noun, as exemplified in (25).

(24) (a) ...women carry **brass bottomless bowls** ... (AEA 171)
(b) ...velcro and **canvas brown trousers** ... (ACP 1032)
(c) ...the **wax hermaphroditic torso** ... (CKW 481)

(25) (a) ...a mere **£10 annual subscription** ... (GXA 1057)
(b) ...using **15mm laminated chipboard** ... (ECJ 335)
(c) ...the **seventy-acre industrial site** ... (APP 824)

As Bauer & Huddleston point out (2002: 1660), these integer plus noun combinations are not nominals, since the noun is not inflected for number: in their analysis these types constitute compound adjectives. If this analysis is correct, then these constructions are irrelevant to the status of NN: the tagging of e.g. *15mm laminated chipboard* as NAN is a mistake, and *15mm* should actually be labelled 'adjective'. Combinations of integer plus noun can then be regarded as similar to, or perhaps even as a sub-class of, the 'compound adjectives' described in section 6.4.1, in which the head of the adjective is morphologically a noun. These three classes then, namely material nouns, proper nouns and nouns that incorporate an integer, may tend to give a phrasal flavour to constructions tagged as NN, in which they occur as first constituent. If so, we would expect to find constructions of the form N[AdvAN], in which the first 'noun' falls into these classes.

Finally, there are 73 types in the data, representing 6.8% overall, which seem to be potentially phrasal despite the fact that the first noun does not fall into any of these three classes. Some examples are shown in (26).

(26) (a) ...with **minority Russian populations** ... (K5H 3602)
(b) ...Martin's **trademark hang-dog mooch** ... (CAE 1317)
(c) ...the **twin heart-shaped pockets ...** (FRF 3387)
(d) ...the **majority communist faction** ... (HLH 800)
(e) ...punished his **rebel Celtic mercenaries** ... (H0K 916)
(f) ...a **weekend residential session** ... (ALB 166)

What is striking about these types is that many of first nouns are listed in the OED as both noun and adjective, and it may be that they represent intermediate types between prototypical nouns and prototypical adjectives. To test the hypothesis that these items are distributionally similar to attributive adjectives, I calculated the family size ratio for each N1 in the residual group. This sample was then compared against the N1 family size ratios of a random sample of 1000 NN types produced by Bell (2012). The hypothesis is that the first nouns in the potentially phrasal types exemplified in (26) typically occur as modifiers rather than heads, and will therefore have a higher family size ratio than N1 in

the average NN. This prediction turns out to be correct: the mean family size ratio for all first nouns in the random sample is 0.338, whereas the mean family size ratio for the first nouns in this group is 0.823. The potentially phrasal types therefore have a significantly higher N1 family size ratio (t=4.1285, p=8.767 e-05). This suggests that the extent to which any NN has a phrasal nature may depend on the identity of N1. Where N1 is a MODIFIER NOUN, the NN will be more loosely bound and more phrase-like, in the sense that adjectives modifying N2 can occur between N1 and N2. What is meant by the term 'modifier noun' is that such nouns occur as N1 in a large number of NN combinations but rarely if ever occur as the head of such combinations. Semantically these nouns also tend to be adjective-like in the sense that they often have adjectival near synonyms: for example, *characteristic* for *trademark, identical* for *twin, rebellious* for *rebel.*

5.2. Noun[AdverbAdjectiveNoun]

A search for strings labelled 'noun adverb adjective noun' returned 2622 hits in 1432 different texts. On inspection, the majority of these turned out to be mistags of various sorts. For example, the first word was often one that would normally be classed as an adjective e.g. *an initial slightly guilty mistrust* (H9H 2969), or the 'adverb' was actually a preposition before a final noun phrase, e.g. *the slope below High Wood* (HPO 1054). Only 69 tokens out of the whole corpus of 100 million words were found to have the structure N[AdvAN]. Furthermore, within these there was considerable repetition, so that they represented only 47 types of N[AdvAN]and a mere 30 types of AN. In 29.2% of the N[AdvAN] types, there was evidence of lexicalisation: either the whole construction, e.g. (27a), or the AdvAN constituent, e.g. (27b), was a proper name, or the AdvAN constituent constituted a lexicalised expression. In (27c), for example, *directionally selective ganglion cells* occurs frequently in the domain of neuroscience and is abbreviated to *DSGC*.

(27) (a) Garnier … Dry **Skin Daily Nourishing Cream** … (C8A 667)
(b) … the **draft Less Favoured Areas** Directive … (B02 34)
(c) the preferred directions of the **on-type directionally selective ganglion** cells (FBD 90)

Table 8: Distribution of patterns in N[AdvAN]

N[AdvAN] or AdvAN is (part of) proper name	4	8.33%
not name, but AdvAN is lexicalised	10	20.83%
sub-total	14	29.17%
none of the above, but N1 is a proper noun	16	33.33%
none of the above, but N1 is a 'material noun'	5	10.42%
none of the above, but N1 has an incorporated number	7	14.58%
residual types	6	12.50%
total	48	100.00%

The frequencies of the various patterns of N[AdvAN] are shown in Table 8. A look at the types that are not lexicalised confirms the hypotheses of the previous section: in almost all cases N1 is either a proper noun, e.g. (28a), or a material noun, e.g. (28b), or has an incorporated number, e.g. (28c).

(28) (a) Spread the bread with **Lurpak slightly salted butter** (H06 1145)
(b) **UPVC double glazed side** window (G2A 793)
(c) we have arranged a 3 **course typically Dutch meal** (EBN 670)

The remaining 7 types are shown in (29). In two cases, (29a) and (29b), the first noun is part of a compound adjective. In another two cases, (29c) and (29d), the first noun, *minimum*, is adjective-like, and may be better analysed as an adjective. In the remaining three types, it is striking that N1 is part of a lexicalised phrase: *fan- in- fin* (29e), *sealed*

unit (29f) and *third world* (29g). This suggests the possibility that phrasal compounds may be amongst the more loosely bound types.

(29) (a) ...a **low-income primarily hispanic area** ... (FBH 385)
(b) ...high **quality financially oriented specialist** ... (CBY 173)
(c) ...following **minimum perfectly coordinated steps** ... (J52 1507)
(d) ...the **minimum legally required number** ... (JNH 15)
(e) ...a **fan-in-fin mainly composite 12-seater** ... (CAU 130)
(f) ...sealed **unit double glazed windows** ... (G2A 152)
(g) ...third **world rapidly expanding populations** ... (HUM 495)

Overall, the results of the N[AdvAN] search provide further evidence that certain noun tend to give a phrase-like quality to NNs in which they occur as first constituent.

6. Conclusion

6.1. Summary of findings

6.1.1. [Adjective Noun] Noun

Evidence from listedness, spelling ratio and other frequency measures has shown that, in the great majority of cases, [AN]N constructions contain an institutionalised or lexicalised AN constituent. In most cases, the AN constituent has the distribution of a noun and cannot therefore be adverbially modified. In such cases, the overall construction can be represented by (30a). In some cases, however, the AN constituent seems to have the characteristics of an attributive AP, and can be adverbially modified. The structure of such constructions can be represented by (30b): evidence for AN strings that function as adjectives comes from their high spelling ratio, frequent occurrence in attributive position and the large number of nouns they modify. A significant proportion of this type involve the adjective *high* and can be represented by the schema shown in (30c). In a few cases, [AN]N combinations represent appositional constructions with the pattern shown in (30d).

(30) (a) $[[AN]_{N}[N]_{N}]_{N}$
(b) $[[AN]_{AP}[N]_{N'}]_{N'}$
(c) $[[high N]_{AP}[N]_{N'}]_{N'}$
(d) $[Det[AN]_{N'}]_{NP}[Nprop]_{NP}$

For a subset of (30a), it was shown that the frequency of AN significantly exceeds that of NN. It may be that this is true of [AN]N constructions in general: in cases where NN is more frequent than AN, the natural interpretation is that the adjective modifies the head noun, N2, or the compound as a whole.

6.1.2. Noun [Adjective Noun]

Where AN forms a highly institutionalised or otherwise lexical unit, the structure of N[AN] can be represented by (31a). The adjectival element is not available for adverbial modification, since it forms part of a noun. In other cases, however, adverbial modification does seem to be admissible and, in these cases, N1 tends to fall into one of a limited number of categories. In the majority of such cases, N1 is a proper noun, and in the majority of these cases, the overall construction is itself a proper name: the structure of the construction is therefore represented by (31b) or (31c).

(31) (a) $[[N]_{N}[AN]_{N}]_{N}$
(b) $[[Nprop]_{N'}[AN]_{N'}]_{N'}$
(c) $[[Nprop]_{NP}[AN]_{N'}]_{NP}$

(d) $[[\text{NumN}]_{AP}[\text{AN}]_{N'}]_{N'}$
(e) $[[\text{Nmod}]_{N'}[\text{AN}]_{N'}]_{N'}$
(f) $[[\text{material}]_{N'}[\text{AN}]_{N'}]_{N'}$

In other cases, the first noun is preceded by a numeral with which it forms a compound. These compounds have the distribution of adjectives since they can be pre-modified by adverbs, and they therefore resemble the AN constituents in (30b) in having the distribution of adjectives despite being headed by morphological nouns. The structure of these constructions is shown in (30d). In other cases where adverbial modification of AN is possible, it is hypothesised that N1 constitutes what I have called a 'modifier noun'. Such nouns modify a wide range of head nouns but are themselves rarely modified by other nouns, i.e. they have a large family size in N1 position relative to their family size in N2 position. The structure of the resulting constructions is shown in (30e). Another recognisable group amongst those N[AN] constructions where adverbial modification is possible are those where N1 is material noun (31f): this group may be a subset of (31e).
Baayen (2010) finds that, to a considerable extent, the order in which English nouns occur in compounds can be described in term of an acyclic directed graph. That is to say that, for a large set of nouns $\{N_1, N_2,...,N_N\}$, it is possible to find an order such that for any compound of the form N_iN_j, N_i precedes N_j in the order for any i and j. As would be expected from such an ordering, nouns at one end of the graph are found only in N1 position while those at the other end of the graph are attested only in N2 position. In other words, nouns can be largely ordered according to the extent to which they typically occur as the modifiers or heads of NN combinations. A hypothesis that arises from the results presented here, is that the further up the graph a noun occurs, i.e. the more typically it behaves as a modifier, the more phrase-like are NNs in which it occupies the first position.

6.2. Discussion

Given NN, the possibility of [AN]N depends on the availability of a relevant AN that is more highly institutionalised than NN, not just on the availability of an adjective that could potentially modify N1. If such an AN combination is not available, then the interpretation of any string in which an attributive adjective precedes NN is that the adjective modifies the second noun, or the compound as a whole. The existence of an [AN]N combination therefore tells us little about the status of the corresponding NN, expect perhaps as a reflection of its frequency and degree of semantic lexicalisation.

The availability of N[AN] depends largely on the nature of N1. Where N1 is a proper noun or has an incorporated numeral or occurs high up on the directed compound graph (Baayen 2010), the NN has phrase-like characteristics, and N1 and N2 can be separated by an adjective that modifies N2. It is of course possible that in these cases too, the availability of the pattern depends on there being an AN combination that is more highly institutionalised than NN, but this has not been tested here and must remain a question for future research.

One possible interpretation of the results is that those nouns that I have called 'modifier nouns', including material nouns, represent a category similar to the one said to be represented by associative adjectives. In this analysis, modifier nouns would be regarded as having the distribution and semantics of adjectives but the morphology of nouns, just as associative adjectives have the distribution and semantics of nouns but the morphology of adjectives. Similarly, those AN combinations and NumN combinations that I have called compound adjectives can also be regarded as examples of category mismatches, since they have the distribution and semantics of adjectives but are headed by morphological nouns.

On the 'category mismatch' view, the differences in distributional frequencies between 'modifier nouns' and nouns in general could reflect an underlying categorical distinction, rather as a general difference in height between men and women reflects an underlying binary distinction in genetic makeup. The classification of NNs as compounds or phrases might then be based on the category of N1, albeit in some cases a 'mismatched' category. However, as discussed in section 1, there is evidence that the distinction between morphological and syntactic objects is not in fact categorical, and selecting any test as criterial runs the risk of circularity. An alternative is to view the frequency and distributional data as the fundamental type. On this view, categories such as 'adjective', 'modifier noun' and 'noun' are more like shoe sizes, imposing a discontinuous classification on an essentially continuous variable (foot length). In this analysis, the availability of N1 and N2 for independent modification in any NN would be probabilistically determined depending on the frequencies with which the two nouns occur together, and in combination with other nouns and adjectives. To the extent that the possibility of such modification reflects a difference between compound-like and phrase-like types, this analysis would be compatible with a non-modular view of morphology and syntax: the difference between morphological and syntactic objects would be a matter of degree. The choice between these two analyses could be made on the basis of statistical modelling, by comparing the success of categorical and probabilistic approaches in predicting which NNs allow modification. For the time being, however, this must remain a question for future research.

References

Alexiadou, Artemis, Liliane Haegeman, & Melita Stavrou 2007. *Noun phrase in the generative perspective* (Studies in Generative Grammar 71). Berlin & New York: Mouton de Gruyter.

Arnaud, Pierre 2008. Adjective + Noun sequences in attributive or NP-final positions: Observations on lexicalization. In Sylvianne Granger & Fanny Meunier (eds.), *Phraseology: An Interdisciplinary Perspective,* 111—25. Amsterdam, Philadelphia: John Benjamins.

Baayen, R. Harald 2008. *Analyzing linguistic data. A practical introduction to statistics.* Cambridge: Cambridge University Press.

Baayen, R. Harald 2010. The directed compound graph of English. In Susan Olsen (ed.) *An exploration of lexical connectivity and its processing consequences: New impulses in word-formation*, 383—402. Hamburg: Buske.

Baayen, R. Harald, Victor Kuperman, & Raymond Bertram 2010. Frequency effects in compound processing. In *Cross-Disciplinary Issues in Compounding*, Sergio Scalise and Irene Vogel (eds.), 257-270. Amsterdam: John Benjamins.

Bauer, Laurie 1983. *English word-formation.* Cambridge: Cambridge University Press.

Bauer, Laurie 1998. When is a sequence of two nouns a compound in English? *English Language and Linguistics* 2. 65-86.

Bauer, Laurie 2001. *Morphological Productivity.* Cambridge: Cambridge University Press.

Beard, Robert 1991. Decompositional composition: The semantics of scope ambiguities and 'bracketing paradoxes'. *Natural Language & Linguistic Theory* 9(2), 195—229.

Bell, Melanie J. 2005. Against nouns as syntactic premodifiers in English noun phrases. *Working Papers in English and Applied Linguistics* 11, 1-48.

Bell, Melanie J. 2011. At the boundary of morphology and syntax: Noun noun constructions in English. In Alexandra Galani, Glynn Hicks and George Tsoulos (eds.) *Morphology and its interfaces*. Amsterdam and Philadelphia: John Benjamins.

Bell, Melanie J. 2012. *The English NN construct: its prosody and structure*. PhD thesis, University of Cambridge.

Bell, Melanie J. & Ingo Plag. 2012. Informativeness is a determinant of compound stress in English. *Journal of Linguistics*. Available on CJO 2012 doi:10.1017/S0022226712000199. http://journals.cambridge.org/abstract_S0022226712000199.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education Limited.

Bloomfield, Leonard 1935. *Language*. London: George Allen & Unwin.

Booij, Geert 1985. Coordination reduction in complex words: a case for prosodic phonology. In Harry Van der Hulst & Norval Smith (eds.) *Advances in non- linear phonology,* 143—160. Dordrecht: Foris.

Booij, Geert 2009. Phrasal names: A constructionist analysis 1. *Word Structure* 2(2), 219-40.

Bresnan, Joan & Sam Mchombo 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13, No.2181—254.

Croft, William 2001. *Radical construction grammar*. Oxford: Oxford University Press.

Culicover, Peter W. & Ray Jakendoff 2005. *Simpler Syntax.* Oxford: OUP.

Dahl, Östen 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Dalrymple, Mary & Irina Nikolaeva 2006. Syntax of Natural and Accidental Coordination: Evidence from Agreement. *Language* 82:4 824-849

Davies, Mark 2004-. BYU-BNC: The British National Corpus. http://corpus.byu.edu/bnc

Davies, Mark 2008-. Corpus of American English (COCA): http://corpus.byu.edu/coca/

De Jong, Nivja H. 2002. *Morphological Families in the Mental Lexicon*. MPI Series in Psycholinguistics, Max Planck Institute of Psycholinguistics, Nijmegen.

Di Sciullo, Anna-Maria & Edwin Williams 1987. *On the definition of word*. Cambridge, MA: MIT Press.

Dixon, Robert M. W. & Alexandra Y. Aikhenvald 2002. Word: a typological framework. In Dixon & Aikhenvald (eds.), 1-41.

Dixon, Robert M. W. & Alexandra Y. Aikhenvald (eds.). 2002. *Word: a cross-linguistic typology.* Cambridge: Cambridge University Press.

Downing, Pamela 1977. On the creation and use of English compound nouns. *Language* 53, 810—42.
Giegerich, Heinz 2005. Associative adjectives in English and the lexicon–syntax interface. *Journal of Linguistics* 41(3), 571-91.
Giegerich, Heinz 2009. Compounding and lexicalism. In *The Oxford handbook of compounding,* Lieber, Rochelle & Pavol Štekauer (eds.), 178-200. Oxford: Oxford University Press
Harris, Zellig 1946. From Morpheme to Utterance. *Language* 22:3, 161–183.
Hoffmann, Sebastian & Stefan Evert 2006. BNCweb (CQP-edition): The marriage of two corpus tools. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods,* 177—95. Frankfurt am Main: Peter Lang.
Huddleston, Rodney & Geoffrey Pullum (eds.) 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Jesperson, Otto 1914. *A modern English grammar*, *Part II, Syntax*. *First volume.* Heidelberg: Carl Winter's Universitatsbuchhandlung.
Jespersen, Otto 1942. *A modern English grammar on historical principles, Part VI: Morphology*. London: George Allen & Unwin Ltd.
Kabak, Baris 2007. Turkish suspended affixation. *Linguistics* 45:2 311-347.
Keizer, Evelien 2011. English proforms: an alternative account. *English Language and Linguistics* 15(2): 303-334.
Levi, Judith N. 1978. *The syntax and semantics of complex nominals.* New York; San Francisco; London: Academic Press.
Lewis, Geoffrey L. 1967. *Turkish Grammar.* Oxford: Oxford University Press
Lieber, Rochelle 1992. *Deconstructing morphology: Word formation in syntactic theory*. Chicago: University of Chicago Press.
Lieber, Rochelle 2009. A lexical semantic approach to compounding. In Lieber & Štekauer (eds.), 78—104.
Lieber, Rochelle & Sergio Scalise 2006. The lexical integrity hypothesis in a new theoretical universe. *Lingue and Linguaggio* 1: 7-32.
Lieber, Rochelle, & Pavol Štekauer 2009. Introduction: status and definition of compounding. In Lieber & Štekauer (eds.), 3—18.
Lieber, Rochelle, & Pavol Štekauer (eds.) 2009. *The Oxford Handbook of Compounding.* Oxford: Oxford University Press.
Lipka, Leonhard 1994. Lexicalisation and institutionalisation. In Ron Asher (ed.), *The encyclopaedia of language and linguistics.* Vol 4. 2164-2167. Oxford: Pergamon.
Lipka, Leonhard, Susanne Handl & Wolfgang Falkner 2004. Lexicalisation and institutionalisation: The state of the art in 2004. *SKASE Journal of Theoretical Linguistics* 1(1).
Marchand, Hans 1969. *The categories and types of present-day English word-formation: A synchronic-diachronic approach*, 2nd edn. Munich: C.H. Beck'sche Verlagsbuchhandlung.
Matthews, Peter H. 1991. Morphology, 2nd edn. Cambridge: Cambridge University Press.
Matthews, Peter H. 2002. What can we conclude? In Dixon & Aikhenvald (eds.), 266-281.
Olsen, Susan. 2000. Compounding and stress in English: A closer look at the boundary between morphology and syntax. *Linguistische Berichte* 181. 55—70.
Payne, John & Rodney Huddleston 2002. Nouns and noun phrases. In Huddleston & Pullum (eds.), 323—524.
Payne, John, Rodney Huddleston & Geoffrey Pullum K. 2010. The distribution and category status of adjectives and adverbs. *Word Structure* 3(1), 31—81.
Plag, Ingo 2003. *Word-Formation in English.* Cambridge: Cambridge University Press.
Plag, Ingo 2010. Compound stress assignment by analogy: The constituent family bias. *Zeitschrift für Sprachwissenschaft* 29.2
Plag, Ingo, Gero Kunter & Sabine Lappe 2007. Testing hypotheses about compound stress assignment in English: A corpus-based investigation. *Corpus Linguistics and Linguistic Theory* 3(2), 199—233.
Plag, Ingo, Gero Kunter, Sabine Lappe, & Maria Braun 2008. The role of semantics, argument structure, and lexicalization in compound stress assignment in English. *Language* 84 (4), 760—94.
Postal, P. 1969. Anaphoric Islands. In *Papers from the fifth regional meeting of the Chicago Linguistic Society.* Chicago: Chicago Linguistic Society.

Pullum, Geoffrey K., & Huddleston, Rodney 2002. Adjectives and Adverbs. In Huddleston, & Pullum (eds.), 525—596.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik 1985. *A comprehensive grammar of the English language*. Harlow: Longman.

Sadler, Louisa & Douglas Arnold J. 1994. Prenominal adjectives and the phrasal/lexical distinction. *Journal of Linguistics* 30(1), 187—226.

Spencer, Andrew. 2003. Does English have productive compounding? In *Topics in morphology: selected papers from the third Mediterranean morphology meeting Barcelona, September 20—2, 2001*, 329—41.

Spencer, Andrew. 2005. Word-formation and syntax. In *Handbook of word-formation,* Štekauer, Pavol & Rochelle Lieber (eds.) 73-98*.* Dordrecht: Springer.

Spencer, Andrew. 2011. What's in a compound? *J. Linguistics* 47, 481-507

Wälchli, Bernhard. 2005. *Co-Compounds and Natural Coordination. Oxford Studies in Typology and Linguistic Theory.* Oxford: Oxford University Press.

Ward, Gregory, Richard Sproat & Gail McKoon 1991. A pragmatic analysis of so-called anaphoric islands. *Language* 67(3), 439—74.

Morphological Complexity in Maltese: A divergence from canonicity

Maris Camilleri
Surrey Morphology Group
maris.camilleri@surrey.ac.uk

1. Introduction

The paper discusses the notion of *morphological complexity* in Maltese. Morphological complexity is here understood in the same sense as Aronoff's (1994) *morphology by itself*, where the morphology is considered as a separate component in the grammar; a notion that has been recently referred to as *autonomous morphology* (refer for example to Maiden et al. 2011). The aim of this paper is to illustrate how Maltese exhibits a number of phenomena which are complex in the way understood here, i.e. pertaining to the language's morphological component. The complexity discussed will be mostly paradigm-internal, but will also involve interesting accounts of what takes place across lexemes that have long been traditionally classified as belonging to the same set, and will involve phenomena such as stem patterns, which come about as a result of stem allomorphy within the paradigm (Vogel 1994, Booij 1996, Fabri 2009), overabundance, and heteroclisis, and the interactions of these together. Apart from displaying the complexity at hand, the presence of stem allomorphy internal to paradigms falsifies definitions of stems along the lines of Nakov et al. (2004), who define the stem as 'the common part shared by all inflected word-forms' within a paradigm. The work presented here will also go against paradigm definitions that consider *form* relatedness as an essential criterion, e.g. Kenstowicz (2005), who defines the inflectional paradigm as 'words sharing the same stem and differ in the exponence of formal features ...' (p. 47). This results in an underrepresentation of what would have also been regarded as a paradigm, considering that every lexeme 'may have a multitude of distinct stems' (Stump, 2001, p. 33), which can come about as a result of stem-alternations, suppletion or semi-suppletion that results from heteroclisis. Much of the data presented and analysed in this paper has not been discussed in the literature on Maltese. While the language has been characterised by its mix of Semitic and Romance influences, in this paper only data from the Semitic part of the language will be provided, as research on the Romance set of data is the topic of present ongoing research.

That which makes a given language morphologically complex can be measured from a prior expectation which is not met. To analyse our morphologically-complex phenomena along these lines, the canonical typology framework as set out in Corbett (2005, 2007a, 2009, 2011) will be used, which framework has also been applied in the syntactic domain by Polinsky (2003), Seifart (2005), and Suthar (2006) amongst others. Spelling out some of the framework's claims and how its measure of analysis operates, is what follows in §2 below. §3 includes a description of the Maltese verbal paradigm, which will be the locus from where to analyse morphological complexity in Maltese. A segmentation analysis will also be provided, since there has yet been no fixed segmentation account for Maltese. An analysis of non-canonical behaviour, particularly illustrating the non-canonical behaviour of stem-form alternations internal to the paradigm is provided in §4, where it will be postulated that inflection in Maltese is not solely realized by inflectional affixes, but also by the same alternating stem-forms. The study here will build on the work in Corbett & Baerman (2006), Corbett & Baerman (2010), and Baerman & Corbett (2012), where the *lexical* material's real function, as well as the end result brought about by the actual non-canonical behaviour of having non-inert stems will be probed into. In §5 we will then see how complexity internal to the inflectional verbal paradigm can cut across different

binyanim verb-forms in the language. §6 then summarises the key points and concludes the paper.

2. Canonical Typology

Applying a canonical approach to the analysis of language means that definitions of elements, entities, and/or phenomena are taken 'to their logical end point' (Corbett, 2005, p. 25). From there, the language data instances are set against the logical definition/instance, and the *theoretical distance* of the *real instance* from the *canon*, is measured, resulting in a gradience of degrees of non-canonicity (Corbett, 2007, p. 9). The *canonical* illustration functions as a fixed point towards which one can always return to as a standard of measurement, i.e. *the canon*, even though this 'may even be non-existent' (Corbett, 2011a, p. 446). What is then required are measures/dimensions that are able to grade the data accordingly. Morphological complexity can thus be understood as an outcome of the divergence from the *canon*, where the further away from the canonical requirement a given example is, the more non-canonical, and the more morphologically complex it is. Since the focus of this discussion is on the non-canonicity internal to the paradigm, what follows below is a canonical account of what one expects to find in a *canonical paradigm*. The reason for doing so is such that the divergence from canonicity to be illustrated for Maltese in §4 and §5 can be compared in the realm of what one expects to find in this morphological paradigmatic entity.

2.1. Canonical Paradigms

In a canonical paradigmatic system, one would expect to have a product of the multiplication of the features and their values, resulting in the expected total number of cells (refer to Spencer's (2003) notion of *exhaustivity*), each with a distinct word-form (Corbett, 2011b, 2009).[1] In this regard, therefore, a canonical paradigm that realizes four distinct morphosyntactic values, (be they portmanteau or not), is expected to have a structure as in table (1) below. If it happens to be the case that the expected number of cells does not match up with the total number of cells, then violations of the canon would involve *defectiveness* (refer to the references and articles in Baerman et al. 2010) on the one hand, and *overdifferentiation* on the other.[2]

X-a
X-b
X-c
X-d

Table 1: Illustrating the canonical behaviour of a four-celled paradigm

[1] All forms are here understood as surface form structures (see Anderson, 2011).

[2] *Defectiveness* occurs when the exhaustive set of morphosyntactic features in the language, (at least when comparing across the same set/class of lexemes and their verbal paradigms), are multiplied out, the result is such that we get less cells, hence a paradigm which does not include *all* the expected number of cells. Non-canonical *overdifferentiated* paradigms on the other hand are illustrations of paradigms that have additional paradigmatic cells, when one compares the number of cells, representative of the number of features, associated with the rest of the lexicon (Corbett, 2000). Also refer to Gauci & Camilleri (2011) for discussions on this phenomenon in Maltese. When on the other hand different cells do not involve distinct word-forms, and assuming that the features involved are all syntactically relevant, then the non-canonical occurence of *syncretism* is present (Baerman et al., 2005).

The analysis of canonical paradigms falls under the rubric of canonical inflection that is concerned with paradigm-internal behaviour. Internal to the paradigm there is a lexical vs. grammatical material dichotomy. The lexical material, which *should* be the stem, *should not* express any grammatical features and is expected to be inert, non-alternating (Baerman & Corbett 2012: 1). If we consider our simplified canonical paradigm representation in table (1) we can see that the invariable X in all the cells represents the lexical content. On the other hand, the grammatical information usually expressed by the affixal material *should* be distinct in all stems, as illustrated through the four distinct suffixal forms in table (1). While a violation of the distinct affixal material in each cell results in *syncretism* (Baerman et al. 2005), a violation of the inert lexical material canonical requirement results in a stem-form that is not only lexical, but can itself be an exponent of grammatical features, since as long as something displays a change in form within a paradigm, this will, in some way or another serve as an exponent of some kind of morphosyntactic feature or value distinction (Corbett & Baerman 2006, Baerman & Corbett 2010, 2012)**.** As a result of the fact that in places where we ought to have *sameness*, one gets distinct forms, or vice-versa, this is taken to imply an 'increased complexity and/or redundancy' (Corbett, 2009, p. 2).[3] Table (2) below first illustrates a representation of the canonical requirements as reviewed above, and table (3) then illustrates the derivations that result out of this.

Canonicity internal to the paradigm and across lexemes

	Comparison across *cells* of a lexeme	Comparison across *lexemes*
Composition/structure	same	same
Lexical material (stem-shape)	same	different
Affixal material (affix-shapes/ forms)	different	same
Realisational outcome	Different cell-forms	Different cell-forms

Table 2: A representation of canonical inflection internal to the paradigm and across lexemes (Corbett 2011)

	The content of the paradigmatic cell	Deviations	Comparisons across different lexical paradigms	Deviations
Composition/structure	different	fused exponence periphrasis	different	defectiveness overdifferentiation
Lexical material (stem-shape)	different	**stem-alternation**[4] suppletion	same	**heteroclisis**
Affixal material (affix-shapes/ forms)	same	syncretism uninflectability	different	deponency inflectional classes

Table 3: Illustrating the array of non-canonicity in Maltese verbal paradigms (adapted from Corbett 2007b)

Apart from calibrating the actual paradigmatic stem-form behaviour vis-à-vis the canonical requirement, the other dimension to this study includes an analysis that looks

[3] In §4. We will also be looking at another additional dimension to the non-canonical paradigm, following Thornton's (2010, 2011) work on *overabundance*, which involves a cell-internal violation that involves the presence of a number of word-forms in a context where one ought only find one.

[4] The bolded non-canonical behaviours/deviations: stem-alternations and heteroclisis will be among the divergent non-canonical illustrations of morphological complexity that will be discussed in this study.

at stem-behaviour across different lexemes, in analogy to the analysis of non-canonical inflectional classes (Corbett 2009). The different patterns of organisations of stem-form alternations across the different lexemes will be referred to as *stem pattern classes*. Maltese verbs will be classified on the basis of their paradigmatic stem-form behaviour and the stem pattern class they fit in. It is important to note that part of the canonical divergence discussed here, which is independent of any phonological-conditioning, will itself be based on the segmentation analysis provided in §3.2 below.

3. The Maltese verbal paradigm

In this section the Maltese verbal paradigm is described, where some additional non-canonical behaviour, apart from those focused upon in §4 and §5 will be highlighted in §3.1. In §3.2 the segmentation issue will be discussed.

3.1. Getting acquainted

The verbal paradigm in Maltese consists of three sub-paradigms; the indicative perfect and imperfect sub-paradigms and the imperative sub-paradigm.[5] This study will be mainly concerned with the first two sub-paradigms, particularly because the forms in the imperative sub-paradigm are themselves a principal part for the word-forms in the relevant cells of the imperfect sub-paradigm, whereby we are thus dealing with the same set of forms, and which we will not need to represent additionally, here.[6] From the Maltese verbal paradigm representation in table (4), one thus observes that the perfect and imperfect sub-paradigms involve three PERS feature values {1, 2, 3}, and two NUM values {SG and PL}. In the 3rd PERS SG cells there is GEND specification that distinguishes across masculine and feminine values. In the imperative sub-paradigm one only finds two word-forms; one in the 2.SG cell and the other in the 2.PL cell.

Morphosyntactic feature values	kiteb 'write'		
	PERF	IMPERF	IMPERATIVE
1.SG	ktibt	nikteb	
2.SG	ktibt	tikteb	ikteb
3.SG.M	kiteb	jikteb	
3.SG.F	kitbet	tikteb	
1.PL	ktibna	niktbu	
2.PL	ktibtu	tiktbu	iktbu
3.PL	kitbu	jiktbu	

Table 4: The paradigm for *kiteb* 'write'

From table (4) it can already be seen that the stem-form across the different cells and sub-paradigms differs, e.g. *kitb-* in the 3rd PERS feminine cell and in the 3rd PERS PL cell, and *-ktb-* in the imperfect PL cells. Accounting for the pattern of stem-form alternations will be the task in §4.1. Additional non-canonical behaviour one can observe from table (4) is the instance of systematic syncretism across the 1.SG and 2.SG word-forms in the perfect sub-paradigm. Recall from §2.1 that in a canonical paradigm one should expect

[5] Regarding what we have here as aspectual paradigms in the indicative mood, follows the work of Borg (1981, 1988) and Fabri (1995). Refer to Hetzron (1997) for a distinct view on the matter however, who considers these sub-paradigms in Semitic languages to realise temporal feature values: PAST and PRESENT respectively.

[6] There is only one lexical item that does not pattern in this way, and that is COME, whose stem-form in the imperfect SG and imperative SG is *n-i-ġi* 'I come' and *ejja* 'come.SG.IMPER', respectively, involving an instance of suppletion.

different word-forms in the different cells, since each cell is understood as realizing a set of distinct morphosyntactic feature values that differ across the different cells. Furthermore, since the canonical stem is invariant, the part of the word-form that is *expected* to differ is the affixal material. From table (4) one sees that the same suffix *-t* is used across the perfect 1SG and 2SG cells. Following the segmentation account provided below in §3.2, this is taken to illustrate an instance of a directional syncretism whereby the 2.SG form itself also becomes the exponent of the 1.SG values. In parallel to this syncretism in the perfect sub-paradigm, we get a similar non-canonical effect in the imperfect sub-paradigm, this time across the 2.SG and 3.SG.F cells. If we combine both patterns of syncretism, as in table (5) below, one sees that, the PERS and NUM values of the form which intersects both patterns, which offers the base for the directional syncretism in both sub-paradigms, are actually non-autonomous (Corbett 2011a, a term attributed to Zaliznjak 1973). In other words, the 2.SG values never have a unique form within the inflectional paradigm in Maltese.[7]

Morphosyntactic feature values	kiteb 'write'	
	PERF	IMPERF
1.SG	ktibt	nikteb
2.SG	ktibt	tikteb
3.SG.F	kitbet	tikteb

Table 5: The non-autonomous illustration of the combination of the 2nd PERS and SG NUM values

3.2. The segmentation adopted in this study

The segmentation adopted in this study is illustrated in table (6), building upon, but moving away from the segmentation analyses provided in Mifsud (1995), Fabri (2009) and Spagnol (2011). The largest variation is found between the segmentation analysis here, and that of Mifsud's. While his segmentation analysis involves what one may wish to refer to as an inflectional class account for Maltese that essentially splits the inflectional classes on the basis of a broad Semitic vs. non-Semitic lexical split. The segmentation is much simpler, and a coherent aspect of the segmentation here is that the analysis does not present different segmentations depending on whether the lexicon is integrated in the Semitic morphology or not. As a result of this, while Mifsud sets the tradition that Semitic verbs and early integrated Romance loans are based on a consonantal root, and the non-Semitic influenced Romance loans involve paradigms built on bases/stems, my account here invokes no such analytic distinction, where through the presence of just one inflectional verbal class in the language, most of the idiosyncrasy is ascribed to the stem.

Morphosyntactic feature values	PERFECT	IMPERFECT
1.SG	-t	n- ~ m-
2.SG		t-
3.SG.M	ø	j- ~ i-
3.SG.F	-(V)t	t-
1.PL	-na	-u ~ -w
2.PL	-t-u	
3.PL	-u ~ -w	

Table 6: The segmentation to be adopted in this study

[7] It is worth highlighting that it is the combination of the NUM and PERS features which is giving us the non-autonomous combination of SG and 2nd PERS values, as in essence, when we consider the imperative sub-paradigm, the 2nd PERS form is actually autonomous there.

From the segmentation in (6), one observes that Maltese involves suffixes in the perfect sub-paradigm and prefixes and suffixes in the imperfect sub-paradigm. This is in itself a non-canonical manifestation. On the basis of the *cell's composition and structure property* (Corbett 2009, p. 2), in a canonical paradigm one expects that if suffixal material is used in a cell, then the paradigm should retain such a position for the inflectional exponents across all the paradigmatic cells. The perfect sub-paradigm can be considered as canonical in this regard, displaying suffixes throughout, realizing PERS, NUM (and GEND) features. When it comes to the exponents involved, it should here be mentioned that unlike previous analyses, the ø in the perfect 3SGM cell in table (6) does not mean that the existence of zero morphs is being acknowledged here. Rather, the ø should be taken to represent the fact that in the 3SGM cell, it is the stem-form itself that is an exponent of these features, and not any additional affixal material.

What is not found in previous segmentation accounts for the language in table (6) is the additional syncretic exponent we get in the perfect sub-paradigm: the *-t* allomorph in the 3.SG.F cell. Another difference is that while Maltese and Arabic literature regard the *-tu* as a suffix, (refer to Lowenstamm 2011 for example), it is implicitly or explicitly taken for granted that the *t* in *-tu* is the exponent of the 2nd PERS, and the *-u* is an exponent of NUM. In table (6) *-tu* is segmented further, in turn providing a rather neat analysis of the *-t* exponent as a default 2nd PERS in the language, and where the presence of syncretism in the paradigm, it comes to realize other feature values, and depending on ASP, it gets its differing placements, either on the left or right edge of the stem-form. From this segmentation analysis, it is only the exponent *-na* that realizes PERS and NUM cumulatively. The imperfect sub-paradigm, on the other hand, neatly involves PERS-realizing prefixes and NUM-realizing suffixes, when present, and is closer to Fabri's (2009) and Spagnol's (2011) segmentation analysis. The exponent of the PL value alone (*-u~ -w*) is shared across both sub-paradigms.

4. Complexity internal to the Maltese verbal paradigm

What follows in this section is an account which delves deeper into the paradigmatic complexity that does not have to do specifically with the affixal material, but is rather concerned with the stem-form behaviour in a sample of Maltese verbal paradigms. Such an account will provide another dimension to the language's divergence from canonicity. Recall that on the basis of a canonical typology account, as mentioned in §2.1, the stem's role should be that which imparts lexical meaning, and in doing so, it must be inert. As mentioned when discussing the paradigm for *kiteb* 'write' in table (4), Maltese paradigms involve alternating stem-forms. In this section I will show how the alternation that is present cannot be considered random, and can be perceived as an interaction of both a phonological and a systematic output of morphological conditioning. Through the analysis provided, following Camilleri (2012), the stem-form in Maltese is considered to be imparting grammatical information that is realized simultaneously with that which is realized by the inflectional affixes themselves. This claim goes against a number of accounts in the Maltese literature, such as the claim in Spagnol (2011) saying that inflection in Maltese is concatenative. The alternating stem-forms will here be treated as a non-concatenative illustration that reduplicates part or all of the affixes' realization, and are taken to be multiple exponents internal to the word-form. This analytical account will in turn show that it is not the case that verb inflection in the language solely 'involves prefixation and suffixation to a stem-base' (Spagnol, p. 37). Furthermore, this non-concatenative dimension to inflection in Maltese (as well as in other Arabic dialects) comes to show that non-concatenative morphology in Semitic languages is not restricted to the *binyanim* system of verb-form alternations, as often asserted in the literature (refer to Booij 2009, for example). If this is really so, then the non-concatenative analysis makes Semitic languages appear more similar to other languages, such as German, for

example, whose non-concatenative system of ablaut-changes can be considered as analogical to the changes we observe within the non-concatenative system of stem-form alternations in Maltese.

In §4.1 the variation that exists in the stem-form alternation behaviour and the patterns observed across two different verbal bases in the language will be exemplified, which will in turn also illustrate how morphological complexity, at least viewed synchronically, cannot be wholly attributed to phonological constraints, in turn analysing that which is unexplicable to be the outcome of the morphological component. §4.2 then presents a discussion with some representative illustrations of *overabundance* in the language, which will also provide us with some interesting supplementary effects that relate to our stem behaviour analysis.

4.1. Stem patterns

In this section two distinct verbal bases in the language will be discussed.[8] From these two distinct classes of verbs, different patterns formed as a result of distinct stem-alternation behaviours will be shown. We will see however that a distinct stem pattern need not cross-classify with a distinct verbal base type. Rather, different types of distinct behaviour will be shown to exist across the individual members of the different verbal bases. The verbal bases which will be looked at here are the: CVCVC verbal bases, illustrated by *laqat* 'hit' and *ħataf* 'grab', and the CV:C type, illustrated by the verbs *mar* 'go' and *żar* 'visit'. There are a number of reasons why these two verbal bases were chosen. Interestingly they manifest distinct illustrations of stem-form behaviour across lexemes of the same type. Through the CVCVC verbal base set it will be shown how verbs of the same type can involve a different sub-pattern of stem-form alternations, which however retain a co-membership in the same stem pattern class. In the case of the CV:C verbs, we will see that the lexemes chosen here do not solely display a distinct sub-pattern of alternation, but rather belong to distinct stem pattern classes altogether. Another distinction associated with the choice of the different verbal bases is also interesting in that in terms of their traditional underlying representation, these differ. A distinction cross these verbal base types based on the nature of their triconsonantal underlying representation will not be pursued here. It we will in fact be shown that this underlying representation does not hold, when based upon surface-form data. The CVCVC set is traditionally considered to belong in the strong *class* of verbs, which do not include a *w/j* radical in their UR, while the CV:C-verbal-base-classified verbs are analysed as *weak* since they do not involve a *w/j* radical, which for this set of *weak* verbs happens to be in the medial position. As will be shown, the reasons why a triconsonantal underlying representation analysis is not upheld here is essentially because it is stem-forms that are in focus here, as we will see a number of discrepancies across what is said to be the underlying representation against what is actually found at the inflectional paradigm, exists. However, the stem-form, under a consonant-root based account can be conceived as outputs of mechanisms that apply on a consonantal root (McCarthy 1982, Fabri 2009), which in turn build up stem-forms and lexical items (e.g. refer to Müller's 2009 account for the Maltese lexicon).

While the analysis of stem patterns formed by a pattern/organisation of internal stem allomorphy requires us to look closely at phonology to see how it conditions stem-forms in the paradigm, phonological facts will not be delved into deeply here, as this would require that we focus on other elements which are not the subject of the discussion in this paper. The notion of the stem pattern class will here be analysed further, and the stem-form is considered as an output available for scrutiny, without

[8] With verbal bases, what I mean here is the phonological shape of the surface base; such that *kiteb* 'write' represented in table (4) belongs to the CVCVC verbal base.

delving in unnecessary detail as to how this pattern comes about, and what forces are responsible for this. It is what is clearly inexplicable, and the effect/consequence of an autonomous morphological layer that will be mostly dealt with and given due attention.

4.1.1. Comparing across CVCVC verbal-based verbs

If we consider the paradigms of the verbs *laqat* 'hit' and *ħataf* 'grab' in table (7) below, we see that notwithstanding the same phonological properties, these differ.

Morphosyntactic Feature values	*laqat* 'hit'		*ħataf* 'grab'	
	PERFECT	IMPERFECT	PERFECT	IMPERFECT
1SG	lqat-t	n-o-lqot	ħtaf-t	n-a-ħtaf
2SG	lqat-t	t-o-lqot	ħtaf-t	t-a-ħtaf
3SGM	laqat	j-o-lqot	ħataf	j-a-ħtaf
3SGF	laqt-et	t-o-lqot	ħatf-et	t-a-ħtaf
1PL	lqat-na	n-o-lqt-u	ħtaf-na	n-a-ħtf-u
2PL	lqat-t-u	t-o-lqt-u	ħtaf-t-u	t-a-ħtf-u
3PL	laqt-u	j-o-lqt-u	ħatf-u	j-a-ħtf-u

Table 7: The paradigms for *laqat* and *ħataf*

These two verbs do not just share the same verbal base, but they also belong to the same ablaut class, i.e. *a-a*.[9] Notwithstanding the difference across the two verbal paradigms, which has to do with the number of stem-forms present, and whether there is any redundant morphological interventions involved, as we will see below, there is however a unifying pattern across the two verbs. This pattern is what we are here referring to as the *stem pattern class*, > i.e. which comes about as a result of the way in which the different morphosyntactic features are conflated within the individual sub-paradigms on the basis of the stem-form alternations and feature value conflations. The stem pattern class represented by the verbs in table (7) conflates the 1^2 cells in the perfect sub-paradigm.[10] It also involves a distinct 3SGM stem-form, as well as a stem-form conflation across the 3SGF^3PL cells. The imperfect sub-paradigm, on the other hand, involves a stem-form alternation that is NUM-based, when we set our data against Corbett & Baerman's (2012) typology of paradigmatic stem-form splits. The pattern that results involves five slots across the whole paradigm. These five slots are split (3 vs. 2) on the basis of an ASP distinction. The abstracted stem pattern class can be represented as in table (8) below.

[9] What is meant by ablaut class *a-a* here is the vocalic pattern in the perfect 3SGM stem/word-form. Note that this differs from the Arabic tradition. An *a-a* classification of CVCVCa verbal bases in Classical Arabic is not used to refer to the 3SGM's vocalic pattern. Rather, *a-a* in the Arabic trandition refers to the V_2 in the 3SGM stem/word-form and the stem-vowel in the imperfect sub-paradigm, which may involve an ablaut-change. The 3SGM's V_1 is not given much importance in Classical Arabic, as this is always an invariable *a*. Hence, as a result of the possible V_2 differences across the Classical Arabic verbal lexicon, the ablaut classes available are *a-a kataba* 'write', *a-i xariba* 'drink', and *a-u jabuna* 'act cowardly'. In Maltese, specifying both Vs within a given two-voweled verbal base may entail a different paradigmatic behaviour. In the case of the CVCCVC verbal base type, on the other hand, knowledge of the V_1 is irrelevant, in that nothing hinges upon it. Following the analysis of the two verbs in table (7), we will eventually end up with a new analysis of Maltese that does not solely consider the ablaut class, but which provides us with an analysis that brings us closer to that of Classical Arabic. It will be shown that belonging to the same ablaut class may involve further sub-classifications, and these are dependent on the nature of the stem-vowel in the IMPERF sub-paradigm.

[10] It is here interesting to see that in terms of word-forms we initially had a systematic word-form syncretism across the perfect 1^2 SG cells, which conflation then extends across both the SG and PL cells that realize PERS 1^2, when the stem-form analysis is involved.

Paradigmatic slot distribution

PERFECT		IMPERFECT	
1^2	1	SG	4
3SGM	2		
3SGF	3	PL	5
3PL[11]			

Table 8: Representing the *stem pattern class* in which all the CVCVC verbal base set members belong

The stem-form split in the imperfect sub-paradigm is based on a SG vs. PL NUM value distinction. In the perfect sub-paradigm, on the other hand, the split is less coherent, in that while the PERS 1^2 values conflate together, the 3rd PERS cells do not form a natural class that displays the same stem-form in all the cells, at least in this type of verbal base. While the stem-form in the 3SGM cell realizes the exact same features as those realized by the inflectional affixes (refer to Baerman & Corbett 2010), on the contrary, what we have across the 3SGF and 3PL cells is a morphomic stem-form that conflates these PERS, NUM, GEND, (as well as ASP) feature values together.

In terms of the schema in table (8), what differences we have across the two verbs is such that while the *laqat* paradigm has five distinct stem-forms to fill in the five-slotted paradigmatic pattern, the *ħataf* paradigm only has four. The difference is attributed to the fact that while *laqat* has a separate stem-form in all five slots, *ħataf* involves a syncretic stem-form that cuts across slots 1 and 4 (-*ħrab*-). This distinction comes about as a result of an ablaut-change in the stem-vowel of the imperfect sub-paradigm's stem-forms in the *laqat* paradigm. Instead of sharing a stem-form across the imperfect SG cells and the perfect 1^2 cells, as is the case in *ħataf*, which implies that there is no phonological motivation for the change, we get a redundant perfect 3SGM V_2 ablaut-change from *a* to *o*. From this redundant distinction across the two verbs, we end up with what we can refer to as the *a-a – a* vs. the *a-a – o a-a* ablaut class verb sub-sets, paralleling analyses of Classical Arabic verbal taxonomy, where verbs are sub-classified in terms of the imperfect stem-vowel, apart from the ablaut-class distinction classification. As a consequence, through the stem-form -*lqot* in slot 4, instead of the expected -*lqat*-, (since this has the same stem-shape as the stem-form in slot 1), the *laqat* paradigm comes to realize ASP and NUM features through its imperfect SG stem-form, unlike the morphomic stem-form in the *ħarab* paradigm. In the imperfect sub-paradigm, the ASP feature is realized both by the stem and the inflectional affixes, and hence in Baerman and Corbett's (2010) typology, the feature is shared. In the case of the NUM feature, the SG value is realized solely by the stem-form. This substantiates my analysis in §3.2 that no zero-morph analysis is being upheld in this study, but rather, the absence of any affixal exponent is taken to imply that the stem-form is the exponent itself, and contributes to the feature realizations. The PL value is realized by both the stem-form and the -*u* suffix, although the stem-form in these cells also realizes ASP, something which the -*u* does not. Recall that this is because from the segmentation analysis proposed in table (6), the -*u* is shared across the sub-paradigms, implying that the suffix only realizes NUM out of the set of features realized by the stem-form. In the imperfect sub-paradigm, the prefixes realize PERS and ASP.[12] In fact there appears to be a tendency

[11] The reason why the perfect 3SGF and 3PL cells have not been conflated, as is the case with the perfect 1^2 cells is because they do not form a natural class, and in fact this conflation need not be the case in other paradigms.

[12] The question which I leave unresolved is whether one can consider the imperfect *t*- prefix as being solely an exponent of PERS across both paradigms, and then it is solely when surfacing with an imperative principal part, instead of the perfect 3SGM one, that it then comes to surface at one edge of the stem-form instead of the other.

for the stem-form to carry more grammatical information than the individual affixal exponents, (except where the stem is morphomic) In the case of *ħataf*, the syncretic stem-form across slots 1 and 4 results in a morphomic exponence where the imperfect SG and perfect 1^2 values are realized simultaneously in a conflated manner.

The presence or absence of a redundant morphologically-induced stem-form alternation illustrated through ablaut-changes, is what renders the difference between a four-slotted or five-slotted paradigm, at least in the case of these two verbs. This additional stem-form increases the complexity of the type being discussed here, further adding to the non-canonical behaviour, in terms of our analysis within the canonical typology framework. This is because having an extra redundant stem-form moves the stem-form behaviour further away from the canonical state of inertness. At the same time, however, this results in a better mapping across the actual number of stem-forms within the paradigm and the number of slots designated by the *stem pattern class* itself. Furthermore, this can also be thought of as morphology's drive towards more coherent feature realizations, i.e. the realization of natural class features, whereby morphology's redundant intervention acts as a means with which to avoid getting morphomic exponence, which we would have otherwise had, as is the case in the *ħarab* paradigm. By this simple comparison across these two verbs, apart from illustrating the intricate nature of the morphological component, which is rather loosely related with phonology, and is that which conditions the further gradience away from the canon, we have seen that having the same phonological properties does not entail sameness, in terms of paradigmatic behaviour; hence a divergence from what one expects to be the case across lexemes of the same type. Furthermore, a *stem pattern class* membership does not entail that all members involve the same stem-form alternations. It is rather the organisations of these stem-forms in the pattern's designated slots that render their co-membership within the same *stem pattern class*. This divergence away from the canon, and the different behaviours across apparently same members will also be explored when comparing across the verbs *mar* 'go' and *żar* 'visit' below.

4.1.2. Comparing across CV:C verbal-based verbs

The verbs, *mar* 'go' and *żar* 'visit', while displaying the same phonological properties, as shown in table (9), have conflicting statuses in traditional grammar.

Morphosyntactic feature values	*mar* 'go'		*żar* 'visit'	
PERF	PERFECT	IMPERFECT	PERFECT	IMPERFECT
1SG	mor-t	m-mūr	żor-t	n-żūr
2SG	mor-t	t-mūr	żor-t	ż-żūr
3SGM	mār	j-mūr	żār	j-żūr
3SGF	marr-et	t-mūr	żār-et	ż-żūr
1PL	mor-na	m-morr-u	żor-na	n-żūr-u
2PL	mor-t-u	t-morr-u	żor-t-u	ż-żūr-u
3PL	marr-u	j-morr-u	żār-u	j-żūr-u

Table 9: The paradigms for *mar* and *żar*

Under a consonantal-root based account, both verbs are treated without distinction, and are classified as *weak*-hollow verbs, i.e. having a *weak* consonant in their underlying consonantal-root account, *żar* is traditionally said to involve a *weak-j* medial radical, and *mar* a *w*. Sutcliffe (1936: 138) however also treats *mar* as irregular, saying that it does not behave like the rest of its class. The notion of irregularity in the grammars that follow Sutcliffe, such as that of Borg & Azzopardi-Alexander (1997), and pedagogical grammars, differs from that applied in Sutcliffe, and thus, *mar*, in these grammars, is not regarded as *irregular*, and is classified with the rest of the CV:C verbal base class, as we will also be

referring to it here. From this surface-based representation, what we find is that *mar* does in fact pattern distinctly from *żar* in terms of its paradigmatic stem-form behaviour. Having said this, it will not be considered as *irregular*, as Sutcliffe considers it to be. Rather, it is only the case that it displays exceptional behaviour in its verbal base set.

If we consider these two verbs' paradigms, we see that while the pattern of stem-form alternations in *mar* is the same as that of the verbs of the CVCVC verbal base set in table (8), particularly patterning the *laqat* sub-type, with five stem-forms fitting in the five-slotted organisation of stem-form alternations, in the case of *żar*, the stem pattern employed is shown in table (10) below, which involves an invariable stem-form in the imperfect sub-paradigm, and a PERS-based split in the perfect sub-paradigm.

Paradigmatic slot distribution

PERFECT		IMPERFECT	
1^2	1	SG	
3	2		3
		PL	

Table 10: Representing the stem pattern across the CV:C verbal base set

What we see therefore is that in the *żar* paradigm, although the syllable-structure across the 3rd PERS cells in the perfect sub-paradigm and in the imperfect sub-paradigmatic cells is the same: CV:C, once again we see a morphological effect, such that, an ablaut-change is involved, where the perfect 3SGM vowel redundantly changes to *ū*. As was the case in the ablaut-changes across the perfect 1^2 and imperfect SG cells in the *laqat* paradigm, the change renders a more feature coherent stem-form realization, instead of the morphomic exponence we would have otherwise had. Thus while the stem-form *żār* realizes perfect 3rd PERS, *żūr* realizes imperfect ASP and SG NUM. The ASP feature is hence realized by multiple exponents, as this is also the function of the prefixes, which also carry a PERS feature (refer to Camilleri *forthcoming*). The ablaut-changes taking place within the paradigm for *mar* parallel those discussed for *laqat*, except that in terms of stem-shapes we have a heteroclite paradigm, where the stem-shape of the perfect 1^2 and 3SGM cells, as well as the imperfect SG, are the ones expected for a CV:C verbal base, whereas the CVCC stem-shape across the perfect 3SGF^3PL and imperfect PL cells patterns with CVCVC verbal base types that involve a resonant as their second stem-form consonant.

Mar 'go' is the only exception in this CV:C verbal-based set, and there is no available synchronic explanation as to why it patterns differently. In doing so, it is still not treated as *irregular*, unlike Sutcliffe's (1936) treatment. The reason for this is because *irregularity* is here considered to have to do solely with when a given stem pattern class only involves one lexeme as its member. Consequently, a lexeme is *irregular* if it displays a unique stem pattern class which differs from the other typical stem pattern classes. For this reason, *mar* is not *irregular* in this regard, since it patterns with what appears to be the most common stem pattern class in the language, when one considers all that we have in the Semitic verbal data. The complexity provided here is to show that it is not only the case that we may have a different organisatory pattern within the same stem pattern class in which *laqat* and *ħataf* are co-members. Rather, what additional complexity we have here across *mar* and *żar* is that although phonologically identical and belonging in the same verbal base set, these verbs participate in distinct stem pattern classes altogether. *Mar* patterns with *laqat*, as in the stem pattern displayed in table (9), whereas *żar* displays its own pattern, as in table (10). Morphological complexity is thus manifest rather clearly when different behaviours are present across verbs with phonological sameness. Furthermore, it is interesting to see that a pattern of stem-form alternations need not cross-classify a given verbal base type, and a given stem pattern

class or sub-pattern of stem-alternations internal to that class may cut across different verbal bases.

4.2. Overabundance

An additional dimension to the canonical paradigm that results in a further divergence from the canon, is the phenomenon of *overabundance*, given most prominence in Thornton (2010, 2011).[13] This non-canonical phenomenon is present when there are 'two or more forms realizing the same cell in an inflected paradigm' (Thornton, 2011: 362). In a canonical paradigm one expects to find one word-form filling in a paradigmatic cell. From this definition we see that her focus is mostly on word-form overabundance. In the account provided here it will be shown that Maltese illustrates a case of overabundant word-forms that are derived as a result of stem-form overabundance.[14] This means that there are different patterns of overabundance, and such patterns in Maltese can combine in different ways, resulting in different word-forms altogether. Table (11) below provides a representative but non-exhaustive illustration of stem-form overabundance in two different verbal base types: the CVCC verbal base, represented by *ħass* 'feel', and the CV:C verbal base, represented by *sam*. Note that the stem-overabundance being illustrated here in the different verbal bases does not entail that all members of these verbal bases should have the same overabundant pattern. Rather, to further add to the complexity, it is somewhat of a lexical idiosyncrasy to see whether a given lexeme's paradigm within these verbal bases will actually involve overabundance or not, and whether overabundance is present in all the cells that display this phenomenon in other verbal paradigms, when it does.

Morphosyntactic feature values	*ħass* 'feel'		*sam* 'fast'	
	PERFECT	IMPERFECT	PERFECT	IMPERFECT
1SG	ħassej-t	n-ħoss	som-t ~ somej-t	n-sūm
2SG	ħassej-t	t-ħoss	som-t ~ somej-t	s-sūm
3SGM	ħass	j-ħoss	sām	j-sūm
3SGF	ħass-et	t-ħoss-u	sām-et	s-sūm
1PL	ħassej-na	n-ħoss-u	som-na ~ somej-na	n-sūm-u
2PL	ħassej-t-u	t-ħoss-u	som-t-u ~ somej-t-u	s-sūm-u
3PL	ħass-u ~ ħassē-w	j-ħoss-u	sām-u ~ samē-w	j-sūm-u

Table 11: The overabundant paradigmatic cells in the paradigms for *ħass* and *sam*

If we consider the overabundance in *ħass* we see that the target for overabundance is the perfect 3PL cell. At a glance we can already see that this targeting is morphologically-conditioned, in the sense that one cannot explain why the availability of overabundance does not target the imperfect PL stem-form as well, considering that the same stem-shape and the same *-u* suffix is involved. Therefore, while we get *ħassu* ~ *ħassew* we do

[13] Cappellaro (2010, 2012), has also worked on overabundance, where however she mostly focuses on overabundance in Italian.

[14] Although word-form overabundance within the paradigm will not be considered here, Maltese does allow for this, as shown in the the three-fold possibility in the IMPERF PL cells, of the paradigm for *marad* 'be sick', for example: j-i-mird-u (most common form) ~ j-i-mord-u (9 *google* hits) ~ j-o-mord-u (5 *google* hits), where although the latter two appear to be the least common forms, assuming that the *google* hit numbers may be taken as representative of their use in spoken and written language, these are still forms available at the native speakers' disposal. While the last two forms do not involve stem-form overabundance, they differ on the basis of the formative vowel, *o* vs. *i*, which in turn results in word-form overabundance. The difference across the first and last two word-forms displays a case of stem-form overabundance, similar to what will be discussed here.

not get *jħossu ~ *jħossew*, even if nothing hinders this from taking place. From the stem-alternation internal to the perfect PL cell, one may wish to argue that the alternation is not as redundant as one may want to assume. One may want to say that the trigger is phonological, where phonology tries to adjust the stem-shape, in turn resulting in the allomorphic suffix-form changes from *-w* to *-u*. Although re-shaping is required, as we cannot have *CCC cluster, (**ħassw*), a lengthened V: is inserted. While the allomorphic difference is phonologically-conditioned, this phonological conditioning is only triggered by the same redundant morphological allomorphic change.

Within the paradigms in table (11) we have a number of divergences taking place. As a result of the presence of overabundance, additional stem-forms are introduced, which in turn render a different stem pattern altogether, where from a three-slotted stem pattern for both the CVCC and CV:C verbal bases, we end up with an extension/expansion of the number of paradigmatic slots, illustrating another drive towards further non-canonicity. All this therefore also implies that the stem pattern is not fixed. Rather, it changes and varies. The paradigm for *ħass*, when this does not involve stem-form overabundance, patterns with that of *żar* 'visit' in table (9), shared by most of the members of the CV:C verbal base set. When overabundant cells figure in our paradigmatic analysis, we get the pattern in table (12), whereby the stem-form in the 3rd PERS cells in the perfect sub-paradigm does not solely realize the PERS feature, but also comes to realize NUM, hence adding another feature that is realized by the stem-form, which is however shared with the suffixal NUM-realizing material, when present. Thus, it is not only the case that we have an additional stem-form, adding to the non-canonical behaviour in a rather redundant manner, but furthermore, we also see that the stem-form comes to realize additional grammatical information. This is in fact what we also get in the case of *sam*, even though this belongs in a distinct verbal base set and involves additional overabundant cells.[15]

Paradigmatic slot distribution

PERFECT		IMPERFECT	
1^2	1	SG	
3SG	2		4
3PL	3	PL	

Table 12: Representing the new stem pattern of feature value organisation that results as a consequence of the non-canonical overabundance present in the paradigm

What I wanted to display, particularly through the paradigm for *sam* is that the set of verbs that pattern with it display two paradigmatic instances of stem-form overabundance; one in the perfect 3PL cell, as is also the case with *ħass*, along with an additional overabundant stem-form across the perfect 1^2 cells. It is somewhat interesting, and perhaps also pointing towards a morphologically complex network, to see that across the perfect 1^2 cells, what we have is a morphologically-triggered stem-extension whose requirement cannot be explained phonologically, since unlike what happens in the perfect 3PL cell, allomorphy was involved, in turn resulting in a partial phonological conditioning, as explained above. This redundant *ej* stem-form extension may itself well be the result of the form's analogous modelling based on other paradigms and verbal bases, which need not concern us here. Note that apart from being a clear illustration of a morphologically-conditioned stem-form overabundance, it also illustrates the way in which the morphological component appears to distinguish across the paradigmatic cells, such that while overabundance in the perfect 3PL cell results in a

[15] Note that the shifting of the paradigm's stem pattern class is not a rare and one off occurrence across verbal paradigms in Maltese. Stem pattern classes may shift freely depending on whether bound pronouns are present, and/or the paradigm's polarity.

stem pattern class shift, the stem pattern is not affected by an overabundant stem-form in the PERF 1^2 cells.

4.3. Summary

This section has shown that morphological complexity exists independently of phonological sameness and independent of morphology's own requirement and stipulations, as shown through the redundant stem-form extensions, and the ablaut-changes across stem-forms, leading to overabundance which only targets specific cells, even though other cells could have also been targeted, but in fact, are not. Furthermore, while stem patterns cut across distinct verbal base types, same verbal-based members may pattern distinctly, either by belonging to a separate stem pattern class altogether, or by involving a distinct stem-form realization of the same stem pattern class, as shown to be the case across *laqat* and *ħataf*. From the data it was also shown that having overabundance in the different cells results in different behaviours. The stem pattern is only affected when we have overabundance in the 3PL cell, which makes all of this further morphologically-complex, in that, overabundance-induced stem pattern class shifts are only related with a particular cell, and not with all of the overabundant cells present in the paradigm.

5. The Maltese *binyanim* system

This section considers the *binyanim* system and illustrates an instance of canonical divergence that takes place within it, which just as with the paradigmatic complexity described above, has never been discussed before in the literature on Maltese. The presence of a *binyanim* system, i.e. the templatic construction of verb-forms, is what has long characterised Maltese as a Semitic language along with its genealogical descendent Arabic (Comrie, 2009). The *binyanim* system has traditionally been considered as an illustration of a derivational morphological system.[16] The aim of this section is to present data exhibiting the phenomenon of heteroclisis (§5.1), and the interaction of heteroclisis along with overabundance, in (§5.2), which introduce other diverse non-canonical behaviours in the language. Heteroclisis is 'the property of a lexeme whose inflectional paradigm contains forms built on stems belonging to two or more distinct inflectional classes' (Stump, 2006, p. 278). Through an illustration of such a morphologically complex instance, a challenge to derivational accounts of the *binyanim* system is provided. This problematic issue is raised when the inflectional paradigm of an idiosyncratic lexeme in a given *binyan* may in fact involve word-forms from another *binyan*, either in the imperfect sub-paradigm, or in different cells within the perfect sub-paradigm. We will see that the ASP-cloven paradigm, i.e. a stem-alternation that splits on the basis of an ASP-feature, parallels an instance that takes place across the Hebrew *binyanim* system for the verb *approach*, as mentioned in Stump (2006, p. 314). Following these paradigmatic accounts, we will then see what effects, if any, there will be on the syntax, when this complexity interacts with argument-structure alternations themselves.

5.1. Inflection across *binyanim*

The first paradigm which we will be dealing with is that of *sieħ* 'call someone', found in the dialect of Naxxar, but obsolete in the Standard variety. From table (13) below, we see that the perfect sub-paradigm involves stem-forms related with the Ist *binyan* lexeme

[16] While we will here not be delving in this argument, for a more detailed account of the system in Maltese, the reader is referred to Borg (1988), Borg & Mifsud (1999), Hoberman & Aronoff (2003), and Spagnol (2011).

sieħ, while the imperfect sub-paradigm involves semi-suppleted stem-forms that belong to the IInd *binyan* counterpart *sejjaħ* 'call out to someone'. The variation across these two forms, under a traditional derivational account, is analysed as the formation of distinct lexemes. The *binyanim* variation, in our analysis here is considered as a mere morphophonological difference, Ist *binyan* CV:C stem-form may alternate with a IInd *banyan* CVCCVC stem-form, which allomorphy results in a different derivationally related lexeme. When considering the imperfect 1SG cell for example, one sees that there is no phonological explanation as to why the form **nsieh* is not possible, at least synchronically. It may have existed in earlier phases of the language, but became obsolete, paving the path for the IInd *binyan* stem-forms to take over. What we have here is a case whereby IInd *binyan* stem-forms fit inside the Ist *binyan* paradigm. In the verbal instance that will follow, we will have the opposite taking place, where Ist *binyan* forms are fitting within a IInd *binyan* paradigm.

Morphosyntactic features values	*sieħ* 'call someone' PERFECT	IMPERFECT
1SG	siħ-t	n-sejjaħ ~ *n-sieħ
2SG	siħ-t	s-sejjaħ
3SGM	sieħ	j-sejjaħ
3SGF	sieħ-et	s-sejjaħ
1PL	siħ-na	n-sejħ-u
2PL	siħ-t-u	s-sejħ-u
3PL	sieħ-u	j-sejħ-u

Table 13: The paradigm for dialectal *sieħ* 'call someone'

Before discussing what is going on in the dialectal paradigm of *sieħ* 'call someone', tables (14-15) represent the respective Ist *binyan* and IInd *binyan* paradigm of the verb *bies* 'kiss', which patterns in the same verbal base as *sieħ*, i.e. a CV:C verbal base, to show what one actually finds in a non-heteroclite paradigm of lexemes of the same verbal base type within the same *binyan*.

Morphosyntactic features values	*bies* 'kiss someone' PERFECT	IMPERFECT
1SG	bis-t	n-bus
2SG	bis-t	t-bus
3SGM	bies	j-bus
3SGF	bies-et	t-bus
1PL	bis-na	n-bus-u
2PL	bis-t-u	t-bus-u
3PL	bies-u	j-bus-u

Table 14: The paradigm for Ist *binyan bies* 'kiss someone'

Morphosyntactic features values	*bewwes* 'kiss' PERFECT	IMPERFECT
1SG	bewwis-t	n-bewwes
2SG	bewwis-t	t-bewwes
3SGM	bewwes	j-bewwes
3SGF	bews-et	t-bewwes
1PL	bewwis-na	n-bews-u
2PL	bewwis-t-u	t-bews-u
3PL	bews-u	j-bews-u

Table 15: The paradigm for IInd *binyan bewwes* 'kiss'

As a result of the heteroclite paradigm in table (13), that involves stem-forms from distinct verbal bases (as a consequence of belonging to the different *binyanim*); CV:C for the Ist *binyan sieħ* and CVCCVC for the IInd *binyan sejjaħ*, we do not solely end up with

heteroclite paradigmatic forms, but we also end up with a heteroclite stem pattern. The perfect sub-paradigm's pattern of stem-form alternations is what one would expect to find given a CV:C verbal base in the language, as displayed for *żar* 'visit' (table 9), *sam* 'fast' (table 11), and *bies* 'kiss' (table 14) above, where the sub-paradigm involves a neat PERS-based stem-form split. In the imperfect sub-paradigm, on the other hand, instead of the expected invariable stem-form, as is the case for the CV:C verbal base set (with the exception of *mar* 'go', as mentioned above), we have a pattern of stem-form alternations that is the same as that which we had in the CVCVC verbal base set in table (8), which also extends for *mar* and other non-CVCVC-verbal-based verbs that pattern in the same way. The pattern of stem-form alternations in the IInd *binyan* also happens to be the IInd *binyan* pattern of CV:C Ist *binyan* counterparts, as displayed in table (15).[17] The heteroclite stem pattern that results is illustrated in table (16) below. It exhibits a rather neat feature-based split within both sub-paradigms, showing a 1^2 vs. 3rd PERS-based distinction in the perfect sub-paradigm and a SG vs. PL NUM-based distinction in the imperfect sub-paradigm, all embedded within an ASP morphosemantic split. Recall that we would not have had this pattern, were it not due to the presence of the morphologically complex and divergent illustration of heteroclisis in this paradigm. In resulting in more feature-coherent stem-forms than morphomic ones, heteroclisis also results in the addition of a stem-form within the paradigm, when one compares *sieħ* with the rest of the CV:C verbal base set, at least if we keep excluding *mar*. All this results in additional non-canonicity, not only on the basis of the way it includes stem-forms from other paradigms, but also in that it has an additional stem-form in the imperfect sub-paradigm, if we measure complexity on the basis of how many stem-forms exist in a lexeme's paradigm, and considering that in theory nothing hinders a CV:C-shaped invariable stem-form across the imperfect sub-paradigm cells. Furthermore, the additional stem-form in the imperfect sub-paradigm does not solely realize ASP grammatical information, but also NUM.

Paradigmatic slot distribution

PERFECT		IMPERFECT	
1^2	1	SG	3
3	2		
		PL	4

Table 16: Representing the heteroclite stem pattern for the heteroclite *sieħ* paradigm

What one needs to add here is that the difference across the Ist and IInd *binyan* word-forms *sieħ* and *sejjaħ* is solely formal, in that there is really no syntactic or semantic distinction or argument-structure differences across these two verbal forms, (though this may not have been the case diachronically). This synchronic state of affairs has resulted in an optionally overabundant perfect sub-paradigm. It is important to mention here that it is not because we have no difference in the argument-structure alternation that we get overabundance. As we will see in the paradigm for *ħabbeb* 'cause to love/befriend' below in §5.2, we still get overabundant cells, even though we have an argument-structure distinction across the stem-forms from the different *binyanim* being used. This hence presents us with a paradigm in (17), which Corbett (2011) would refer to as an instance of a higher order exceptionality.

[17] Under this account here, no *binyan* distinction needs to be made across Ist *binyan* CVCVC verbs and IInd *binyan* consonant-final verbal bases, considering that CVCVC and CVCCVC verbal bases share the same pattern of stem-form alternations, and hence, irrespective of consonantal root representations or templatic formations, morphology merely considers phonological bases, and how to get to the paradigm, accordingly.

Morphosyntactic feature values	*sieħ* 'call someone' PERFECT	IMPERFECT
1SG	siħ-t ~ sejjaħ-t	n-sejjaħ
2SG	siħ-t ~ sejjaħ-t	s-sejjaħ
3SGM	sieħ ~ sejjaħ	j-sejjaħ
3SGF	sieħ-et ~ sejħ-et	s-sejjaħ
1PL	siħ-na ~ sejjaħ-na	n-sejħ-u
2PL	siħ-t-u ~ sejjaħ-t-u	s-sejħ-u
3PL	sieħ-u ~ sejħ-u	j-sejħ-u

Table 17: The overabundant paradigm for dialectal *sieħ*

Just as exhibited in §4.2 for *ħass* 'feel' and *sam* 'fast', this overabundance results in a different pattern of stem-form alternations based on the way in which the distinct morphosyntactic feature values are conflated across stem-forms. Thus we see that in parallel to the PERS-based stem-split in the perfect sub-paradigm, we also have an alternating CVCVC-based stem pattern of alternation, paralleling that which was displayed in table (8), involving: 1^2, 3SGF^3PL and 3SGM feature value conflations. From this paradigmatic stage, one may want to say that just as in the Standard variety, the I[st] *binyan* form has become obsolete, one may hypothesise that this overabundant stage is intermediary and is the stage that precedes the actual loss of the I[st] *binyan* form altogether in the dialect, which would result in the paradigm's levelling, where it becomes entirely a II[nd] *binyan* paradigm.

5.2 Heteroclisis, overabundance and argument-structure alternations

What follows below is the overabundant paradigm of the II[nd] *binyan* verb-form *ħabbeb* 'cause to love/befriend' (table 18). For expository purposes, the I[st] *binyan* counterpart *ħabb* 'love' is provided below in table (19).

Morphosyntactic feature values	*ħabbeb* 'cause to love someone' PERFECT	IMPERFECT
1SG	ħabbib-t II ~ ħabbej-t I	n-ħabbeb II ~ *n-ħobb I
2SG	ħabbib-t ~ ħabbej-t	t-ħabbeb
3SGM	ħabbeb II	j-ħabbeb
3SGF	ħabb-et I/II	t-ħabbeb
1PL	ħabbib-na II ~ ħabbej-na I	n-ħabb-u
2PL	ħabbib-t-u ~ ħabbej-t-u	t-ħabb-u
3PL	ħabbē-w I/II[18]	j-ħabb-u

Table 18: The overabundant paradigm of the II[nd] *binyan ħabbeb* 'cause to love someone'

Morphosyntactic feature values	*ħabb* 'love someone' PERFECT	IMPERFECT
1SG	ħabbej-t	n-ħobb
2SG	ħabbej-t	t-ħobb
3SGM	ħabb	j-ħobb

[18] The stem-form in this cell illustrates another occurrence of interesting morphological complexity, where while I[st] *binyan ħabb* 'love' patterns just like *ħass* 'feel' in table (11), involving stem-form overabundance in the perfect 3PL cell: *ħabbu ~ ħabbew* 'they love', when it comes to the II[nd] *binyan*, it is only *ħabbew* that is used, and an overabundant stem-form is not allowed, in turn implying that the presence of overabundance, as well as a given stem-form instead of another can give morphological cues for a distinct argument-structure, which would have otherwise been ambiguous. Furthermore, a unifying factor across the I[st] and II[nd] *binyanim* paradigms in this regard, at least when comparing across the verbs *ħabb* 'love' and *ħabbeb* 'cause to love/befriend', is that in the presence of an attached pronoun, in the perfect 3PL cell, it is only the stem-form *ħabbē-* that can be used, and not *ħabb-*; *ħabbew-h ~ *ħabbu-h* 'they loved him'.

3SGF	ħabb-et	t-ħobb
1PL	ħabbej-na	n-ħobb-u
2PL	ħabbej-t-u	t-ħobb-u
3PL	ħabb-u ~ ħabbē-w	j-ħobb-u

Table 19: The paradigm of the Ist *binyan ħabb* 'love'

What we can see from the overabundant paradigm in table (18) is that if we exclude the overabundant state of affairs, this IInd *binyan* paradigm already involves stem-forms/word-forms which are shared with those in the Ist *binyan* paradigm counterpart, as one can see when comparing table (18) and (19), facilitated by the use of the roman numerals I and II, next to the word-form in the different cells, as is the case with the morphologically ambiguous forms in the 3SGF and 3PL cells. What happens as a result of stem-form overabundance is that all the cells in the perfect sub-paradigm, with the strict exception of the 3SGM form, which bears the verbal-base-stem-shape related with the given IInd *binyan*: CVCCV(C), we get quasi-levelling in the use of the Ist *binyan* stem-/word-forms across all the paradigmatic cells, whereby when the overabundant Ist *binyan* forms are used in the IInd *binyan* paradigmatic context, morphological ambiguity is increased. It is once again interesting to see that the presence of overabundant stem-forms across the perfect 1^2 cells, as was also shown to be the case for the overabundant *sam* 'fast' paradigm in table (11), do not result in a stem pattern class shift. Recall that in the overabundant *sieħ* 'call someone' paradigm in §5.2 we had a stem pattern class shift as a consequence of having all perfect paradigmatic cells being the target for overabundance, also mentioned in §4.2.

In the case of *ħabbeb*, considering that as mentioned earlier in §5.1, the distinction across the Ist and IInd *binyanim* verb-forms *ħabb* ~ *ħabbeb*, unlike *sieħ* ~ *sejjaħ*, involve an argument-structure distinction. The Ist *binyan* predicate takes a SUBJ and an OBJ as its subcategorised grammatical functions, while the IInd *binyan* predicate takes SUBJ, OBJ and OBL grammatical functions. What we end up with, as a result of this morphological complexity, is the situation illustrated in sentences (20) below, where it is now not the morphological forms which are giving us the argument-structure distinction, but it is rather the syntax itself which now helps disambiguate morphologically ambiguous forms. From a robust morphological system that brings about argument-structure alternations, (although of course not necessarily always the case), a larger weight on syntax has now to be imposed.

20. a. *Habbe-w* *lil* *xulxin*
loved-3.PL ACC each other
They loved each other *ħabb* 'love' <SUBJ, OBJ>
*They caused to love each other

b. *Habbe-w-hom* *ma'* *xulxin*
loved-3.PL-3PL.ACC with each other
They made them love each other *ħabbeb* 'cause to love' <SUBJ, OBJ, OBL>
*They loved them each other

The 3rd PERS PL form in (20a-b) is morphologically ambiguous, an ambiguity that is a property of the Ist *binyan* CVCC-derived IInd *binyan* verbal bases. It is only the nature of the argument-structure expressed in the syntax, which reflects the verb-form interpretation. When the additional morphological complexity manifest by overabundance is added on top of this, as is the case in the perfect PERS 1^2 cells, we get a similar effect, when the word-form common across both *binyanim* is used:

c. *Habbej-t* *lit-tfal*
loved-1SG ACC.DEF-children

I loved the children — *ħabb* 'love' <SUBJ, OBJ>
*I caused to love Mary

d. *Ħabbej-t/ħabbib-t* *lit-tfal* *ma'* *xulxin*
love.CAUSE-1SG ACC.DEF-children with each other
I made them love each other — *ħabbeb* 'cause to love' <SUBJ, OBJ, OBL>
*I loved Mary with each other

Therefore, when the semantic interpretation of the causative II[nd] *binyan* verb-form *ħabbeb* is intended, when the perfect 1[st] PERS SG verb-form *ħabbejt* is used instead of *ħabbibt*, it is not the morphological form that is denoting the syntactic valence of the verb, but it is rather the presence/absence of a preposition-headed constituent that functions as an OBL grammatical function, that in fact provides the semantic interpretation of the morphological form.

6. Conclusion

The study aimed to show that Maltese displays a number of interesting morphologically complex phenomena, with data also illustrating the interactions of these. All the occurrences of canonical divergence were here interpreted and understood as a complexity that is solely derived from an autonomous morphological component. This was particularly manifest through phenomena that result out of no obvious phonological motivations. It was shown that a stem pattern class need not be as complex, if ablaut-changes were not to be involved, as for most of the cases, these are derived out of a number of interacting hierarchically-ordered set of phonological constraints. Furthermore, different syllable structures redundantly result when overabundance is involved, as is also the case with heteroclisis, where nothing can synchronically account for why a non-heteroclite form is not present in the imperfect sub-paradigm. Moreover, an independence from phonology was also shown to be the case through the different paradigmatic behaviours of verbs which are in fact grouped together under the same verbal base classification as a result of their phonological make-up.

The fact that such members differ does not only reflect this morphological complexity, but it also aims to show that looking at an underlying representation, rather than at surface structure paradigms, the truth of what actually goes on in the paradigms, similar to the surface phonological classifications do not at times contribute to homogeneous morphological outcomes. Differences between the *laqat* and *ħataf*, and the *mar* and *żar* set of verbal base types, particularly illustrated this point. The former participate in the same stem pattern class, but the way the different paradigmatic slots are realized differs, whereas *mar* and *żar* were shown to belong to the same verbal base but differ in their stem pattern class membership, such that *mar* patterns with *laqat*, showing that the phenomenon of stem pattern formations cuts across verbal base types, and is not restricted to the member's phonological structure. The availability of stem-form overabundance (§4.2) showed us that morphology plays an important role in actually determining which cells are targeted and whether such overabundance needs involve a stem pattern class shift. By the availability of such shifts we see that a lexeme need not be a member of just one stem pattern class, and that there is some level of flexibility internal to the stem-form behaviour. It was also highlighted that stem allomorphy and the morphologically-induced conditions that change the stem-form, constitute another dimension to Maltese inflection that is non-concatenative, which coexists with the concatenative affixal exponents. Furthermore, all divergence from that which is canonical suggests that in Maltese, the morphological component is an important innovation that distances the Maltese paradigm from what one expects to find under a canonical account. This was shown to be the case not only through the increase in the number of stem-

forms, but also by having stem-forms that carry grammatical information. Additional complexity and non-canonical behaviour, via redundant ablaut-changes and overabundant forms, result in stem-forms that involve more feature-coherent realizations, moving the stem-form further away from the inert paradigmatic requirement and the lexical material function, adding to further divergence from the canon.

References

Aronoff, Mark 1994. *Morphology by itself: Stems and inflectional classes.* Massachusetts: MIT Press.

Baerman, Matthew, Brown, Dunstan, & Corbett, Greville G. (eds.) 2005. *The Syntax-Morphology Interface: A study of syncretism.* Cambridge: CUP.

Baerman, Matthew, Corbett, Greville G., & Brown, Dunstan (eds.) 2010. *Defective paradigms*. Oxford: OUP.

Baerman, Matthew, & Corbett, Greville G. (2012) Stem alternations and multiple exponence. Paper presented at the workshop 'Stems in Inflection and Words'. To appear in *Word Stricture.*

Booij, Geert 1996. Autonomous morphology and paradigmatic relations. In G. Booij, & J. van Marle (eds.) *Yearbook of Morphology* (pp. 35-53). Dordrecht: Kluwer.

Booij, Geert 2009. Morphological analysis. In B. Heine, & H. Narrog (eds.), *The Oxford Handbook of Grammatical Analysis* (pp. 563-589). Oxford: OUP.

Borg, Albert J. 1981. *A study of aspect in Maltese.* Ann Arbor: Karoma Publishers.

Borg, Albert J. 1988. *Ilsienna.* Malta: Ħas Sajjied.

Borg, Albert J., & Azzopardi-Alexander, Marie 1997. *Maltese*. New York: Routledge.

Camilleri, Maris 2012. Stem patterns across Maltese verbal paradigms. Paper presented at the *IMM 15*, February, Vienna.

Camilleri, Maris forthcoming. Multiple exponence: The case of stems and affixes in Maltese. *Workshop on the selection and representation of morphological exponents*, June, Trømso.

Cappellaro, Chiara 2010. Pronominal overabundancy in Italo-Romance. Paper presented at the *IMM 14*, May, Budapest.

Cappellaro, Chiara 2012. The one form-one meaning principle and the emergence of overabundance in noun paradigms: Evidence from Rhaeto-Romance. Paper presented at the *IMM15,* February, Vienna.

Comrie, Bernard 2009. Maltese and the World Atlas of Language Structures. In: In B. Comrie, R. Fabri, E. Hume, M. Mifsud, T. Stolz, & M. Vanhove (eds.) *Introducing Maltese Linguistics* (pp. 3-12). Amsterdam/Philadelphia: John Benjamins

Corbett, Greville G. 2000. *Number.* Cambridge: CUP.

Corbett, Greville G. 2005. The canonical approach in typology. In Z. Frajzyngier, A. Hodges, & D. S. Rood (eds.), *Linguistic diversity and language theories* (pp. 25-50). Amsterdam/Philadelphia: John Benjamins.

Corbett, Greville G. 2007a. Canonical typology, suppletion and possible words. *Language, 83*, 8-42.

Corbett, Greville G. 2007b. Deponency, syncretism, and what lies between. *Proceedings of the British Academy*, *145*, 21-43.

Corbett, Greville G. 2009. Canonical Inflectional Classes. In: F. Montermini/G. Boyé/J. Tseng (eds.), *Selected proceedings of the 6th Decembrettes* (pp. 1-11). Somerville: Cascadilla Proceedings Projects.

Corbett, Greville G. 2011a. The prenumbra of morphosyntactic feature systems. *Morphology 21,* 445-480.

Corbett, Greville G. 2011b. Higher order exceptionality in inflectional morphology. In: H. J. Simon & H. Wiese (eds.) *Expecting the unexpected: Exceptions in grammar* (Trends in Linguistics. Studies and Monographs 216), (107-126). Berlin: De Gruyter Mouton. [With comments by Stephen Anderson, "An I-language view of morphological 'exceptionality': comments on Corbett's paper", (127-134), and reply "Exceptions and what they tell us: reflections on Anderson's comments", (135-137).]

Corbett, Greville G., & Baerman, Matthew 2006. Prolegomena to a typology of morphological features. *Morphology 16,* 231-246.

Corbett, Greville G., Baerman, Matthew 2010. "Deep Irregularity": When the expected lexical splits follow the wrong pattern. Paper presented at the *14th International Morphology Meeting*, Budapest, Hungary.

Fabri, Ray 1995. The tense and aspect system of Maltese. In R. Thieroff (ed.) *Tempussysteme in europaeischen Sprachen II* (pp. 327-343) Niemeyer: Tübingen.

Fabri, Ray 2009. Stem allomorphy in the Maltese verb. In T. Stolz (ed.), *Ilsienna – Our Language, 1,* (pp. 1-20). Amsterdam/Philadelphia: John Benjamins.

Gauci, Phyllisienne, & Camilleri, Maris 2011. Syncretism and morphological complexity in Maltese: The case of the Dual. Paper presented at the *3rd International Conference on Maltese Linguistics,* April, Malta.

Hetzron, Robert 1997. Preface. In R. Hetzron (ed.), *The Semitic Languages*. London: Routledge.

Hoberman, Robert D., & Aronoff, Mark 2003. The verbal morphology of Maltese: From Semitic to Romance. In J. Shimron (ed.) *Language Processing and Language Acquisition in Languages of Semitic, Root-Based Morphology,* (pp. 61–78). Amsterdam/Philadelphia: John Benjamins.

Kenstowicz, Michael 2005. Paradigmatic uniformity and contrast. In Downing, L. J., Hall, A., & Raffelsiefen, R. (eds.) *Paradigms in Phonological Theory* (pp. 145-171). Oxford: OUP.

Lowenstamm, Jean 2011. The phonological pattern of phi-features in the perfective paradigm of Moroccan Arabic. *Brill's Annual of Afroasiatic Languages and Linguistics, 3*(1), 140-201.

Maiden, Martin, Smith, John C., Goldbach, Maria, & Hinzelin Marc-Oliver (eds.) 2011. *Morphological Autonomy: Perspectives from Romance Inflectional Morphology* (p. 358-381). Oxford: OUP.

Mifsud, Manwel 1995. *Loan Verbs in Maltese: A descriptive and comparative study.* Leiden: Brill.

Müller, Stefan 2009. A Head-Driven Phrase Structure Grammar for Maltese. In B. Comrie, R. Fabri, E. Hume, M. Mifsud, T. Stolz, & M. Vanhove (eds.) *Introducing Maltese Linguistics* (pp. 83-112). Amsterdam/Philadelphia: John Benjamins.

Nakov, Preslav, Bonev, Yury, Angelova, Galia, Cius, Evelyn, & von Hahn, Walther 2004. Guessing morphological classes of unknown German nouns. In N. Nicolov (ed.), *Recent Advances in Natural Language Processing III* (pp. 346-356). Amsterdam/Philadelphia: John Benjamins.

Polinsky, Maria 2003. Non-canonical agreement is canonical. *Transactions of the Philological Society, 101*(2), 279-312.

Seifart, Frank 2005. The structure and use of shape-based noun classes in Miraña (North West Amazon). Nijmegen: Radboud University dissertation.

Spagnol, Michael 2011. A Tale of Two Morphologies: Verb structure and alternations in Maltese. Ph.D Dissertation, University of Konstanz.

Spencer, Andrew 2003. Periphrastic paradigms in Bulgarian. In U. Junghanns, & L. Szucsich (eds.), *Syntactic Structures and Morphological information* (pp. 249-282). Berlin: Mouton de Gruyter.

Stump, Gregory T. 2001. *Inflectional Morphology*. Cambridge: CUP.

Stump, Gregory T. 2006. Heteroclisis and paradigm linkage. *Language, 82,* 279-322.

Sutcliffe, Edmund F. 1936. *A grammar of the Maltese language with chrestomathy and vocabulary.* London: Oxford University Press.

Suthar, Babubhai K. 2006. Agreement in Gujarati. Philadelphia: University of Pennsylvania dissertation.

Thornton, Anna M. 2010. Towards a typology of overabundance. Paper presented at Décembrettes 7: International Conference on Morphology, University of Toulouse, December.

Thornton, Anna M. 2011. Overabundance (multiple forms realizing the same cell): a noncanonical phenomenon in Italian verb morphology. In M. Maiden, J. C. Smith, M. Goldbach, & M. Hinzelin (Eds.). *Morphological Autonomy: Perspectives from Romance Inflectional Morphology* (p. 358-381). Oxford: OUP.

Vogel, Irene 1994. Verbs in Italian morphology. In G. Booij, & J. van Marle (eds.) *Yearbook of Morphology* (pp. 219-254). Dordrecht: Kluwer.

Zaliznjak, Andrej A. 1973. O ponimanii termina 'padež' v lingvističeskix opisanijax. In: A. A. Zaliznjak (Ed.). Problemy grammatičeskogo modelirovanija. Moscow: Nauka, 53-87. [Reprinted in: Andrej A. Zaliznjak. 2002. Russkoe imennoe slovoizmenenie: s priloženiem izbrannyx rabot po sovremennomu russkomu jazyku i obščemu jazykoznaniju, 613-647. Moscow: Jazyki slavjanskoj kul´tury.]

Prediction and Generalisation in Word Processing and Storage

Claudia Marzi
Institute for Computational Linguistics"A. Zampolli"- CNR Pisa,
University of Pavia - Dept. of Theoretical and Applied Linguistics
claudia.marzi@ilc.cnr.it

Marcello Ferro
Institute for Computational Linguistics"A. Zampolli"- CNR Pisa
marcello.ferro@ilc.cnr.it

Vito Pirrelli
Institute for Computational Linguistics"A. Zampolli"- CNR Pisa
vito.pirrelli@ilc.cnr.it

1. Introduction

Word storage and processing have traditionally been modelled according to different computational paradigms, in line with the classical corner-stone of "dual-route" models of word structure assuming a sharp dissociation between memory and computation (Clahsen 1999, Di Sciullo & Williams 1987, Pinker & Prince 1988, Parasada & Pinker 1993). Even the most radical alternative to dual-route thinking, connectionist one-route models, challenged the lexicon-grammar dualism only by providing a neurally-inspired mirror image of classical base-to-inflection rules, while largely neglecting issues of lexical storage (Rumelhart & McClelland 1986, McClelland & Patterson 2002, Seidenberg & McClelland 1989). Recent psycho- and neuro-linguistic evidence, however, supports a less deterministic and modular view of the interaction between stored word knowledge and on-line processing [Baayen et al. 1997, Hay 2001, Maratsos 2000, Stemberger & Middleton 2003, Tabak et al. 2005, Ford et al. 2003, Post et al. 2008). The view entails simultaneous activation of distributed patterns of cortical connectivity encoding redundant distributional regularities in language data. Furthermore, recent developments in morphological theorising question the primacy of grammar rules over lexical storage, arguing that word regularities emerge from independent principles of lexical organisation, whereby lexical units and constructions are redundantly stored and mutually related through entailment relations (Matthews 1991, Corbett & Fraser 1993, Pirrelli 2000, Burzio 2004, Booij 2010). We endorse here such a non modular view on Morphology to investigate two basic behavioural aspects of human word processing: morphological prediction and generalisation. The investigation is based on a computer model of morphology acquisition supporting the hypothesis that they both derive from a common pool of principles of lexical organisation.

2. Background

2.1. Generalisation

Morphological generalisation is at the roots of the human ability to develop expectations about novel lexical forms, so that some words (say *plipped* as the past tense of the nonce verb *plip*) are perceived by speakers as more acceptable than other potential competitors (e.g. *plup* or *plept* for the same base). These expectations can be used to produce novel forms from familiar bases. After Berko's seminal work on children mastering *wug* words (1958), linguists have put considerable effort into trying to unravel conditions for generalisations in the morphological competence of both learners and mature speakers (Bybee and Pardo, 1981; Bybee and Slobin, 1982; Bybee and Moder, 1983). After the advent of connectionism (Rumelhart and McClelland, 1986), the question of what

structural and formal conditions affect morphological generalisations in humans was coupled with the substantially different question of whether artificial neural networks eliminate or rather sub-symbolically implement algebraic productive rules of some kind (Smolensky 1988). The ensuing debate went through controversial issues of grammar architecture, centred on the hypothesis of a sharp separation between lexicon (functionally related to storage) and rules (functionally related to processing). A recent reformulation of the problem of morphological generalisation is due to Albright and Hayes (2003): given that many morphological processes are known to be productive in limited contexts, what sort of computational mechanisms are needed to account for context-sensitive restrictions on morphological generalisations? Albright and Hayes suggest that speakers conservatively develop structure-based rules of mapping between fully-inflected forms. In the orthographic domain, a mapping pattern such as X*s* → X*ing* accounts for the word pairs *talks-talking*, *plays-playing*, *forms-forming* etc., but would wrongly yield *puts-*puting* and *gives-*giveing*. These patterns are based on a cautious inductive procedure named "minimal generalisation", according to which speakers are confident in extending a morphological pattern to other forms to the extent that i) the pattern obtains for many existing word forms and ii) there is a context-sensitive difference between those word forms and word forms that take other patterns. For example, a speaker has to induce the more specific pattern X*ts* → X*tting* to cover *puts-putting*, *sits-sitting*, *hits-hitting* etc. An important point made by Albright and Hayes is that patterns apply to similar word pairs, with word similarity being based on a context-sensitive structural mapping', rather than on a pre-theoretical notion of "variegated" analogy. Finally, the level of confidence of a speaker in the pattern is defined by the ratio between the number of forms undergoing the pattern change and the number of forms meeting the context for the pattern change to apply.

A number of interesting theoretical implications follow from Albright and Hayes' approach. Unlike traditional dual-route models of morphological competence, their minimally generalised patterns are not committed to a derivational conception of morphological generalisation, according to which rules define base-to-form mapping relations only. Patterns may underlie any pair of intra-paradigmatically related forms. This view is easily amenable to a word-and-paradigm conception of the morphological lexicon where fully inflected forms are redundantly stored and mutually related through entailment relations (Matthews 1991; Pirrelli 2000; Burzio 2004; Blevins 2006). Accordingly, the speaker's knowledge of word structure is more akin to one dynamic relational database than to a general-purpose automaton augmented with lexical storage. Nonetheless, mapping patterns adhere to a rule-like manner of stating generalisations, providing the necessary and sufficient conditions that a form must meet in order for the pattern to apply. Albright and Hayes argue that this does not have to be true for variegated analogy to apply, as, in principle, forms undergoing the same pattern change may be similar to one another in many different ways, thus going beyond the reach of a structural rule-like description of the needed context. Finally, they claim that sensitivity to context-based similarity is not a specific condition of unproductive morphological processes (as suggested by dual-route modellers), but a hallmark of any morphological pattern change. Cautious generalisation is an inherent feature of morphological productivity as such.

2.2. Prediction

Morphological prediction defines the human capacity to anticipate upcoming known words. Unlike generalisation, which refers to the ability to go beyond available evidence and compensate for gaps in lexical competence, prediction appears to functionally maximise available linguistic evidence (including, but not limited to, lexical competence) to entertain hypotheses about the upcoming flow of language input, and make language

comprehension easier and more efficient. From a more general perspective, experimental studies based on event-related potentials and eye-movement evidence, for example, show that people use prior (lexical and semantic) contextual knowledge to anticipate upcoming words (Altmann and Kamide 1999; Federmeier 2007). DeLong et al. (2005) demonstrate that expected words are pre-activated in the brain in a graded fashion, reflecting their expected probability. This provides the empirical ground to probabilistic approaches to lexical prediction and gaze planning in reading. Ferro et al. (2010) offer a computational model of the interlocked relationship between processes of lexical self-organisation and active sensing strategies for reading that exploit expectations on stored lexical representations to drive gaze planning. This can explain why the capacity to repeat non words is a good predictor of whether or not the child is likely to encounter reading problems (Baddeley and Gathercole 1992; Gathercole and Pickering 2001).

There have been attempts to explain the role of prediction in facilitating language comprehension on the basis of the argument that highly predictable words are easier to integrate into the linguistic context (e.g. because unexpected words in test experiments often violate the grammatical constraints imposed by the context itself). In fact, recent evidence goes against this simpler explanation, suggesting that predictions can be made at many different levels of language comprehension, including strong biases against perfectly grammatical but somewhat rarer or less likely lexical alternatives (Staub & Clifton 2006, DeLong et al. 2005). A more intriguing explanation comes from evidence of mirror neurons (Wilson & Knoblich 2005) pointing to the observation that perceiving other people's behaviour activates covertly imitative motor plans. The use of covert imitation to facilitate perception of other's people behaviour could occur in any domain where upcoming behaviour is at least sometime predictable and where the perceiver can also perform that behaviour. Language in general, and lexical access in particular, are cases in point.

2.3. Grammar, Memory and the Lexicon

At its core, the lexicon is the store of words in long-term memory. Any attempt at modelling lexical competence must take into account issues of string storage. In this respect, the rich cognitive literature on short-term and long-term memory processes (Miller 1956; Baddeley and Hitch 1974; Baddeley 1986, 2006; Henson 1998; Cowan 2001; among others) has had the unquestionable merit of highlighting some fundamental issues of coding, maintenance and manipulation of strings of symbols. It is somewhat surprising that the linguistic literature on lexical access and organisation, on the one hand, and the psycho-cognitive literature on memory processes on the other hand, have so far made comparatively sparse contact. This is arguably due to the strong influence of the calculator metaphor (Baayen 2007) on mainstream conceptions of the role of the lexicon in the grammar architecture. According to the metaphor, the lexicon is only storage, an inert repository of item-based, unpredictable information whose nature and structure is predetermined, and considered as relatively unproblematic. The combinatorial potential of lexical items, on the other hand, is defined by the rules of grammar, taking care of processing issues. Contrary to what is commonly held, connectionism has failed to offer an alternative view of such an interplay between storage and processing. There is no place for the lexicon in classical connectionist networks: in this respect, they seem to have adhered to a cornerstone of the rule-based approach to morphological inflection, thus providing a neurally-inspired mirror image of derivational rules.

In this paper, we entertain the substantially different view that memory plays a fundamental role in lexical modelling, and that computer simulations of memory processes can go a long way in addressing issues of lexical acquisition and processing. Recent studies of cortico-cortical evoked potentials show a functional bidirectional

connectivity between anterior and posterior language areas (Matsumoto et al., 2004), pointing to more integrated and dynamic mechanisms underlying language functioning in the brain than previously acknowledged. In addition to its well-established linguistic functions, Broca's area appears to be engaged in several other cognitive domains such as music (Maess et al., 2001), working memory and calculation, as well as action execution and understanding (Buccino et al., 2001; Fadiga et al., 2009). Taken together these results suggest the intriguing possibility that Broca's areas could provide the neural structures subserving context-dependent sequence processing in general, and that these structures shed considerable light on what we know about lexical organisation, access and productivity.

Human lexical competence is known to require the fundamental ability to retain sequences of items (e.g. letters, syllables, morphemes or words) in the working memory (Gathercole and Baddeley, 1989; Papagno et al., 1991). Speakers appear to be sensitive to frequency effects in the presentation of temporal sequences of verbal stimuli. Items that are frequently sequenced together are stored in the Long-Term (LT) Memory as single chunks, and accessed and executed as though they had no internal structure. This increases fluency and eases comprehension. Moreover, it also explains the possibility to retain longer sequences in Short-Term (ST) Memory when familiar chunks are presented. The ST span is understood to consist of only a limited number (ranging from 3 to 5 according to recent estimates, e.g. Cowan 2001) of available store units. A memory chunk takes one store unit of the ST span irrespectively of length, thus leaving more room for longer sequences to be temporarily retained. Furthermore, chunking produces levels of hierarchical organisation of the input stream: what is perceived as a temporal sequence of items at one level, may be perceived as a single unit on a higher level, to become part of more complex sequences (Hay and Baayen 2003). Finally, parts belonging to high-frequency chunks tend to resist being perceived as autonomous elements in their own right and being used independently. As a further implication of this "wholeness" effect, frequently used chunks do not participate in larger word families (e.g. inflectional paradigms).

From this perspective, generalisation and prediction can be seen as being in competition. Prediction presupposes LT entrenchment of memory chunks as a result of repeated exposure to frequent sequences of letters/segments. LT entrenchment eventually drives word recognition through anticipatory activation of frequently-activated chunks. Prediction is thus most accurate when concurrent activation of LT chunks is minimised. In information theoretic terms, this is equivalent to minimising the entropy over lexical choices. Generalisation, on the other hand, requires that the lexicon contains recurrent sub-lexical chunks, which recombine for novel words to be recognised as well-formed. This is equivalent to keeping entropy high in the lexicon, making room for novel word stimuli.

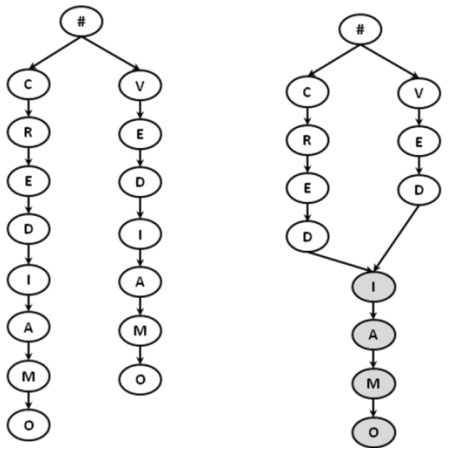


Figure 1: A word-trie (left) and a word-graph (right), for the Italian forms *VEDIAMO* ('we see') and *CREDIAMO* ('we believe').

Figure 1 shows examples of lexical structures that can account for these effects. So-called "word-tries" (left) encode symbol sequences as rooted hierarchies of labelled nodes connected through arcs, under the constraint that no node can be reached by two different descending arcs. So-called "word-graphs" (right), on the other hand, allow the same node to be reached by multiple arcs, thus using up fewer nodes in representing partially overlapping forms. For our present concerns, word-tries can be seen as encoding deeply entrenched, dedicated memory structures, whereby partially overlapping forms are nevertheless assigned independent representational resources. On the other hand, word-graphs allow for shared substrings to be assigned identical memory units. As shown below, the two graph types can in fact be conceived of as different developmental stages in lexical acquisition. In the ensuing sections, we offer a computational model of dynamic memories that can explain the emergence of such lexical structures in terms of common computational principles of self-organisation and time-bound prediction.

3. Hebbian SOMs

Kohonen's Self-Organizing Maps (*SOMs*) (Kohonen, 2001) define a class of unsupervised artificial neural networks that mimics the behaviour of small aggregations of neurons (*pools*) in the cortical areas involved in the classification of sensory data (*brain maps*). In such aggregations, processing consists in the activation of specific neurons upon presentation of a particular stimulus. A distinguishing feature of brain maps is their topological organisation (Penfield and Roberts, 1959): nearby neurons in the map are activated by similar stimuli. Although some brain maps can be pre-determined genetically, there is evidence that at least some aspects of their neural connectivity emerge through self-organisation as a function of cumulated sensory experience (Kaas et al., 1983). Functionally, brain maps are thus dynamic memory stores, directly involved in input processing, exhibiting effects of dedicated long-term topological organisation.

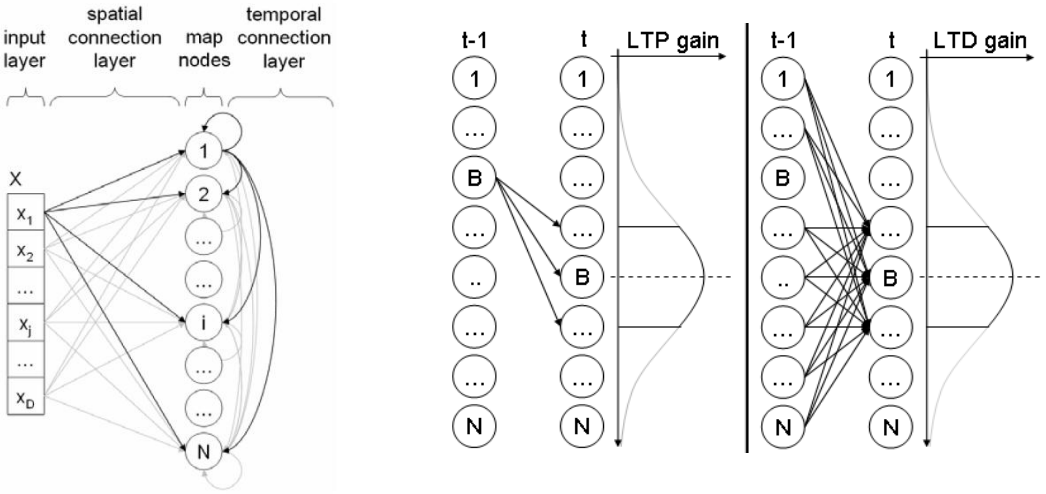


Figure 2: Left: Outline architecture of a T2HSOM. Each node in the map is connected with all nodes of the input layer. Each connection is a communication channel with no time delay, whose synaptic strength is modified through training. Connections on the temporal layer are updated with a fixed one-step time delay, based on activity synchronisation between *BMU(t–1)* and *BMU(t)*. **Right**: Long Term Potentiation (LTP) and Long-Term Depression (LTD) of Hebbian connections between consecutively activated nodes in the learning phase. (from Ferro et al., 2010)

In its typical configuration (Kohonen, 2001), a *SOM* is a grid of parallel processing nodes fully connected to an *input layer* where incoming stimuli are encoded. Input connections are modelled as weighted communication channels with no time delay, defining a spatial layer (SL) of connectivity. In the present work we make use of Topological Temporal Hebbian Self-Organizing Maps (*T2HSOMs*) (Koutnik, 2007; Ferro et al., 2010, 2011), an extension of traditional SOMs augmented with re-entrant Hebbian connections defined over a temporal layer (TL), encoding probabilistic expectations of time series. Each map node is linked to all other nodes through a delayed connection that provides, at time *t*, the activity of all nodes at time *t-1* (Fig. 2, left).

Nodes exhibit a short-term dynamic, based on equation (1) below, and a long-term dynamic, based on adaptive learning (Ferro et al., 2010). Upon presentation of an input stimulus at time *t*, all map nodes are activated synchronously at different levels $h_i(t)$, but only the most highly activated one, called the Best Matching Unit (or *BMU(t)*), is selected. Node that the activation equation (1) of node n_i at time *t* is the sum of two functions. The first function, $h_{S,i}(t)$, measures how similar the input vector weights of node n_i are to the current input, and the second one, $h_{T,i}(t)$, how predictable the current input is on the basis of past input. Parameters α and β (in 1) determine the relative contribution of the two functions to the overall activation score; high values of α make the map sensitive to the specific content of the current input stimulus, while high values of β make the map sensitive to its timing.

(1) $$BMU(t) = n_i = \arg\max_{i=1,\dots N} \{ h_i(t) \}$$

$$\text{where } h_i(t) = \alpha \cdot h_{S,i}(t) + \beta \cdot h_{T,i}(t)$$

In the learning phase, at each time *t*, *BMU(t)* adjusts its connection weights on both layers (SL and TL) and propagates adjustment to neighbouring nodes. On SL, adjustment makes connection weights closer to values in the input vector. On TL, adjustment of Hebbian connections i) potentiates the strength of association from *BMU(t-1)* to *BMU(t)* (and its neighbouring nodes), and ii) depresses the strength of association from all other nodes to *BMU(t)* (and its neighbouring nodes) (Fig. 2, right). This amounts to logically enforcing the entailment *BMU(t)* → *BMU(t-1)*, thereby inducing the emergence of dedicated patterns of activation over nodes that are reminiscent of word graphs (Fig. 1, left).

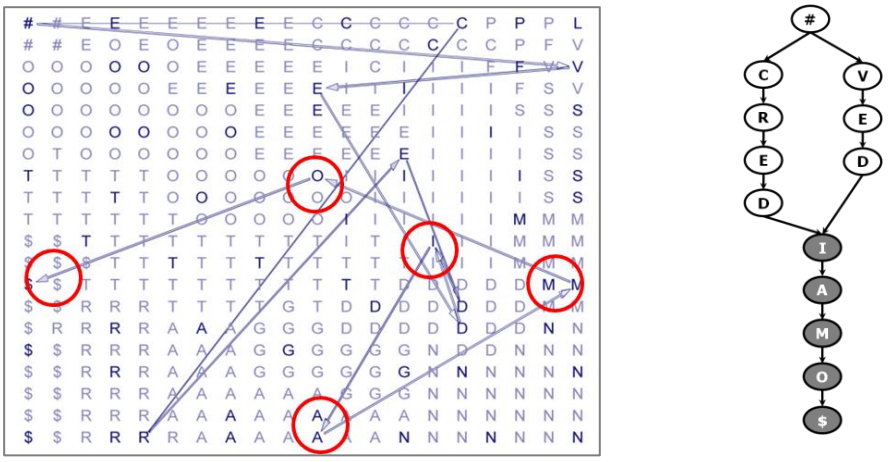


Figure 3: BMU activation chains for *vediamo-crediamo* on a 20×20 map (left) and their word-graph representation (right). Shared processing nodes are circled on the map and shaded in grey in the word graph.

When a string of letters is presented to the map one character at a time, a temporal chain of *BMU*s is activated. Figure 3 illustrates two such temporal chains, triggered by the

Italian verb forms *crediamo* and *vediamo* ('we believe' and 'we see') presented to a 20×20 nodes map trained on 30 verb paradigms, sampled from the Italian Treebank corpus (Montemagni et al., 2003) by decreasing values of cumulative paradigm frequency. In the figure, each node is labelled with the letter the node is most sensitive to. Pointed arrows represent temporal connections linking two consecutively activated nodes, thus depicting the temporal sequence of node activation, starting from the beginning-of-word symbol '#' (anchored in the top left corner of the map) and ending to '$'. Activation chains allow us to inspect the memory patterns that a map develops through training.

Although temporal learning is based on first-order re-entrant Hebbian connections only (i.e. connections emanating from the immediately preceding *BMU*), nodes can propagate information of their immediate left-context over longer activation patterns, thereby simulating orders of memory longer than 1. In Figure 3, both letters *D* in *VEDIAMO* and *CREDIAMO* are preceded by *E*. Nonetheless they recruit two topologically close but distinct nodes on the map which thus "store" information that the two *E*s were in turn preceded by a different symbol (second-order memory). A Hebbian map can thus enforce longer orders of memory through a profligate use of dedicated nodes, trading space for time. The trade-off is based on learning, and depends on available memory resources and distribution of training data. Upon hitting the ensuing *I* in *VEDIAMO* and *CREDIAMO*, the map in Figure 3 recruits the same *BMU*, showing that the map cannot retain higher-order memory events (at least for this specific sequence). The distance on the map between two *BMU*s that respond to identical symbols in different input contexts thus reflects the extent to which the map perceives them as similar. By the same token, the topological distance between chains of activated *BMU*s responding to similar input strings tells us how well the map is aligning the two strings. This is a general problem for morphology induction, arising whenever known symbol patterns are presented in novel arrangements, as when speakers are able to spot the Arabic verb root shared by *kataba* ("he wrote") and *yaktubu* ("he writes"), or the German verb root common to *machen* ("make") and *gemacht* ("made" past participle).

3.1. Inductive Bias and Input recoding

It is useful at this stage to focus on the inductive bias of Hebbian SOMs under specific parameter configurations and training conditions. Figure 4 (right) shows the topological configuration of a map trained on a uniformly distributed data set of 64 binary strings.

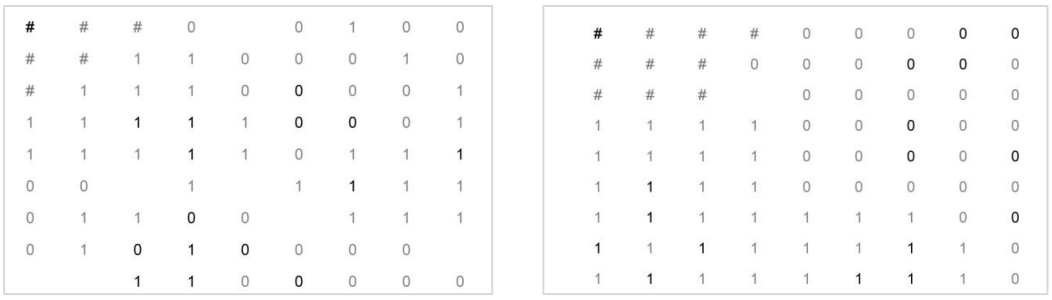


Figure 4: Topologies of a *T*-map (left) and a *ST*-map (right) trained on binary strings. Only the top left corner of both maps is shown.

The map (henceforth referred to as a Spatio-temporal map or *ST*-map for short) presents a comparatively high value for α, causing equation (1) to be more sensitive to symbol coding than to symbol timing. Conversely, the leftt-hand map of Figure 4 (called Temporal map or *T*-map for short), is more sensitive to weights on the temporal connection layer due to lower α. The resulting topologies of the two maps are different. Symbol coding defines the most external level of clustering for the *ST*-map, with all nodes

fired by a specific bit being clustered in the same connected area. In fact, due to the uniform distribution the training data, '0' and '1' take the top right corner and bottom left corner of the map respectively, parting the map's topological space into two halves. Within each half, several nodes are recruited for different instances of the same bit, as a function of their position in the training sequences. The *T*-map reverses the clustering hierarchy, with timing defining the most external level of nesting (Fig. 4, left). Within each such external cluster, nodes are specialised for being sensitive to different, similarly-distributed bits. Hence, '0' and '1' appear to be scattered around the map's topological space, depending on their time-bound distribution in the training corpus.

Differences in topological organisation define the way the two map types categorise input symbols (and eventually input sequences). A *T*-map recodes symbols positionally, by recruiting nodes that are sensitive to – say – a *1* in first position (1_1), a *1* in second position (1_2), a *0* in first position (0_1) and so on and so forth. An *ST*-map tends to categorise instances of the same symbol across different positions in the input. Nonetheless, since *ST*-maps are also sensitive to timing, they are able to distinguish instances of the same symbol on the basis of its left context. For example, a *1* preceded by a *0* will activate a dedicated 0_*1* node, whereas a *1* preceded by another *1* will activate a different 1_*1* node in the same cluster. The difference is shown in Figure 5, plotting the topological dispersion of map nodes by position of input bits for two maps.

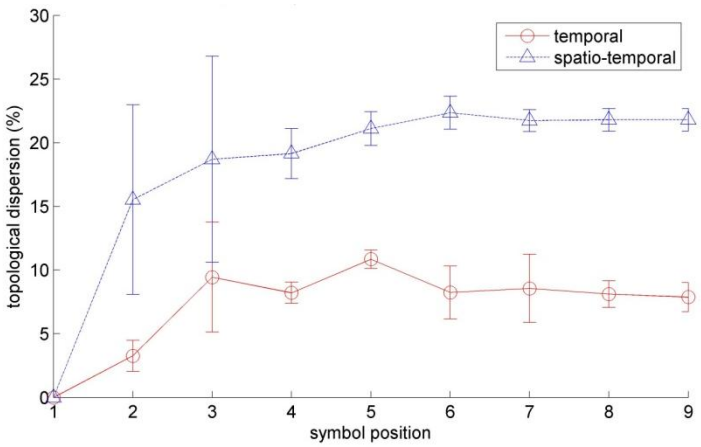


Figure 5: Topological dispersion of symbols on *T*- and *ST*-maps, plotted by their position in input words.

The characteristically distinct ways the two maps recode symbols are reminiscent of the encoding schemes known in the literature as "positional coding" and "Wickelcoding" (Sibley et al. 2008, Davis 2010). It is important to bear in mind, however, that in Hebbian maps symbol encoding is not wired-in in the network's input layer, as customary with connectionist architectures. Rather it is the result of a recoding process based on training data. The number of positional slots, or the length of the left-context affecting symbol recoding is adjusted dynamically on i) the map's memory resources, and ii) the combinatorial complexity of the training input and its frequency distribution.

Differences in the map's inductive bias and recoding scheme have interesting effects on the way input forms are organised and processed through memory structures. *T*-maps are more sensitive to time and can build up stronger expectations over an upcoming symbol in activation. Therefore, they are slightly less accurate than *ST*-maps in perceiving known words (as they trust more their own expectations than actual input stimuli) and considerably less accurate than *ST*-maps in perceiving novel words, for which they built no expectations in the learning phase. When it comes to recalling stored words, however, *T*-maps are more accurate, as they can rely on more accurate positional coding of symbols. On the other hand, *ST*-maps are weaker in capitalising on past events

and thus more tolerant towards unexpected symbols. The implication of this is that they recall novel input sequences more accurately. More importantly for our present concerns, *ST*-maps, unlike *T*-maps, can develop pools of nodes that are specifically sensitive to position-independent *n*-grams. As we shall see in what follows, the notion of position-independent *n*-gram is the closest approximation to the notion of morpheme a Hebbian map can possibly get and has a bearing on the map's ability to recognise morphologically complex novel word forms.

4. Experimental design and materials

To investigate the interplay between prediction and generalisation in the morphological lexicon, we trained instances of a Temporal (*T*) and Spatio-temporal (*ST)* 40×40 T2HSOM on Italian and German text excerpts of about 3,000 word tokens, sampled from two books of child literature: *Pinocchio's Adventures* and the brothers Grimm's *Fairy Tales*. To simulate low-level memory processes for serial order and their impact on morphological organisation, only information about raw forms was provided in training. Such a preliminary step in the process of morphology acquisition is intended to investigate the important but often neglected connection between input word recoding and perception of morphological structure.

Word forms are encoded as strings of capitalised letters preceded by '#' and ended by '$': e.g. '#IST$' for *ist*. Word forms are input to a T2HSOM one letter at a time, with memory of past letters being recoded through re-entrant Hebbian connections that are reset upon presentation of '#'. All letters common to the German and Italian alphabets are written in upper-case. Umlauted characters are written as lower-case digraphs (e.g. '#BRueCKE$' for *Brücke*) and the sharp s '*ß*' as 'ss' (e.g. '#BEIssEN$' for *beißen*). In both cases, pairs of lower-case letters are processed as one symbol. Letters are encoded on the input layer as mutually orthogonal, binary vector codes. Identical letter codes were used for upper-case letters in both German and Italian. 'Five T-maps and five ST-maps were trained on each language for 100 epochs.'. In the five *T*-maps, α= 0.5 and β= 1.0. In the five *ST*-maps, α= 0.087 and β= 1.0. After training, we probed the memory content of the maps on two basic tasks, using both known word forms (i.e. words belonging to the map's training set) and unknown word forms (forming the test set). Both Italian and German test sets contain unseen word forms belonging to word families partially attested in the training set.

4.1. Recoding

The task consists in testing the accuracy of the activation function (1) on both known and unknown word forms. For each symbol *s* shown to the map at time *t*, we test if the map recodes the symbol correctly by activating an appropriate *BMU*(*t*) labelled with *s*. An input word is taken to be recoded accurately if all its letters are recoded accurately. Activation requires faithful memory traces of the currently input symbol, but is also a function of how well the current input symbol is predicted on the basis of past symbols.

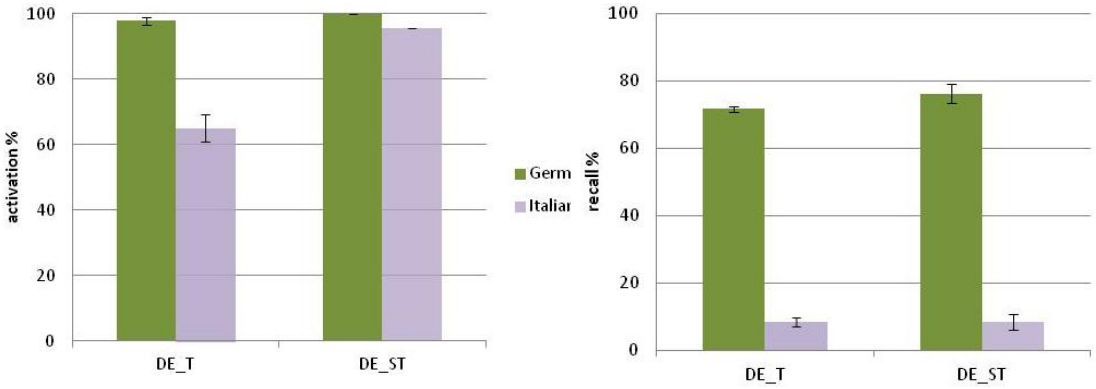


Figure 6. **Left**: Activation scores of German Temporal (DE_T) and Spatio-Temporal (DE_ST) maps, averaged across multiple instances, tested on unknown German words and Italian word forms. **Right**: Recall scores for German *T*-maps and *ST*-maps on unknown German words and Italian word forms.

That more than just storage is involved in recoding is shown by the diagram to the right of Figure 6, providing accuracy scores for both temporal and spatio-temporal German maps, tested on unknown German words and unknown (and unfamiliar) Italian words. Although all Italian letters are present in the German alphabet, 35% of Italian words are wrongly recoded by the German *T*-map (DE_T). This is in striking contrast with the 96% accuracy of German *ST*-maps (DE_ST) on the same task and witnesses the higher sensitivity of *T*-maps to built-in expectations over letter *n*-grams.

4.2. Recall

After Baddeley (1986), we model lexical recall as the task of reinstating a word form soon after a map is exposed to it. The experimental protocol is intended to highlight the dynamic interaction between short-term integration/sustainment of memory traces and long-term storage of lexical information. A *N*-nodes map is first exposed to an input word *w* of length n_w. Its resulting integrated activation pattern $\hat{Y}=\{\hat{y}_1,\ldots,\hat{y}_N\}$, with

$$(2) \quad \hat{y}_i = \max_{t=2,\ldots,n_w} \{y_i(t)\}, \; i = 1, \ldots, N$$

is input to the same map (n_w*-1)* times. At each time step *t*, the map's *BMU*(*t*) is calculated according to the activation function (1). A word *w* is taken to be recalled accurately if for each *t* ranging from 2 to n_w, the label of *BMU*(*t*) matches the *t*-th letter in *w*. The protocol is thus intended to assess how well the map can output the appropriate sequence of symbols in *w* upon presentation of the whole activation pattern triggered by *w*. Results of recalling words from the training set are shown in Figure 7, grouped by language and map type, and averaged across instances of map type.

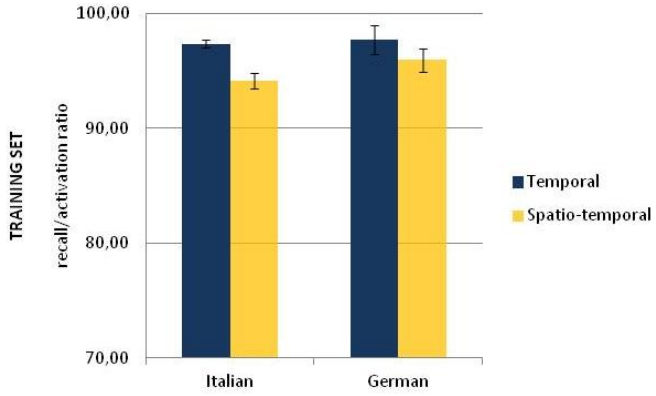


Figure 7: Recall/activation scores on the German and Italian training sets averaged across 5 instances of Temporal and Spatio-Temporal maps.

Note that both temporal and spatio-temporal maps are fairly good at recalling familiar words (training set), with a marginally significant but consistent advantage for temporal maps. This pattern of results is distinctly reversed in Figure 8, plotting the recall/activation ratio on test words, with temporal maps performing consistently worse in both languages. Incidentally, both map types perform considerably worse when tested on recalling unfamiliar unknown words, as is the case of German maps recalling Italian words (see Fig. 6, right).

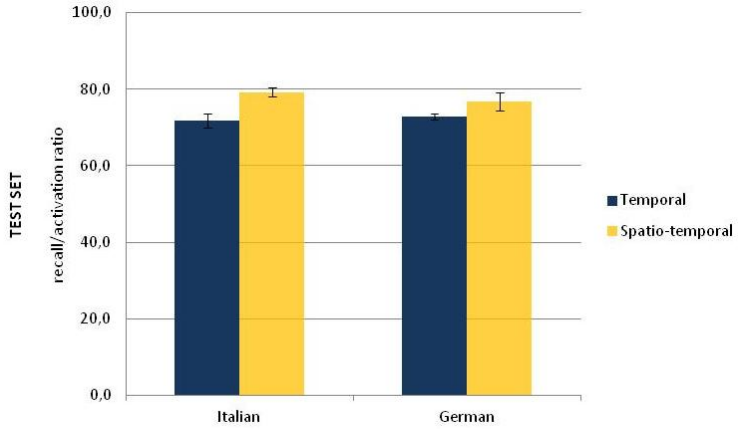


Figure 8: Recall/activation scores on the German and Italian test sets. Scores are averaged across 5 instances of Temporal and Spatio-Temporal maps.

5. Data Analysis

To better understand how generalisation works in Hebbian maps, it is useful to look at Figure 9, where we assume that a (*ST)* map trained on three Italian verb forms (*VEDIAMO* 'we see', *VEDETE* 'you see', and *CREDIAMO* 'we believe') is prompted to recall *CREDETE* 'you believe' afresh. The connection patterns highlighted by grey and red arrows on the trained *ST*-map to the left, are unfolded and vertically arranged in the word graph to the right, to emphasise what is shared and what is not shared by activation patterns.

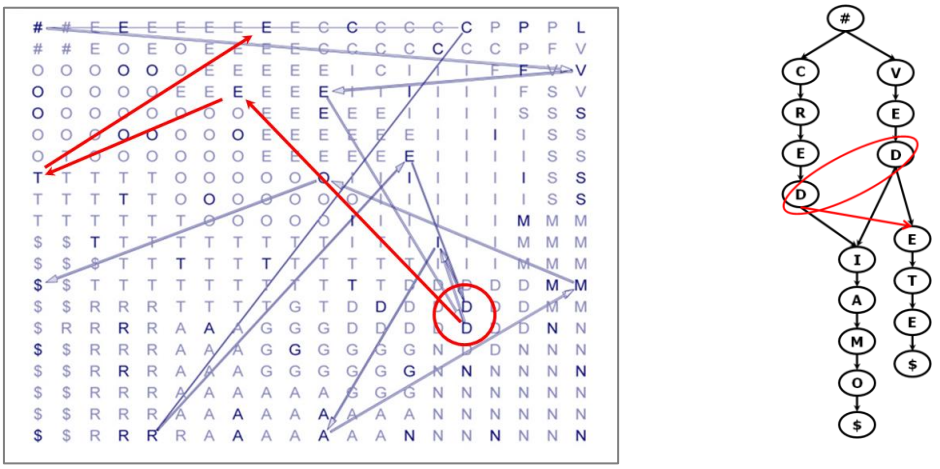


Figure 9: BMU activation chains for *VEDIAMO-VEDETE-CREDIAMO* on a 20×20 map (left) and their word-graph representation (right).

The crucial generalisation step here is represented by a red directed arc in the word graph, and involves the unattested connection between the root *CRED* and the ending –*ETE*. For the map to entertain this connection, it has to be able to i) generalise over the two instances of *D* in *CRED* and *VED*, and ii) align the ensuing ending –*IAMO* in *VEDIAMO* and *CREDIAMO*. The difference in generalisation potential between *T*-maps and *ST*-maps demonstrate that both steps are more likely if symbols are recoded in a context-sensitive but position-independent way, in keeping with the minimal generalisation requirement that rules mapping fully inflected forms are based on the immediate structural context of the change (Albright and Hayes 2003).

Positional encoding appears to be a more effective strategy in lexical recall, suggesting that generalisation and prediction are indeed complementary processing functions, serving different purposes. This is quantitatively summarised in Figure 10 and Figure 11, where we relate the difference in recall accuracy between the two map types to perception of morphological structure, and measures of topological organisation, such as length of *receptive fields*, average per node number of input words and relative number of used-up nodes (*BMU*s).

Following Voegtlin (2002), the receptive field of a map node *n* is defined as the common end of all input strings triggering *n*. For example, if a single node is triggered by '*O*' in the forms *VEDIAMO* and *CREDIAMO* only, its receptive field will be –*EDIAMO*. Accordingly, evidence that *ST*-maps have i) significantly longer receptive fields than *T*-maps (Fig.10, top), ii) more words triggering a single node on average (Fig. 10, centre), and iii) fewer *BMU*s (Fig. 10, bottom), confirms that they are better at finding recurrent substrings in input words. Figure 11 shows how this evidence relates to morphological structure and map's performance. Misalignment (Figure 11, top) tells us how badly activation chains of morphologically-related forms are aligned on shared morphemes (Marzi et al. 2012). High values here indicate that – say – *VEDERE* ('to see') and *CREDERE* ('to believe') activate distinct nodes on their common endings. Lower values indicate a better correlation between activation patterns and shared morphological structure. In turn, this is shown to correlate negatively with accuracy in recalling novel words (Figure 11, bottom). Scores are given for a few verb inflections only.

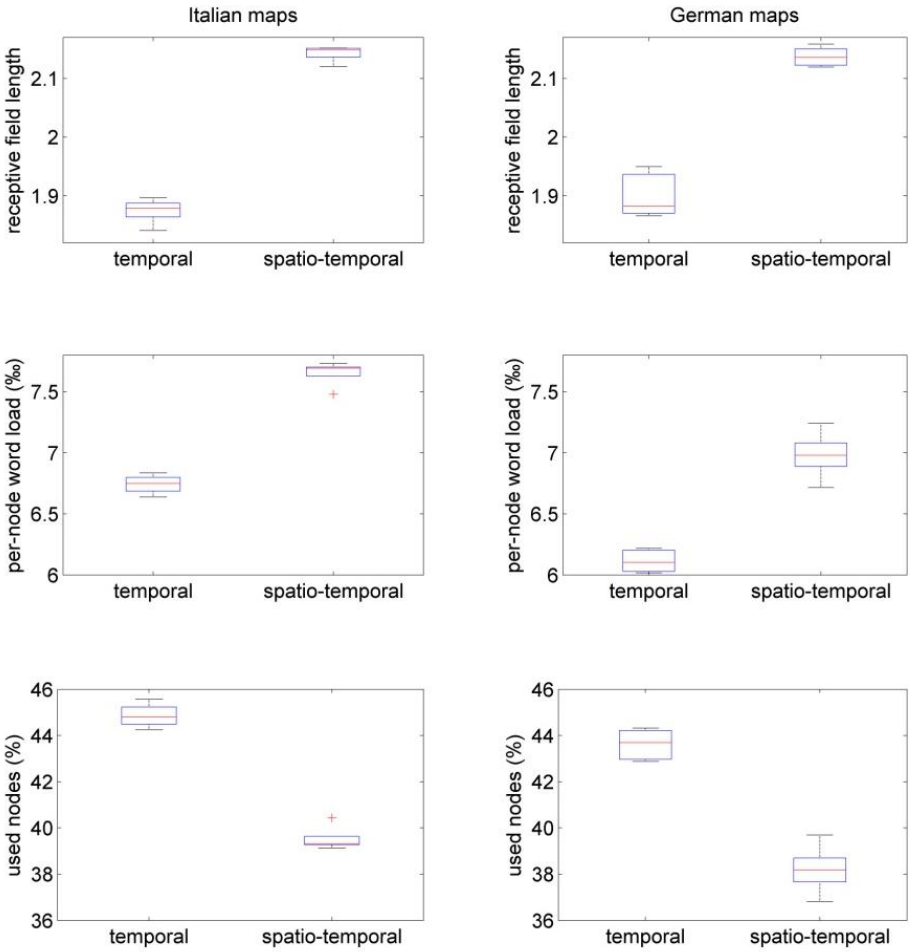


Figure 10: Measures of topological organisation of temporal and spatio-temporal maps on the Italian (left) and German (right) training sets. Scores are averaged across 5 instances of each map type.

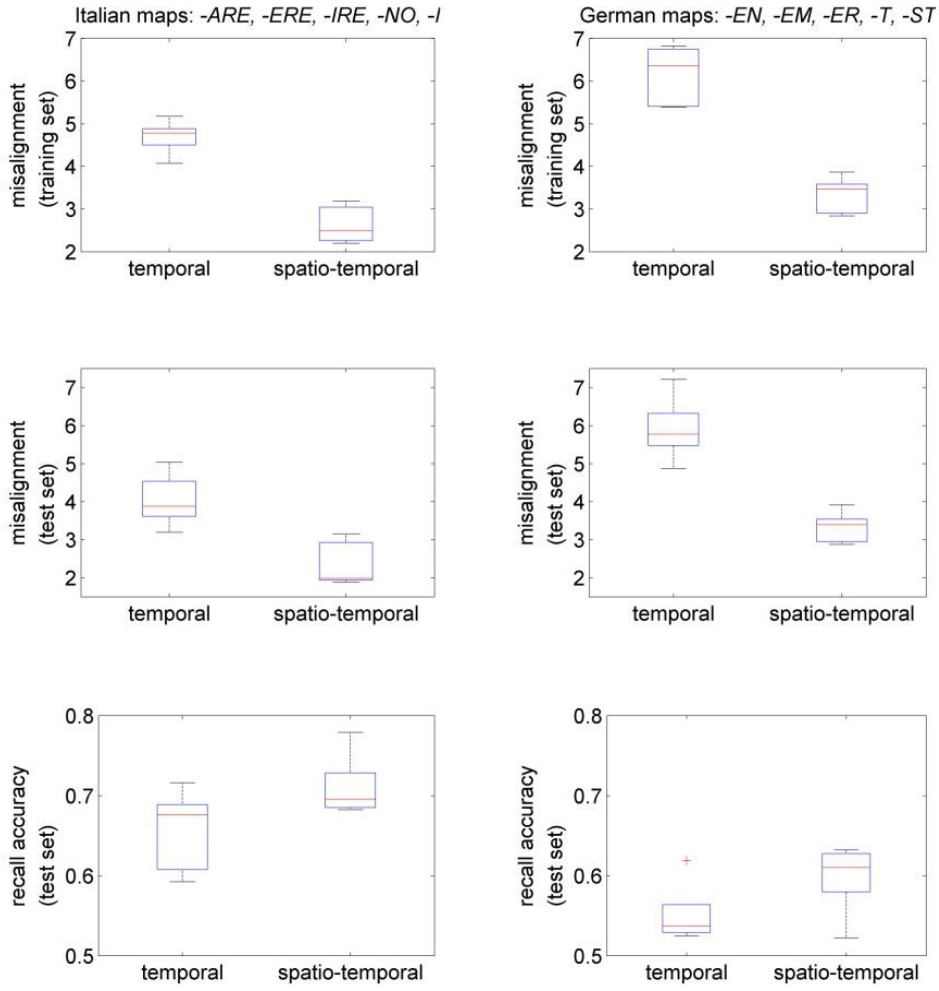


Figure 11: **Top**: misalignment scores across activation patterns triggered by selected inflectional endings on temporal and spatio-temporal maps for both Italian (left) and German (right). **Centre**: misalignment scores for the same set of inflections calculated on the test set. **Bottom**: recall scores of word forms in the test set inflected for the selected endings.

An analysis of the errors made by the two map types in recalling known words indicates that prediction-based errors are more local, involving letter substitution in specific time slots, with no mistakes for letters coming after that slot. Length preservation in the face of local errors is what we would expect for words stored and recalled positionally. This is confirmed by the average per-word percentage of misrecalled symbols in the two maps: about 25% for *T*-maps and more than 37% for *ST*-maps on known Italian words; about 21% for *T*-maps and 34% for *ST*-maps on known German words.

6. Concluding remarks

Prediction affects the way we perceive things and events, through anticipation of upcoming stimuli and integration of missing or noisy information in the input. In Lexical Hebbian maps, prediction is implemented as a process of first-order anticipatory activation of to-be-selected *BMU*s, which presupposes context-driven sensitivity of map nodes to time-bound letters/segments. Thanks to such built-in prediction drive and accurate recoding of time-bound stimuli through training, Hebbian maps show a

remarkable capacity to use past information to process and store the incoming input, offering an interesting model of memory-based word processing and recoding.

In the computational framework offered here, prediction presupposes a bias for past events, under a closed world assumption that what is not currently known to be attested is fairly unlikely, if not impossible. From this perspective, lexical items exhibit minimally redundant patterns which are based on a strictly positional coding of constituent symbols and strong serial connections between them. Morphological generalisation, conversely, seems to require the ability to understand unseen forms based on the discovery of recurrent sub-lexical constituents (morphemes), whose proper coding is context-sensitive but independent of specific positional slots. It is a remarkable aspect of the experimental framework reported here that the two strategies of prediction and generalisation are in fact the outcome of different parameter configurations of a unitary memory model. This has, in our view, a few interesting theoretical implications.

The proposed memory framework radically departs from derivational approaches to morphological competence, by suggesting that principles of lexical organisation may rest on memory self-organisation and recoding, and that rule-like morphological generalisations are the outcome of cautious extension of attested inflections to different word families than those originally attesting them. This move, in our view, blurs not only the traditional linguistic dichotomy between lexicon and rules, but also the related but somewhat more general divide between input/output representations (or knowledge of 'what') and processing principles (or knowledge of 'how'). According to the perspective endorsed in this paper, ways of processing and structural properties of input/output representations are in fact mutually implied, as representations are not given, pre-existing abstract representations but the outcome of an active process of recoding. In turn, processing is memory-driven, with memories of past evidence and already structured information being brought to bear on attentional and combinatorial strategies.

It could be suggested, in line with interactive-activation accounts of word processing (Diependaele et al. 2009), that both strategies for memory organisation (temporal and spatio-temporal) may simultaneously compete in word processing and interact through feedback connections. Temporal maps are better interfaced with the level of lexico-semantic representations, while spatio-temporal maps are more conducive to structured morpho-orthographic representations. Our computational framework allows us to spell out principles of this dynamic interaction in some detail, by putting to extensive empirical test detailed alternative hypotheses. For example, we could test the view that the relative balance between prediction and generalisation is in fact decided dynamically as a function of the stage of acquisition. Item-based learning (Tomasello 2003) may provide an early advantage to children acquiring the lexicon of their own language, as they may find it easier to retrieve and produce a word on the basis of a stronger prediction drive. This can be shown in Figure 12, where we compare the percentage of correctly recalled words by a *T*-map and a *ST*-map through early learning epochs, together with their average length (expressed as a percentage over the average length of all training words). A *T*-map recalls more and longer words at early stages, as item-based storage is relatively local and instantaneous. Finding morphological structure in memorised words, on the other hand, appears to require more time and more evidence for sublexical memory structures to be used in word processing, recoding and retrieval.

Finally, it may well be possible that different languages and morphology types favour either strategy. For example, templatic morphologies and Semitic morphologies in particular may prove to be more conducive to serial, position-based coding of sub-lexical material than more concatenative morphologies are, suggesting that a bias for either strategy could develop as the result of learning. Algorithms for morphology acquisition should be more valued for their general capacity to adapt themselves to the

morphological structure of a target language, rather than for the strength of their inductive morphological bias.

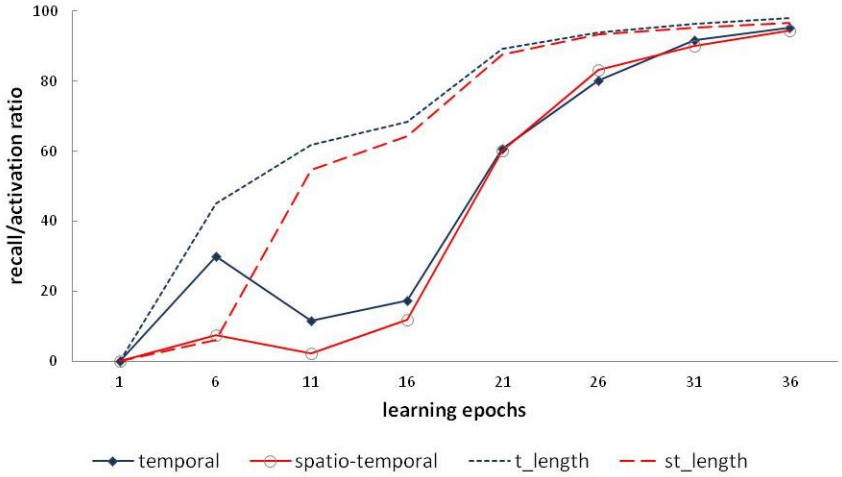


Figure 12: Recall/activation scores and average length of recalled words by early learning epochs for *T*-maps and *ST*-maps.

References

Albright, A., Hayes, B. (2003), Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90, 119–161.

Altmann, G.T.M., and Kamide, Y. (1999), Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73, 247-264

Baayen, H., Dijkstra, T., and Schreuder, R. (1997), Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, 37, 94-117.

Baddeley, A.D., and Gathercole, S.E. (1992), Learning to read: The role of the phonological loop. In J. Alegria, D. Holender, J.J. de Morais and M. Radeau (eds.), *Analytic approaches to human cognition*, Amsterdam, Elsevier, 153-168.

Berko, J. (1958), The child's learning of English morphology. *Word* 14, 150-77.

Blevins, J. P. (2006), Word-based morphology. *Journal of Linguistics*, 42, 531-573.

Booij, Geert (2010), Construction Morphology. Oxford, Oxford University Press.

Burzio, L. (2004), Paradigmatic and syntagmatic relations in italian verbal inflection. In J. Auger, J.C. Clements and B. Vance (eds.), *Contemporary Approaches to Romance Linguistics*, Amsterdam, John Benjamins.

Bybee, J.L., Pardo E. (1981), Morphological and lexical conditioning of rules: Experimental evidence from Spanish. *Linguistics* 19, 937-68.

Bybee, J.L., Slobin D.I. (1982), Rules and Schemas in the development and use of the English Past Tense. *Language,* 58, 265-289.

Bybee, J.L., C. Moder 1983 "Morphological Classes as Natural Categories" *Language* 9, 251-270.

DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8, 1117-1121.

Clahsen, H. (1999), Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22, 991-1060.

Corbett, G., and Fraser, N. (1993), Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics*, 29, 113-142.

Cowan, N. (2001), The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87-185.

Davis, C. J. (2010), The Spatial Coding Model of Visual Word Identification. *Psychological Review*, 117, 3, 713-758.

Di Sciullo, A. M., Williams, E. (1987), *On the Definition of Word*. Cambridge, MA, MIT Press.

Diependaele, K., Sandra, D., & Grainger, J. (2009). Semantic transparency and masked morphological priming: The case of prefixed words. *Memory & Cognition*, 37(6), 895-908.

Federmeier, K.D. (2007), Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology*, 44, 491-505.

Ferro M., Pezzulo G., Pirrelli V. (2010), Morphology, Memory and the Mental Lexicon. In *Lingue e Linguaggio*, V. Pirrelli (ed.) *Interdisciplinary aspects to understanding word processing and storage*, Bologna, Il Mulino, 199-238.

Ferro, M., Marzi, C., Pirrelli, V. (2011), A self-organizing model of word storage and processing: implications for morphology learning, in *Lingue e Linguaggio*, X, 2, Bologna, Il Mulino, 209-226.

Ford, M., Marslen-Wilson, W., Davis, M. (2003), Morphology and frequency: contrasting methodologies. In H. Baayen and R. Schreuder (eds.), *Morphological Structure in Language Processing*, Berlin-New York, Mouton de Gruyter.

Gathercole, S.E., Pickering, S.J. (2001), Working memory deficits in children with special educational needs. *British Journal of Special Education*, 28, 89-97.

Hay, J. (2001), Lexical frequency in morphology: is everything relative? *Linguistics*, 39, 1041-1111.

Maratsos, M. (2000), More overregularizations after all. *Journal of Child Language*, 28, 32-54.

Marzi, C., Ferro, M., Caudai, C., Pirrelli, V. (2012), Evaluating Hebbian Self-Organizing Memories for Lexical Representation and Access. In *Proceedings* of LREC 2012, Istanbul, Turkey.

Matthews, P.H. (1991), Morphology. Cambridge, Cambridge University Press.

McClelland, J., Patterson, K. (2002), Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Science*, 6, 465-472.

Pickering, M.J., Garrod, S. (2007), Do people use language production to make predictions during

comprehension? *Trends in Cognitive Sciences*, 11, 105-110.

Pinker, S., Prince, A. (1988), *On language and connectionism: Analysis of a parallel distributed processing model of language acquisition*, *Cognition*, 29, 195-247.

Pinker, S., Ullman, M.T. (2002), The past and future of the past tense. *Trends in Cognitive Science*, 6, 456-463.

Pirrelli, V. (2000), Paradigmi in Morfologia. Un approccio interdisciplinare alla flessione verbale dell'italiano, Pisa, Istituti Editoriali e Poligrafici Internazionali.

Pirrelli, V., Ferro, M., Calderone, B., (2011), Learning paradigms in time and space. Computational evidence from Romance languages. In M. Goldbach, M.O. Hinzelin, M. Maiden and J.C. Smith (eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*, Oxford, Oxford University Press, 135-157.

Post, B., Marslen-Wilson, W., Randall, B., Tyler, L.K. (2008), The processing of English regular inflections: Phonological cues to morphological structure, *Cognition*, 109, 1-17.

Prasada, S., Pinker, S. (1993), Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes,* 8, 1-56.

Rumelhart, D. E., McClelland, J. L. (1986), On learning of past tenses of English verbs. In J.L. McClelland and D.E. Rumelhart (eds.), *Parallel distributed processing*, Cambridge, MA, MIT Press, 2, 216-271.

Seidenberg, M.S., McClelland, J.L. (1989), A distributed, developmental model of word recognition and naming. In A. Galaburda (ed.), *From neurons to reading*, MIT Press.

Sibley, D.E., Kello, C.T., Plaut, D., Elman, J.L. (2008), Large-scale modeling of wordform learning and representation, *Cognitive Science*, 32, 741-754.

Smolensky, P. (1988), On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11, 1-74.

Staub, A., Clifton, C., Jr (2006), Syntactic prediction in language comprehension: evidence from either. . .or. *Journal of Experimental Psychology, Learning, Memory and Cognition*, 32, 425–436.

Stemberger, J.P., Middleton C.S. (2003), Vowel dominance and morphological processing, *Language and Cognitive Processes*, 18(4), 369-404.

Tabak, W., Schreuder, R., Baayen, R.H. (2005), Lexical statistics and lexical processing: semantic density, information complexity, sex and irregularity in Dutch. In M. Reis and S. Kepser (eds.), *Linguistic Evidence*, Berlin, Mouton de Gruyter, 529-555.

Tomasello, M. (2003), Constructing a Language: A Usage-based Theory of Language Acquisition. Cambridge, MA, Harvard University Press.

Wilson, M., Knoblich, G. (2005), The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131, 460-473.

Meaning resides in fully inflected forms: The Georgian "unwillingness" construction[1]

Alice C. Harris
University of Massachusetts Amherst
acharris@linguist.umass.edu

1. Introduction

Subject and object agreement on the verb are generally highly regular in Georgian, if a bit complex. But in an obscure corner of the grammar, there are forms traditionally referred to as *uneblieroba* 'unwillingness'. These can be used only with certain verbs of dark meaning, including 'kill', 'beat', 'curse', or 'insult', and the forms include the meaning that the agent acted 'unwillingly'. The construction is illustrated by the sentence in (1).

(1) še-mo-g-e-lanʒγ-eb-i[2]
PV-PV-2O-E-insult-SM-SCM
'I will unwillingly make you insult me.'

(The form in (1) is a complete, grammatical sentence.) As indicated by the translation line, the meaning involves three arguments, here 'I', 'you', and 'me'. As shown here, the prefix *g-* in (1) indicates agreement with the "causee", the argument that carries out the forced action, here 'you'. As an indirect object, the causee conditions object agreement, *g-*. Because of slot competition, the matrix agent, here 'I', cannot be marked overtly; nevertheless, it can be unambiguously discerned in this form, just as in other regular forms in Georgian. But what of the direct object 'me'? It cannot be represented as an independent pronoun as most arguments can, and it does not condition agreement in the verb form as other arguments in Georgian do. We may assume that the first person direct object cannot be marked on the verb because of slot competition. But the person of this object is not discernable in the sense that the first person subject is in this same form. Georgian has a special way of marking first person direct object when the indirect object is second person, Object Camouflage, a manifestation of the person case constraint. Object Camouflage cannot be used in (1). This is in stark contrast to ordinary simple sentences and to ordinary causatives formed with these and other verb roots and lacking the meaning 'unwillingly'.

The puzzle here is that nothing obviously represents

- 'me'
- 'unwillingly'
- causation, or even transitivity.

The ordinary way to express 'unwillingly' in Georgian involves one of the independent adverbs *uneburad* or *unebliet* 'unwillingly'. In §2 of this paper, I discuss the meaning of each morpheme of (1) in other contexts and show that the notions bulleted above are not overtly expressed; rather, these meanings emerge only from the

[1] I am grateful to Shukia Apridonidze for calling this construction to my attention many years ago. I would like also to thank my main informants for this material, Marina Zambakhidze and Akaki Kenchoshvili. The research reported here was supported by the National Science Foundation under grant BCS- 0091691.

[2] The following abbreviations are used in this paper: CAUS causative, CISL cislocative, DAT dative, ERG ergative, NOM nominative, O object marker, PV preverb, S subject, SCM screeve (tense-aspect-mood category) marker, SM series marker. All arguments in these examples are singular, and number is not glossed.

construction as a whole. These three meanings are entirely non-compositional. It is not even possible to say here that the meanings are borne by zero morphemes, in the usual sense of that notion, since there is no minimal contrast with another form and no appropriate place for such a morpheme to occur.

In §3 I argue, instead, for a schema which includes certain morphology and occurs only with certain verb roots. It contains the meanings of (i) the presence a direct object coreferential with the subject, (ii) 'unwillingly', (iii) transitivity and causation even though these meanings are not expressed by any specific morpheme. I argue that forms of this kind constitute another argument for Construction Morphology or some other approach in which the structure of the whole word is taken into account in determining meaning.

2. The core meanings of morphemes

2.1. Person agreement

In Georgian, subjects, direct objects, and indirect objects condition verb agreement, as illustrated (for subjects and direct objects) in (2).

(2) (a) m-xat'av-s
1O-paint-3S
'S/he paints me.'
(b) g-xat'av-s
2O-paint-3S
'S/he paints you.'
(c) xat'av-s
paint-3S
'S/he paints him/her/it/them.'

(In order to focus the discussion on agreement more clearly, I have not separated and glossed some morphemes here and below.) (2c) illustrates the fact that third person direct objects do not ordinarily condition agreement, although first and second person objects always condition agreement. All other arguments in (2) are overtly marked for person (with no gender distinction). Some affixes are used in both singular and plural, while others are used only in the singular; this is not indicated in the glosses, in order to keep the material simpler.

The prefix *g-* in (1) indicates agreement with the "causee", the argument that carries out the forced action, here 'you'. Thus, that portion of the meaning is accounted for straightforwardly.

So-called slot competition prevents more than one agreement prefix from appearing, as shown in (3), where both the first person subject marker, *v-*, and the second person object marker, *g-*, would otherwise be expected.[3]

(3)	g-xat'av	*v-g-xat'av	*g-v-xat'av
	2O-paint	1S-2O-paint	2O-1S-paint
	'I paint you.'	'I paint you.'	'I paint you.'

In spite of slot competition, it is generally possible to discern the person of each argument unambiguously, though number is less unambiguously marked. This difference in the person of each form is illustrated in (4).

[3]The slot-competition approach does not account easily for the occurrence of the sequence *v-h-/v-s-*, which occurred in Old Georgian and remains in literary Modern Georgian (see Harris 2006).

(4) (a) g-xat'av
2O-paint
'I paint you.'
(b) v-xat'av
1S-paint
'I paint him/her/it.'
(c) m-xat'av
1O-paint
'You paint me.'
(d) xat'av
paint
'You paint him/her/it.'
(e) m-xat'av-s
1O-paint-3S
'S/he paints me.'
(f) g-xat'av-s
2O-paint-3S
'S/he paints you.'
(g) xat'av-s
paint-3S
'S/he paints him/her/it.'

Because third person subjects are not zero-marked in Georgian, the pattern illustrated in (4) makes it possible to discern the person of each argument unambiguously. For example, in (4a), the object 'you' is clearly marked by the prefix *g-*. The subject of this verb form could not be second person, because a reflexive would be expressed differently. Its subject could not be third person, because this would require a suffix, which is not present. Thus, in this example the subject can only be first person.

The meaning of the 'unwillingness' construction involves three arguments, in (1) 'I', 'you', and 'me'. Because of slot competition, the matrix agent, here 'I', cannot be marked overtly, just as in (3) above. Nevertheless, it can be discerned in this form just as unambiguously as in (3) and (4a). That is, if the matrix subject were third person, it would require a suffix, which is not present. If the matrix subject were second person, a reflexive would be used. By process of elimination the subject can only be first person. Thus the first person subject in (1) is not a problem for us.

I now turn to the direct object of (1), 'me'. In general, in case of a first or second person direct object and an indirect object of any person in the same clause, we resort to the construction called Object Camouflage, illustrated in (5b).

(5) (a) *vano m-adareb-s givi-s.
Vano.NOM 1O-compare-3S Givi-DAT
'Vano compares me to Givi.'
(b) vano □em-s tav-s adareb-s givi-s.
Vano.NOM my-DAT PRO-DAT compare-3S Givi-DAT
'Vano compares me to Givi.' (Harris 1981: 49)

By analogy to (4e), we would expect the form in (5a), which, however, is ungrammatical. Instead, in (5) the first person direct object is expressed as a third person object with the proform *tav-*, derived from and homophonous with the noun 'head', as its syntactic head. This form is grammatically a third person direct object and as such conditions no agreement. The semantic person of the direct object is expressed in the person of the possessor of *tav-*, here *čem-* 'my'.

Although Object Camouflage is expected in (1) from the meaning, it is not used. It cannot be included optionally either.

(6) *še-mo-g-e-lanʒγ-eb-i (čem-s) tav-s
PV-PV-2O-E-insult-SM-SCM my-DAT self-DAT
'I will unwillingly make you insult me.'

The object 'me' cannot be represented as *m-*, as it is in (4e). Nor can the meaning 'me' be represented as an independent pronoun, and indeed it cannot be expressed in any way in the verb form. We may assume that the first person direct object cannot be marked on the verb because of slot competition. But the person of this object is not discernable in the sense that the first person subject is in this same form, and it is not expressed in any other way. This is in stark contrast, not only to (5), but also to ordinary causatives formed with these and other verb roots and lacking the meaning 'unwillingly'.

(7) ga-g-a-xat'-v-in-eb čem-s tav-s
PV-2O-A-paint-SM-CAUS-SM my-DAT self-DAT
'I will make you paint me.'

Although the meaning of (7) is similar to that of (1), in (7) the direct object is obligatorily expressed as a camoflaged object, *čems tavs* 'me'. (1) has nothing comparable; the object is not overtly expressed in any way.

An ordinary causative such as (7) can also be constructed with the "dark" predicates that can appear in the "unwillingly" construction. (8) is such an ordinary causative, lacking the meaning 'unwillingly'.

(8) merab-i ga-g-a-lanʒγ-v-in-eb-s čem-s tav-s
Merab-NOM PV-2O-A-insult-SM-CAUS-SM-3S my-DAT PRO-DAT
'Merab will make you insult me.'

In (8), as in most clauses in Georgian, the person of every argument is unambiguously represented. (Number is not always unambiguously represented.) The matrix subject is expressed by the suffix *-s*, the embedded subject (causee) 'you' by the prefix *g-*, and the direct object by the phrase *čems tavs* 'me' (literally 'my head').

The morphology in (8) is very different from that in (1). First, the verb in (1) must have the preverbs *še-mo-*, discussed in more detail below, while the verb in (8) has the preverb *ga-*, which might be different with a different lexeme. Second, the vowel that precedes the verb root in (1) and similar constructions is *e-*, a characteristic of intransitives, while that in (8) is *a-*, which is found with all standard causatives and with some other transitives. The verb in (8), but not that in (1), contains the causative suffix *-in*. And finally, (1), but not (8), contains the suffix *-i*, another characteristic of intransitive verbs.

Because the direct object is coreferential with the subject, we might expect it to be expressed as a reflexive. With an ordinary causative, reflexivization is possible, as in (9).

(9) ekim-ma alap'arak'a vano tavis tav-ze
doctor-ERG talk.CAUS Vano.NOM self's self-on
'The doctor$_i$ got Vano$_j$ to talk about himself$_{i,j}$.' (Harris 1981: 72)

But with the 'unwillingness' construction, this is not possible.

(10) *še-mo-g-e-lanʒγ-eb-i čemi tavi

PV-PV-2O-E-insult-SM-SCM my self
'I will unwillingly make you insult me.'

In Georgian, reflexives are expressed as *tav-* 'self', or as a possessed phrase, such as *čems tavs* 'myself' (see Amiridze 2006), homophonous with the Camouflaged proform in (5b). It is not used in the 'unwillingness' construction.

Georgian also has a morphological marker, *i-*, which is sometimes associated with reflexivity, as when a benefactive indirect object is coreferential with the subject.

(11) gela i-k'er-av-s axal šarval-s (Harris 1981: 95)
Gela.NOM I-sew-SM-3S new trousers-DAT
'Gela sews himself new trousers.'

Clearly this marker is not present in the forms in (1). Thus reflexive morphology and syntax are entirely absent from the form in (1). The object, 'me' in (1), unlike those in the similar causatives (7), (8), and (9) are not marked in any way.

Thus the object 'me' of (1) is not expressed through direct marking, as in (2a), through the inferences connected with slot competition as the subject 'I' is in (4a) and (1), through Object Camouflage as in (5b) or (7), through syntactic or morphological reflexivization as in (9) and (11). How to account for this aspect of (1) remains a problem.

The form in (1) occurs in other person-number combinations.

(12) (a) še-mo-g-e-lanʒγ-eb-i (=(1))
PV-PV-2O-E-insult-SM-SCM
'I will unwillingly make you insult me.'
(b) še-mo-m-e-lanʒγ-eb-i
PV-PV-1O-E-insult-SM-SCM
'You will unwillingly make me insult you.'
(c) vano anzor-s še-mo-e-lanʒγ-eb-a
Vano.NOM Anzor-DAT PV-PV-E-insult-SM-3SG
'Vano$_i$ will unwillingly make Anzor$_j$ insult him$_i$.'

But it is not possible to use this construction to express any of the following: 'I will unwillingly make Vano insult Anzor', 'Vano will unwillingly make you insult me', 'Vano will unwillingly make Anzor insult Merab'. Another such example is given in (13).

(13) *vano anzor-s še-mo-m-e-lanʒγ-eb-a
Vano.NOM Anzor-DAT PV-PV-1O-E-insult-SM-3S
'Vano will unwillingly make Anzor insult me.'

The sentences in (12) are not ambiguous; in particular (12a) cannot mean 'I will unwillingly make you insult someone'; the only possible meaning is the one glossed, 'I will unwillingly make you insult me'. Thus, the interpretation is consistently that the instigator, 'I' in (1), is also the recipient of the insult. While the person and number of the instigator (subject) and the one making the insult (indirect object) may vary to the full extent possible with other verbs (all three persons and both numbers), the subject must be coreferential with the recipient of the insult, the grammatical direct object.

In conclusion, I note that (i) subjects and indirect objects in the 'unwillingness' construction are marked in ways that are normal and compositional for Georgian, and (ii) the direct object is not marked at all.

2.2. Morphology of unaccusatives

The prefix *e-* occurs only with unaccusative verbs.[4]

(14) (a) c'eril-i i-c'ereb-a
letter-NOM I-write-3S
'The letter is written.'
(b) c'eril-i m-e-c'ereb-a
letter-NOM 1O-E-write-3S
'The letter is written to me.'
(c) c'eril-i g-e-c'ereb-a
letter-NOM 2O-E-write-3S
'The letter is written to you.'
(d) c'eril-i e-c'ereb-a
letter-NOM E-write-3S
'The letter is written to her.'

(1) and (12) appear to be transitive, in that they have direct objects, and causative. True causatives, and many other transitives, use *a-* instead in the position occupied by *e-*.

(15) a-c'erinebs
A-write.CAUS.3S
's/he makes him/her write it'

The prefix *e-* generally occurs when an indirect object is associated with an unaccusative verb, and it has that purpose here. *e-* is entirely unexpected with a verb of transitive causative semantics.
Like *e-*, the suffix combination *-eb-i* is characteristic of unaccusatives.

(16) v-dg-eb-i 'I am standing'
v-i-xat'-eb-i 'I am painted, I am being painted'
v-i-lanʒγ-eb-i 'I am being insulted'

Some transitives use *-eb* or *-i*, but never together. Causatives regularly use *-eb* in the present, as in (17).

(17) g-a-xat'-v-in-eb čem-s tav-s
2O-A-paint-SM-CAUS-SM my-DAT PRO-DAT
'I am making you paint me.'

Yet (1) and (12) have *-eb-i*, characteristic of unaccusatives.

In the third person singular of the present tense, unaccusatives use the suffix *-a*, while transitives and unergatives use *-s*.

(18) c'er-s 's/he writes' Transitive
muša-ob-s 's/he works' Unergative
e-c'er-eb-a 'it is written to him/her' Unaccusative

[4] A few unergative verbs, such as *lap=arak=obs* 'talks, speaks=, correspond to a form with *e-*, such as *elap=arak=eba* 'talks to him/her=. But the latter forms, too, unexpectedly have the grammatical characteristics of unaccusatives (see Harris 1981: 270-273).

The 'unwillingness' construction uses *-a* in the third person singular subject form in the present tense, as shown in (12c), even though it is apparently transitive.

Finally, like most unaccusatives, the 'unwillingness' construction uses the suffix *-od* in certain tense-aspect-mood categories, contrasting with the *-d-* suffix used by transitives and unergatives.

(19)	c'er-d-a	's/he was writing'	Transitive
	mušaob-d-a	's/he was working'	Unergative
	kvd-eb-od-a	's/he was dying'	Unaccusative
	e-c'er-eb-od-a	'it was written to him/her'	Unaccusative
	še-mo-g-e-lanʒγ-eb-od-a	'if he unwillingly made you insult him'	'Unwillingness'

The 'unwillingness' construction is conjugated like an unaccusative verb except that it lacks forms in those tense-aspect-mood forms that are formed without a preverb. The preverb here is *še-mo-*, and it is discussed in detail below. The tenses formed with a preverb are illustrated in (20), keeping the persons of each argument the same.

(20)	še-mo-g-e-lanʒγ-eb-I	'I will unwillingly make you insult me'	FUTURE
	še-mo-g-e-lanʒγ-eb-od-I	'if I unwillingly made you insult me'	CONDITIONAL
	še-mo-g-e-lanʒγ-eb-od-e	'I would unwillingly make you insult me'	SUBJUNCTIVEII
	še-mo-g-e-lanʒγ-e	'I unwillingly made you insult me'	AORIST
	še-mo-g-e-lanʒγ-o	'I would unwillingly make you insult me'	OPTATIVE
	še-mo-g-e-lanʒγ-v-i-a	'I have evidently made you insult me'	EVIDENTIAL

The "unwillingness" construction seems morphologically closest the construction in (21), a relatively isolated (irregular) unaccusative that is sometimes interpreted as involving coreference ('he/she/it will hide [self] from you').

(21) da-g-e-mal-eb-a
PV-2O-E-hide-SM-3S
'he/she/it will hide from you'

I conclude from the evidence above that the 'unwillingness' construction is morphologically an unaccusative, similar to the unaccusative in (21). In spite of the fact that it is semantically causative and has a direct object, it has the grammatical characteristics of unaccusatives, including the prefixal vowel *e-*, the suffix combination *-eb-i*, and the suffix *-od*. Each of these contrasts with the morphology of regular transitive and unergative verbs.

2.3. Preverbs

The morphemes *še-* and *mo-* are both preverbs; *še-* has the core locative meaning 'in', and *mo-* is a cislocative, thus having the core meaning 'hither, toward the speaker or hearer', as illustrated in (22).

(22) (a) še-vid-a
in-went-3S
'S/he went in.'
(b) mo-vid-a
CISL-went-3S
'S/he came.'

The two preverbs can be combined compositionally.

(23) še-mo-vid-a
in-CISL-went-3S
'S/he came in.'

Generally, preverbs in Georgian may have three functions. The adverbial meaning, as illustrated in (22) and (23), is the core meaning. In appropriate contexts, any preverb can make a verb form perfective, often retaining some of its core adverbial meaning. In the grammar of Georgian, futures are the perfective of the present, and the preverb is thus present.

(24) še-it'ans
in-bring.3S
'S/he will bring it in, input it, introduce it'

(25) (a) i-g-eb-s
I-win-SM-3S
'S/he wins, is winning it.'

(b) mo-i-g-eb-s
PV-I-win-SM-3S
'S/he will win it.'

In (1), the preverbs also fill the function of marking perfectivity, and the tense-aspect-mood categories illustrated in (20) above are those that are grammatically perfective.

The third function of preverbs is that sometimes they combine with a specific root to create a lexeme whose meaning cannot be predicted compositionally. For example, *naxa* 'she saw it', combines with *še-* 'in' (and *i-*) to make a new lexeme *šeinaxa* 'she saved it, put it away, kept it'. Usually this function of preverbs occurs with single verbs, but there are also some sets of specialized meanings. For example, the preverb *c'a-* has the core adverbial meaning 'thither, far away' (translocative); but it has a further meaning that falls into this third function, 'for a little while' or 'superficially'. The examples below illustrate this in third person singular future forms.

(26)

it'irebs	'will cry'	c'a-it'irebs	'will cry a little'
daiʒinebs	'will sleep'	c'a-iʒinebs	'will sleep a little'
ivarǰišebs	'will exercise'	c'a-ivarǰišebs	'will exercise a little'
imušavebs	'will work'	c'a-imušavebs	'will work a little'

(Jorbenaʒe et al. 1988: 459)

The preverbs in (26) thus add a semantic element, 'for a little while', but do not alter the argument structure.

While the preverbs in (1) make the verb perfective, their core meanings do not add up to 'unwillingly'. Nevertheless, the 'unwillingly' interpretation is best attributed to these preverbs, since this meaning is of the same general type as that illustrated in (26), and since similarity to other constructions provides some evidence that the meaning of 'unwillingness' is carried here.[5] The selection restriction for "dark" predicates must also be associated with the preverbs, since the other morphology is more general. Thus, the combination of preverbs has a meaning that is non-compositional but still is definable, and an unusual selection restriction.

[5] There are related meanings; see Jorbenaʒe et al. (1988:431) for additional information.

2.4. Interim summary

I summarize by listing the elements of (1) that are straightforwardly accounted for by the regular morphology and those that are not.

(27) Elements of (1) that are straightforwardly accounted for by the regular morphology
- (a) first person subject (instigator)
- (b) second person object (the one who insults)
- (c) root meaning 'insult'
- (d) future tense, indicative mood, perfective aspect

(28) Elements of (1) not accounted for by the morphology
- (a) first person object (the one insulted)
- (b) 'unwillingness' meaning
- (c) limitation to "dark" predicates
- (d) causation
- (e) transitivity

3. Analysis

What is predictable should be accounted for by regular rules. I suggest that simple schemas be used to account for what is not predictable. I propose the schema in (29)

(29) [*še-mo-*PERSON.NUMBER.PREFIX-*e-*ROOT-TAM-PERSON.NUMBER.SUFFIX]$_{\text{CLASS 2 VERB}}$

Semantics:
subject: instigator (agent)
direct object: patient
indirect object: actor (agent)
'unwillingly'
Constraints: Adark predicate
direct object coreferential with subject

"TAM" here covers a sequence of suffixes. "PERSON.NUMBER.PREFIX" and "PERSON.NUMBER. SUFFIX", too, include up to two affixes each. The designation "Class 2" focuses on the morphological class, rather than the semantic class of unaccusatives, which is less easily defined in Georgian (see Harris 1981). By designating the whole schema as a class 2 verb we make it possible to predict the correct endings for the various TAM categories. For example, an unaccusative (Class 2) verb requires the suffix *-od* in the conditional and the subjunctive II in (20), while a transitive or unergative verb would require *-d* instead. Although the prefix *e-* might be predicted from the fact that this is an Aintransitive@ verb with an indirect object, since the verb is really transitive, I have preferred to specify the *e-* in the schema. The semantics of the arguments must be listed with the schema, since an unaccusative verb is not expected to have arguments with these theta roles. The semantics, >unwillingly=, and the constraints must be listed with the schema as well, since they are unique to it. Thus, it is essentially the use of an unaccusative (Class 2) verb form with the non-compositional preverb combination *še-mo-* that creates verb forms meaning >unwillingness=, together with the other idiosyncracies of these forms.

Gurevich (2006) presents a more complete analysis of Georgian grammar showing that many aspects of the grammar are best analyzed as form-meaning pairs (constructions) specifying associated morphology, syntax, and semantics. The construction described in the present paper is very restricted, but it shows that a similar

approach is needed to account for a specific construction that combines predictable and unpredictable elements.

4. Conclusion

The >unwillingness= construction presents a number of peculiarities. Morphologically it has the conjugation characteristics of an unaccusative, including the prefix *e-*, the tense-aspect-mood marker *-od*, and the endings *-eb-i* and *-eb-a*. Morphosyntactically, it requires that the subject and direct object be coreferential, though this is not indicated through any of the means usual for coreference. Semantically it has the meaning 'unwillingness', though this is not indicated compositionally. It has none of the characteristics of a causative, or even of a transitve.

I have proposed that these unexpected features can all be dealt with in terms of a schema that specifies these details in terms of the form of the word. This approach is situated within a construction- or schema-based view of morphology.

References

Amiridze Nino 2006. *Reflexivization strategies in Georgian*. Utrecht: Netherlands Graduate School of Linguistics (LOT).

Gurevich Olga 2006. *Constructional morphology: The Georgian version*. Dissertation, University of California, Berkeley.

Harris Alice C. 1981. *Georgian syntax: A study in relational grammar.* Cambridge: Cambridge University Press.

Harris Alice C. 2006. History in Support of Synchrony. *Berkeley Linguistics Society* 30: 142-159.

Jorbena∞e Besarion, M. K=obai∞e, and M. Beri∞e. 1988. *Kartuli enis morpemebisa da modaluri element=ebis leksik=oni.* [Dictionary of morphemes and modal elements of the Georgian language.] Tbilisi: Mecniereba.

'One Allomorph to rule them all':
The Single Allomorph Selection Constraint in Greek

Athanasios Karasimos
University of Patras, Greece
akarasimos@upatras.gr

1. Stem Allomorphy: Introduction

1.1. Stem Allomorphy under the spotlight

Research on stem allomorphy has been revived in Aronoff (1994), whose work has led to novel approaches of inflectional and derivational phenomena in morphological research by Booij (1997), Thornton (1997), Pirrelli & Battista (2000a, 2000b), Ralli (2000, 2007), Stump (2001), Bonami & Boyé (2003), Maiden (2004) among others. Aronoff's main idea also followed by other morphologists is that the signifiant of a lexeme is not a single phonological representation, but an array of indexed stems, which may stand in relations ranging from identity through regular phonological alternation, arbitrary change to full suppletion. (cf. Maiden 2004).

1.2. Definition of Allomorphy and Theoretical Framework

Expanding the definition of Lieber (1982: 27) for allomorphy, what we define as allomoprhs are the different varieties of the same morpheme, which share such lexical information as semantic representation and argument structure, but that differ unpredictably and arbitrarily in their phonological form and in the morphological environments in which they occur (for example κύμα~ κύματ 'wave', παίρν~ πήρ~ παρ 'take'). However, we have to point out that the term is over-used in the literature, since several allomorphs do not qualify for the proper conditions of Lieber's definition.

Ralli (2000, 2007) emphasizes that stem allomorphy is included in the core of morphology and participates in all word formation processes. She suggests that it is one of the basic features for categorizing verbal inflectional classes and nominal inflectional classes.

1.3. The Allomorphic Behavior Principle

Karasimos (2001, 2011) observes that the allomorphic behavior of a morpheme / lexeme is the same in all word formation processes with the unique and systematic exceptions of the *παιδί*-type and the *τεμπέλης*-type words. These kinds of nominal morphemes participate with new or different allomorphs in derivation and compounding (compared to inflection), since the *παιδί*-type nouns have no allomorphs (e.g. *παιδί / παιδι-ά* 'child / children' in inflection and the *τεμπέλης*-type nouns participate with a different allomorph than the one(s) used in inflection (e.g, INF *τεμπέλης: τεμπελη~ τεμπελησ* vs. DER/COM *τεμπελ* 'lazy').

The morphemes that display allomorphic behavior in word formation processes seem to be depend on their behavior in the process of inflection. The allomorphic behaviour principle determines inflection as the primary field of allomorphic comparison, since most allomorphs 'were created' from morphemes reanalysis and from inactive phonological an morphological rules in inflection (see more Karasimos 2011).

(1) i.

κύμα/ κύματ-α	κυματ-ίζω[1]	κυματ-ο-θραύστης
kima/ kimat-a	kimat-izo	kimat-o-thrafstis
'wave' / 'waves'	'to wave'	wavebreaker
καφέ-ς/ καφέδ-ες	καφεδ-άκι	καφεδ-ο-πωλείο
kafe-s/ kafedh-es	kafedh-aki	kafedh-o-polio
'coffee'/ 'coffees'	'small coffee'	'coffee shop'

ii.

καράβι/ καράβι-α	καραβ-ίσιος	καραβ-ό-σχοινο
karavi/ karavi-a	karav-isios	karavo-o-sxino
'ship'/ 'ships'	'shipborne'	'headrope'
βάρ-ος/ βάρ-η	βαρ-ίδι	βαρ-ό-μετρο
var-os/ var-i	var-iδi	var-o-metro
'weight'/ 'weights'	plumb	barometer
μπακάλη-ς/ μπακάληδ-ες	μπακάλ-ικο	μπακαλ-ό-γατος
bakali-s/ bakalidh-es	bakal-iko	bakal-o-γatos
'grocer'/ 'grocers'	'grocery store'	'employee in a grocery store'
νταή-ς/ νταήδ-ες	νταηδ-άκος	νταηδ-ό-μαγκας
dai-s/ daidh-es	daidh-akos	daiδ-o-magas
'bully'/ 'bullies'	'little bully'	'bully-and-bloke'

2. Comparing the word formation processes

Comparing the word formation processes of inflection and derivation (and compounding) based on the phenomenon of allomorphy, we can observe several trends among languages. There are languages, like German (2.a), where all the allomorphs of an inflectional paradigm participate in derivation and compounding, while in other languages, such as Dutch and Greek, the above behavior is unlikely to be found.

(2) a. German

Vater ~ Väter	Vaterland	-	Vättersitte
'father' – 'fathers'	'homeland'		'ethics of ancestors'
Mutter ~ Mütter	Mutterfreuden	-	Mütterverschickung
'mother' – 'mothers'	'mother's joy'		'mothers' decharge note'
Buch ~ Bücher	Buchbinder	-	Bücherfolge
'book' – 'books'	'bookbinder'		'series of books'

(Lieber 1982)

b. Modern Greek

άνθρωπ(ος)	άνθρωπ(οι)	ανθρωπ-ισμ(ός)	ανθρώπ-ιν(ος)
anθrop(os)	anθrop(i)	anθrop-ism(os)	anθrop-in(os)
'man'	'men'	'humanism'	'human'
αυλή	αυλ(ές)	αυλ-ικ(ός)	αυλ-αία
avli	avl(es)	avl-ik(os)	avl-ea
'yard'	'yards'	'courtier'	'curtain'
βήμα	βήματ(α)	βηματ-ίζ(ω)	βηματ-άρα
vima	vimat(a)	vimat-iz(o)	vimat-ara
'step'	'steps;	'stride'	'big step'

However, analyzing data from derivation in Modern Greek, we discover that all the different forms of a morpheme are not fully available during all the word formation

[1] In these paradigms, I do not separate the inflectional suffixes from the derivational. Also in compounding words, the second component is not separated into their morphemes.

processes. For example the noun *βήμα* 'wave' displays two allomorphs *βημα~ βηματ* in inflection, it displays only one morpheme form in the process of derivation (*βηματ~*), e.g. *βηματάκι* 'small step', *βηματάρα* 'big step', *βηματίζω* 'stride'. The same allomorphic pattern of this noun is observed in the process of compounding, in examples such as *βηματοδότης* 'pacemaker', *βηματομέτρηση* 'step counting'. As it is demonstrated in the following session (3.), such allomorphic behavior is not random and is solely due to a constraint that applies to all nominal and adjectival stems and suffixes.

3. The Single Allomorph Selection Constraint

The process of derivation contains principles and constraints about the categories of stems of derived words that are combined with derivational suffixes. The most common constraint for input-type cases (input constraint) is the number of requirements of a base-stem to be chosen and combined by derivational suffixes. There are a few derivational processes that require even more limited bases. Let's take some examples from German (Riehemann 1998: 54) and Modern Greek:

(3) a. *German*

essen 'eat'	essbar	'eatable'
zahlen 'pay'	zahlbar	'payable'
halten 'hold'	haltbar	'durable'

b. *Modern Greek*

ταξί 'taxi	ταξιτζής	'taxi driver'
καφές 'coffee'	καφετζής	'coffee shop owner'
τζάμι 'glass'	τζαμτζής	'glazier'
σαπούνι 'soap'	σαπουντζής	'soapmaker'

Analyzing the above examples from German, the derivational suffix *-bar* is combined only with base-stems that are transitive verbs to form adjectives; a similar case is the corresponding suffix *-baar* from Dutch (Booij 2006: 62), which requires the same context subcategorization. For example, the word *drink-baar* 'drinkable' is derived from the transitive verb *drink* 'drink'. In Modern Greek there are few derivational suffixes which are attached to specific stems, such as the suffix *-τζη(ς)* in (3.b) which combines only with nominal bases and stems. On the other hand there are several suffixes that do not present any combinational constraints with bases, since they can attach to nominal, adjectival and verbal stems (even adverbial bases). In English the suffix *-er* is combined with verbs (*keep* >> *keeper*, *print* >> *printer*) and nouns (*Berlin* >> *Berliner*). The greek verbal derivational suffix *–ευ(ω)* combines with nominal stems (*χορ-ός* 'dance' >> *χορεύω* 'dance') with adjectival stems (*ήμερ-ος* 'meek' >> *ημερεύω* 'tame') and with adverbial stems (*κοντά* 'close' >> *κοντεύω* 'getting close').

More specifically, we maintain that the non-appearance of all the allomorphs as bases in derivation is not random, occasional or not independent from the morphological environment in which the allomorphs appear. I have previously suggested (Καρασίμος 2011a, 2011b) that nominal bases of derived words and nominal stems as first constituents of compound words allow the appearance of only one allomorph due to a constraint that takes place in the input configuration of a derived or compound word; this constraint prevents the appearance of all allomorphs of each morpheme as basis on a nominal derived word. We call this specific limitation of selection **single-allomorph selection constraint**. I have observed that the Greek derived words fall under the restrictions of this constraint and some allomorphs that participate in inflection are excluded. Since the constraints do not usually apply randomly in a language, but for a reason, I point out that the single-allomorph selection constraint ensures unique

uniformity among derived words with a common basis through the appearance of only one form of the morpheme-base.

In the present article, I suggest that the non-appearance of all allomorphs as bases in derivation does not depend on the process itself, since the phenomenon appears also in compounding. I believe that in Modern Greek the reasons for the occurrence of single-allomorph selection constraint are independent; the selection of the single-allomorph is neither random and arbitrary, nor unpredictable. In fact, this particular constraint applies mainly for morphological and phonological reasons.

3.1. Optional phonological conditions

Checking carefully the phonological structure of a derived word base, the existence of stems with a final-character consonant is statistically more significant than with a final-character vowel. According to Ralli's inflectional model (2000)[2], the nouns in Modern Greek that display allomorphs, have an allomorph that ends with a vowel and another one that ends with a consonant (4.b). The inflectional classes with no allomorphs at all have in majority stems with ending of the XC-type[3] (4.a). The only and systematic exception is the sixth inflectional class of *καραβί*-type neutral nouns (4.c, see Karasimos 2011a). On the other hand, verbs with systematic allomorphy (5.a, second inflectional class according to Ralli's model (2004)) display a stem with a final-character consonant and a stem with a final-character vowel (X ~ XV); other verbs without systematic allomorphy (5.b, first inflectional class, see above) can have all their allomorphic types with a consonant as a thematic character.

(4) a.	άνθρωπ-ος		
	anθrop-os		'man'
	δάσ-ος		
	δas-os		'forest'
	μωρ-ό		
	mor-o		'baby'
b.	πάπα-ς	παπάδ-ες	
	papa-s	papaδ-es	'priest' – 'priests'
	θάλασσα	θάλασσ-ες	
	θalasa	θalasa	'sea' – 'seas'
	κύμα	κύματ-α	
	kima	kimat-a	'wave' – 'waves'
c.	καράβι	καράβι-α	
	karavi	karavi-a	'ship' – 'ships'
	τραπέζι	τραπέζι-α	
	trapezi	trapezi-a	'table' – 'tables'
	μολύβι	μολύβι-α	
	molivi	molivi-a	'pencil' – 'pencils'

(5) a.	αγαπ-ώ	αγάπη-σα	
	aγap-o	aγapi-sa	'love' – 'loved'
	πηδ-ώ	πήδη-σα	
	piδ-o	piδi-sa	'jump' – 'jumped'
	αφαιρ-ώ	αφαίρε-σα	

[2] Ralli (2000) suggests that there are eight inflectional nominal classes in Modern Greek. The first two contain masculine nouns, the third and fourth (include) feminine nouns and the rest neutral nouns.

[3] XC-type: any kind of phoneme (X) and a consonant (C).

	afer-o	afere-sa	'remove' – 'removed'
b.	πλέν-ω	έ-πλυν-α	
	plen-o	e-plin-a	'wash' – 'washed'
	ξεχν-ώ	ξεχα-σα	
	ksexn-o	ksexa-sa	'forget' – 'forgot'
	ρουφ-ώ	ρουφηγ-σα[4]	
	ruf-o	rufiγ-sa	'suck' – 'sucked'

Structurally, over eighty percent of the derivational suffixes that combine with nouns/ or nominal bases and adjectives/ or adjectival bases start/begin with a vowel. To maintain the optimal syllabic structure CV, the base "should" have a morpheme ending with a consonant (if applicable). Therefore, although it does not form a requirement, the allomorph ending with a consonant is qualified as the only and final choice. For example, the verbal suffix *–ιζ(ω),* when combined with the lexeme KYMA 'wave' which has the allomorphs *κυμα~ κυματ*; the allomorph *κυματ* is joined with the derivational suffix and that way the optimal syllabic structure is achieved. The highly productive subclass of diminutive and augmentative derivational suffixes have all their suffixes starting with vowel, like *-άκι*, *-άρα*,*-ούτσικος*, *-ούκλα-*, *-ούλα, -ίτσα* etc, among others.

I have to clarify here that I do not maintain that the derivational suffix selects the proper allomorph, but that the phonological structure of the suffix justifies the "choice" of the specific allomorph. Moreover, there are a few derived words with their optimal syllabic structure violated in the absence of an alternative morpheme, as for example *λα(ός)* 'people' > *λα-ικ(ος)* 'folkish', *μυστήρι(ο)* 'mystery' > *μυστηρι-ακ(ος)* 'mystic', *ράβ(ω)* 'sew' > *ράφ-τη(ς)* 'tailor'. Therefore, the derivational suffix does not require a specific form of a basis-morpheme, since if that was the case, then this characteristic would constitute a universal property of all suffixes, that is to select the proper form of a basis-morpheme.

3.2. Morphological conditions

All the derived words in Modern Greek have two components, a morpheme –base or a derived stem (non-head position) and a derivational suffix (head position) that applies to the basis. In the non-head position, the morphemes with two or three allomorphs cannot display all the possible forms. Therefore, every allomorph is marked with the context information and the morphological environment to avoid the possibilities of ungrammatical derived words (e.g. **ψυχη-ισμ(ός)* instead of *ψυχ-ισμ(ός)* 'psychism', **παίρνσιμο/ * πήρσιμο* instead of *πάρσιμο* 'taking', **βαθύως* instead of *βαθέως* 'deeply'. **κονταεύω* instead of **κοντεύω* 'approach').

The derivational suffixes "adopt" the same allomorphic behavior as the stems. Therefore, we expect that the suffixes share the same context information and morphological environment with their allomorphs. More specifically, when a derivational suffix, that is going to be combined with a simple or a derived stem, is followed by an another derivational suffix, it will be placed in the second level of derivation process in a non-head position and will display only one allomorph. Derived base is called the combination of a stem and a derivational suffix, e.g. [*ανθρώπ-ιν-*] 'human.NoINF[5]', [*χορ-ευ-*] 'dance$_V$.NoINF', [*ομορφ-ότερ-*] 'more beautiful.NoINF'.

Plag (1999) and Hay & Plag (2004) claim that the basis of a derived word demands a specific allomorph of the derivational suffix; we maintain that this fact can be modified in Modern Greek derivation. More specifically, the basis demands a specific allomorph of derivational suffix *if and only if* the derivational suffix participates in another derivational suffixation process; thus it is obliged to obey the single-allomorph

[4] The type 'ρούφηξα' is created after the phonological rule of dissimilation.
[5] NoINF = absence of inflectional suffix.

selection constraint. In English, such constraints apply only to specific groups of morphemes or suffixes. On the contrary, in Modern Greek each base selects the proper allomorph of a suffix, if it is any further suffixation (6.a). A similar procedure applies in cases of suffixation in compound words, such as *χορ-ο-πηδηχ-τ-ούλης* 'gamboler', *δικτυ-ο-τρομο-κρα-τ-ικ-ός* 'cyber-terroristic', *αερ-ο-μαχη-τ-ικ-ο* 'air-fighter plane' (6.b).

(6) a.	κλαίω	κλαι~ κλα	DER	κλά-μα/ κλά-ματ-α	>> κλα-ματ-άκι
	kleo	kle~ kla		kla-ma/ kla-mat-a	kla-mat-aki
	'cry'	'allomorphs of cry'		'crying / cryings'	'short crying'
b.	αέρας (αερα~ αερ) + μάχομαι		COM	αερ-ο-μαχη-τη-ς	αερ-ο-μαχη-τ-ικ-ο
	aeras (aera~ aer) + maxome			aer-o-maxi-ti-s	aer-o-maxi-t-ik-o
	'air'	'fight'		'pilot-fighter'	'air-fighter plane'

No information for the morphological environment of a non-head position is included in languages like German. The allomorphs of/in German are characterized by Lieber (1980) as singular and plural ones; she uses inflectional terms to categorize allomorphy. This characterization is not only morphological concerning the inflectional paradigm, but it is also semantic, because it implies the meaning of the singularity (SINGULAR) and quantity (PLURAL), as it is mentioned by Lieber (2.a). On the contrary, in Modern Greek it is possible for two allomorphs to occur in the same inflectional sub-paradigm, as for e.g. the nouns of IC8 (*βήμα*-type nouns) that 'use' the "plural" allomorph-t (*βήματ-α* 'steps', *κύματ-α* 'waves', *μαθήματ-α* 'lessons') also in the singular genitive (*βήματ-ος* 'of step', *κύματ-ος* 'of wave', *μαθήματ-ος* 'of lesson'). Furthermore, the allomorphs in Modern Greek are characterized by morphological information and do not contain any potentially hidden non-morphological information, such as semantic markedness. Additional morpho-semantic information for a lexeme, such as if it/ that is an Ancient Greek relic, puristic Greek, or a calque, is not attached to its allomorphs, since this information is not helpful for the selection of the proper allomorph in derivation and compounding. According to Booij (1997), the feature [+ CALQUE] in Dutch seems to cause a different behavior in lexemes and the selection of a non-expected allomorph.

3.3. 'Counterexamples' of single-allomorph selection constraint

Drachman (2006) gives some examples from Modern Greek, which seem to form exceptions of the single-allomorph selection constraint. He introduces the term 'shared allomorphs' which 'are produced' from a basic form, have an independent status in the word formation processes involved, but are more or less related to each other, so in essence the notion of 'derived from' or 'created from' is no longer necessary.

(7) a.	κρέ-ας/ κρέατ-α	κρεατ-ερό, κρεατ-ίλα, κρεατ-ινό
	kre-as/ kreata	kreat-ero, kreat-ila, kreat-ino
	'meat'/ 'meats' 'fleshy'	'smell-of-the-meat' 'of-meat'
		κρε-ο-πώλης, κρε-ο-φάγος, κρε-ο-κοφ-τήρας
		kre-o-polis, kre-o-fayos, kre-o-kof-tiras
		'butcher' 'meat-eater' 'minching machine'
		κρεατ-ο-πωλείο, κρεατ-ο-σανίδα, κρεατ-ο-πιτα
		kreat-o-polio, kreat-o-saniδa, kreat-o-pita
		'butcher's shop' 'meat-board' 'mince pie'
	αίμ-α/ αίματ-α	αιματ-άκι, αιμάτ-ωμα
	em-a/ emat-a	emat-aki, emat-oma
	'blood'/ 'bloods'	'hematoma'
		αιμ-ο-σφαίρια, αιμ-ο-δοσία, αιμ-ο-ρραγία

em-o-sferia, em-o-δosia, em-o-rajia
'blood capsule' 'blood donation' 'bleeding'
αιματ-ο-βαμμένος, αιματ-ο-κύλισμα, γλυκ-ο-αίματ-ος
emat-o-vamenos, emat-o-kilisma, γlik-o-emat-os
'bloodstained' 'carnage' 'sweet-blood'

b. (η)μέρα ημερίσιος (*μερίσιος), ημερομίσθιο (*μερομίσθιο)
(i)mera imerisios (*merisios), imeromisθio (*meromisθιο)
'daily' 'wage'
μερόνυχτα
meronixta
'days and nights'

(ε)νοίκι(ο) νοικιάζω, ενοικιαστής (*νοικιαστής), νοικιάρης (*ενοικιάρης)
(e)nici(o) nicjazo, enicjastis (*nicjastis), nicjaris (*enicjaris)
'rent$_V$' 'tenant' 'roomer'

(ο)μιλώ μίλησε, ομιλητής (*μιλητής)
(o)milo milise, omilitis (*militis)
'spoke' 'speaker'

(Drachman 2006: 14, 19)

The examples in (7a) seems that they seem to be counterexamples of the constraint presented here; however, the truth lies somewhere in the middle. First of all, I have to underline that this constraint has no exceptions of derived words with nominal and adjectival stems. The case of the noun αίμα 'blood' with the allomorphs *αιμα~ αιματ~ αιμ*, Karasimos (2001) has exhibited its peculiar behavior in the process of compounding. The derived words from *αίμα* 'blood' use only the allomorph *αιματ~*, obeying to the single-allomorph selection constraint. The compound words from *αίμα* are divided into three groups: (a) compound words with the stem *αιματ~ as* their second component, (b) compound words with the allomorph *αιμ~ as their* first component; these words came from French, Ancient Greek or International Greek and (c) compound words with the allomorph *αιματ~ as their* first component; these words came from English or from Modern Greek. As Ralli & Karasimos (2008, 2009a, 2009b) argue for the bare-stem constraint, the compound words formed in International Greek or by non-native speakers, violate the rules, principles and constraints of the compounding process, as they are created outside of the morphological word formation processes of the Greek language. Regarding the case of the noun *κρέας* 'meat' with the allomorphs *κρεα~ κρεατ~ κρε~*, its derived words follow the constraint without exceptions, while the word participates in compounding with two allomorphs (*κρεατ~ κρε*), since it is a word from Ancient Greek with a double inflectional paradigm (see Economou 1971: 85-86); therefore the 'relic-type allomorph' *κρε~* is used into words that were created in previous phases of Greek. Furthermore, the examples (7b) are in no way allomoprhs, as the optional phonological deletion of the initial vowel, by definition, does not constitute a case of allomorphy (see Karasimos 2011a).

Finally, it is necessary to point out that the single-allomorph constraint makes no exception in the whole process of derivation and applies to all nominal and adjectival stems (we except that there will not are also no exceptions in Modern Greek Dialects) and allows us to predict which allomorph is going to be used. In languages, like German (Lieber 1981) the single-allomorph selection constraint does not exist, but we expect that languages with extended allomorphy may display this constraint.

4. Conclusion

The derivational suffix does not force the base of a derived word to participate with the proper allomorph, as this would be a universal feature of all suffixes. The single-

allomorph selection constraint applies in Modern Greek, is definitely not a universal constraint, but may apply also in other allomorphically rich languages. This constraint refers to a morphological phenomenon (allomorphy) which changes are arbitrary and unpredictable; however the constraint is characterized by predictability and regularity. Furthermore, Karasimos (2011a) claims that this constraint provides us important advantages to analyze computationally this phenomenon, to export allomorphic rules and to improve the performance of a parser through predictability of allomorph selection.

References

Aronoff, Mark 1994. *Morphology by itself*. Cambridge, Mass: MIT Press.

Bonami, Olivie and Boyé, Gilles 2003. "La nature morphologique des allomorphies conditionnès : Les formes de liaison des adjectifs en français". In B. Fradin et al. (eds.), *Les unités morphologiques, tm. 3 de Silexicale*s:. 39-48.

Booij, Geert 1997. "Allomorphy and the Autonomy of Morphology". *Folia Linguistica XXXI/1-2*: 22-56.

Hay, Jennifer and Plag, Ingo 2004. "What Constrains Possible Suffix Combinations? On the Interaction of Grammatical Processing Restrictions in Derivational Morphology". *Natural Language and Linguistics Theory 22:* 565-596

Maiden, Martin 2004. "When lexemes become allomorphs. On the genesis of suppletion". *Folia Linguistica XXXVIII*: 227-256.

Nespor Marina 1999. *Φωνολογία*. Απόδοση: Αγγελική Ράλλη. Αθήνα: εκδόσεις Πατάκη.

Nespor, Marina & Vogel, Irene 1986. *Prosodic Phonology*. Dordrecht: Foris

Pirrelli, Vito and Battista, Marco 2000a. "The paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection". *Italian Journal of Linguistics (Rivista di Linguistica) vol. 12 num. II*: 307-380.

Pirrelli, Vito and Battista, Marco 2000b. "On the interaction of paradigmatic and syntactic stem alternation in Italian Conjugation". *Acta Linguistica Hungarica 47 (1-4):* 289-314.

Plag, Ingo 1999. *Morphological Productivity. Structural Constraints in English Derivation*. Berlin/ New York: Mouton de Gruyter.

Ralli, Angela 2000. "A Feature-based Analysis of Greek Nominal Inflection". *Glossologia 11-12*: 201-228.

Ralli, Angela 2007. "On the Role of Allomorphy in Inflectional Morphology: Evidence from the Dialectal Varieties of Lesvos, Kydonies and Moschonisia". In G. Sica (ed.) *Open problems in Linguistics and Lexicography*. Milano: Polimetrica.

Riehemann, Suzanne 1998. "Type-based Derivational Morphology". *Journal of Comparative Germanic Linguistics 2:* 44-77.

Stump, Gregory 2001. *Inflectional Morphology*. Cambridge: Cambridge University Press.

Thornton, Aanne-Marie 1997. "Stem allomorphs, suffix allomorphs, interfixes or different suffixes? On Italian derivatives with antesuffixal glides". In G. Booij, S. Scalise & A. Ralli (eds.) *Proceedings of 1st Mediterranean Meeting of Morphology*. Mytilene, Greece

Καρασίμος, Αθανάσιος 2008. "Η μη-εποπτευόμενη μορφολογική μάθηση και το πρόβλημα τη αλλομορφίας στην ελληνική γλώσσα". In A. Karasimos et al. (eds.) *Proceeding of 1st PICGL:* 123-138. Patras: University of Patras Press.

Καρασίμος, Αθανάσιος 2011a. *Υπολογιστική επεξεργασία της Αλλομορφίας στην Παραγωγή λέξεων της Ελληνικής*. Διδακτορική Διατριβή. Πανεπιστήμιο Πατρών.

Καρασίμος, Αθανάσιος 2011b. "Θεματική αλλομορφία στην Ελληνική Γλώσσα: ο περιορισμός της επιλογής μοναδικού αλλόμορφου". In M. Andreou et al. (eds.) *Proceedings of 2nd PICGL*, 85-94. Πάτρα: Πανεπιστήμιο Πατρών.

Complex adjectives at the morphology-syntax interface*

Hideki Kishimoto
Kobe University
kishimot@lit.kobe-u.ac.jp

1. Introduction

Japanese has several different types of adjectival expressions where a noun is combined with an adjective (see e.g. Namiki 1988, Yumoto 1990, 2009): among them, there is a fairly large inventory of adjectives with complex forms consisting of a noun plus the adjective base *nai* 'null, empty' (Kudo 2000). The following representative examples illustrate the morphological forms of the complex adjectives.[1]

(1) a. tawai(-ga)-na-i
solidity(-NOM)-null-PRES
'childish'

b. darasi(-ga)-na-i
punctuality(-NOM)-null-PRES
'untidy'

The complex adjectives in (1) show a peculiar morphological property, in the sense that nominative case marking can be placed optionally inside the complex form (without affecting the meaning of the whole).[2] The morphological composition of the compound adjectives suggests that Japanese has the grammatical process whereby complex forms are derived by incorporating the noun part into the adjective base, as illustrated in (2).

(2) $[_{TP}$ N-NOM $[_{Adj}$ *na*]-*i*] > $[_{TP}$ ~~N-NOM~~ $[_{Adj}$ N-NOM-*na*]-*i*] > $[_{TP}$ $[_{Adj}$ N-*na*]-*i*]

The compound forms without case marking can be assumed to emerge as a consequence of incorporating the dependent noun to the adjective base. On the other hand, when the noun bears case marking, it looks as if the noun remains unincorporated. As I will show, however, the case-marked noun may or may not be incorporated to the adjectival base.

The complex adjectives in (1) display one paradoxical property, in that their component noun, when case-marked, is transparent to the syntax in one respect, but it is

* This is a revised version of the paper presented at the 8th Mediterranean Morphology Meeting held at University of Cagliari (September, 2011). Part of the material was also presented in a seminar at the School of Oriental and African Studies, University of London (September, 2011). I am thankful for Akira Watanabe, Kaz Fukushima, Amir Zeldes, Mark Aronoff, Ryosuke Shibagaki, Peter Sells, Shin-Shook Kim, and the audience for comments and suggestions.

[1] To a large extent, complex adjectives are idiomatic. Most nouns included in the complex adjectives do not serve as independent words now, and cannot be used meaningfully in other contexts. For instance, *tawai* originally means 'solidity in ideas or behavior', but it is no longer used as a meaningful word. The types of complex adjectives where the nouns are case-marked are sometimes referred to as "idiomatic adjectives" (see Nishio 1985). The glosses of the component nouns are assigned based on the meanings inferred from the meanings of the whole expressions, or, if possible, on the meanings of their source words, which are no longer in use in contemporary Japanese.

[2] The presence or absence of case marking in the complex forms normally gives rise to semantic differences in the other classes of complex adjectival expressions, i.e. when compound adjectives have non-compound counterparts, the two variants usually do not express the identical meanings. For instance, the compound adjective *ha-gayui* (tooth-itchy) means 'irritating', but *ha-ga kayui* (tooth-NOM itchy) means 'the tooth is itchy', although, in some cases, the meanings of the paired expressions are approximately the same; e.g. *yoku-bukai* (desire-deep) 'greedy' and *yoku-ga hukai* (desire-NOM deep) 'mean'.

not in another respect. I suggest that the puzzling behavior comes from the availability of pre-incorporation structure for some syntactic operations, and claim that the partial sensitivity of the components of the complex adjectives to syntactic operations naturally follows from the morpho-syntactic constraints imposed on noun incorporation. The fact shows that despite lexical integrity, some syntactic operations may affect incorporated elements when their pre-incorporation structures are accessible syntactically.

The subsequent discussion proceeds as follows. In section 2, after advancing a tripartite classification of complex adjectives comprising *nai* (the Classes I, II and III), some notable features of the Class I and II adjectives are discussed. By comparing the Class III adjectives with the Class II adjectives, section 3 provides a confirmation that some syntactic operations are applicable to the Class II adjectives, owing to the syntactic visibility of the noun in the pre-incorporation position. Section 4 is a summary of the discussion.

2. Some properties of the complex adjectives

The complex adjectives under investigation have surface morphological forms in which the noun part appears with or without case marking. This suggests that in order to form compound adjectives with nouns, Japanese makes use of noun incorporation (see Spencer 1991, Mithun 1984 and others for discussion on the properties of noun incorporation), which presumably involves head movement, as often discussed (Baker 1988, 1996 and others), and also that the operation of noun incorporation is optional (at least in certain cases; see below).

The components of complex adjectives have varying degrees of tightness when the noun appears with case marking; according to the tightness of the noun+*nai* sequence, the complex adjectives can be divided into the following three classes.[3]

(3) a. Class I: *yurugi(-ga)-nai* (shake-NOM null) 'unshakable'; *sokke(-ga)-nai* (interest-NOM null) 'curt, blunt'; *atogusare(-ga)-nai* (later.concern-NOM null) 'without later trouble'; *abunage(-ga)-nai* (danger-NOM null) 'safe'; *nukari(-ga)-nai* (fault-NOM null) 'without any mistakes'; etc.

b. Class II: *sikata(-ga)-nai* (doing.way-NOM null) 'unavoidable'; *moosiwake(-ga)-nai* (excuse-NOM null) 'inexcusable'; *darasi(-ga)-nai* (tindiness-NOM null) 'untidy'; *tawai(-ga)nai* (solidity-NOM null) 'childish'; *hugai(-ga)-nai* (worth-NOM null) 'cowardly'; etc.

c. Class III: *syoo-ga-nai* (doing.way-NOM-null) 'cannot help'; *syoo-mo-nai* (doing.way-also-null) 'trivial'; *doo-siyoo-mo-nai* (how-doing.way-also-null) 'no way of doing'; *tohoo-mo-nai* (means-also-null) 'extraordinary'; *totetu-mo-nai* (reason-also-null) 'incredible'

As I will discuss below, the Class I adjectives have their case-marked nouns separate from the bases syntactically. The Class II adjectives have the constituent structure where the case-marked nouns are incorporated into the adjectival bases, but with the apparent partial transparency of their component nouns to the syntax. The Class III adjectives are ones where no syntactic operations are applicable even if their nouns are case-marked.[4]

[3] There is a tendency that the noun loses its transparency in meaning progressively from the Class I to the Class III adjectives. Thus, the native speakers tend to easily understand the meaning of the nouns in the Class I adjectives, but have difficulty in understanding the meaning of some nouns in Class II, if not all. In the Class III adjectives, some speakers do not even recognize that they have the morphological constituency until they are told.

[4] The number of the Class III adjectives, which do not allow any alternation that can be implemented in other classes, is fairly small. The suppression of nominative case marking on *syoo-*

Before proceeding, note that the negative expression *nai* can be used as a grammatical word—the functional negator—or an adjective (Kishimoto 2007, 2008). On the surface, the clause containing a complex predicate bears resemblance to a negated possessive clause, and the two types of clauses often carry similar meanings, as in (4).

(4) a. John-ga yuuki-ga na-i.
John-NOM courage-NOM NEG-PRES
'John does not have courage' (Possessive)
b. John-ga sokke-ga na-i
John-NOM interest-NOM null-PRES
'John is curt.' (Adjective: Class I)
c. Mary-ga tawai-ga-na-i.
Mary-NOM solidity-NOM-null-PRES
'Mary is childish' (Adjective: Class II)

Superficially, it looks as if the clauses in (4b-c) having complex adjectives as their predicates are of the same type as the negated possessive clause in (4a), but they are not.[5] The complex adjective clauses differ from the negated possessive clause, in that positive counterparts cannot be derived by replacing *nai* with *aru* 'be'.[6]

(5) a. John-ga yuuki-ga ar-u.
John-NOM courage-NOM have-PRES
'John has courage'
b. *John-ga sokke-ga ar-u.
John-NOM interest-NOM have-PRES
'John is not curt.'
c. *Mary-ga tawai-ga ar-u.
Mary-NOM solidity-NOM have-PRES
'Mary is adult-like (?).'

Needless to say, the affirmative forms are not available for the complex adjectives without overt case marking on the nouns.

(6) a. *John-ga sokke-ar-u.
John-NOM interest-have-PRES
'John is not curt.'
b. *Mary-ga tawai-ar-u.
Mary-NOM solidity-have-PRES
'Mary is adult-like (?).'

ga-nai is possible in colloquial speech, as in *syaa-nai (naa)*, whereas this shortened form is not derived from *syoo-mo-nai*. In general, when the Class III adjectives comprise the particle *mo*, it is not possible to remove *mo* from them. Note that both *syoo-ga-nai* and *syoo-mo-nai* contain the same noun '*syoo*', which has originated from *si-yoo* 'way of doing', but they are not construed as alternants because of their distinct lexical meanings (see section 3).

[5] The predicate can also be used as an existential predicate, taking a nominative phrase plus a locative argument, but this use is not directly relevant here.

[6] There are some other observable differences between the two types of clauses, but I will not discuss them in this paper. The positive form of *nai* is *aru* 'have'; the negative form *nai* might have been derived from dropping the verb *aru* from **ara-nai* 'have-NEG', which is not a possible form in Standard Japanese (see Kato 1985).

Even though complex adjective clauses sometimes bear resemblance to possessive clauses semantically, describing the attribute of the subject, certain syntactic differences are observed between them.

To return, one notable fact concerning the case-marked variants of the Class II adjectives is that their noun part is susceptible to certain syntactic operations even though it is incorporated into the adjectival base. I propose that in the Class II adjectives, both pre- and post-incorporation configurations are visible, and that some syntactic operations can apply to the component noun located in the pre-incorporation position, insofar as they do not violate the well-formedness conditions imposed on the incorporated head appearing inside the adjectival base by virtue of noun incorporation. In the following discussion, I argue that the difference between the Class I and the Class II adjectives emerges, depending on whether or not the case-marked noun is incorporated to the adjectival base.

2.1. Some divergent behavior

The two classes of complex adjectives with the noun+*nai* sequence (categorized as Classes I and II) display a number of differences in their syntactic behavior, which I argue should be attributed to the presence or absence of the noun incorporation. The main claim in this section is that in the Class I adjectives, when the component nouns are case-marked, their incorporation to the adjectival bases is not instantiated, but that the noun parts in the Class II adjectives need to be incorporated even if they bear case marking.

To begin, *nai* 'null, empty' is a predicate that takes a dative-nominative case-marking pattern when used transitively, but it also allows a nominative-nominative case-marking pattern. Thus, the subject of the Class I adjectives may bear dative as well as nominative marking, when the component noun bears nominative marking, as in (7).

(7) [John-no unten] {-**ga**/-**ni**} abunage-ga na-i
John-GEN driving{-NOM/-DAT} danger-NOM null-PRES
'John's driving is not dangerous/smooth.' (Class I)

Nevertheless, this 'nominative-dative' alternation is not possible with the Class II adjectives, even if the component noun is case-marked, as shown in (8).

(8) Mary{-**ga**/*-**ni**} tawai-ga na-i.
Mary{-NOM/-DAT} solidity-NOM null-PRES
'Mary is childish.' (Class II)

Furthermore, the dative marking on the subject is *not* possible when the noun part of the complex adjective does not appear with case marking.

(9) [John-no unten] {-**ga**/*-**ni**} abunage-na-i
John-GEN driving{-NOM/-DAT} danger-null-PRES
'John's driving is not dangerous/smooth.' (Class I)

In this respect, the Class II compound adjectives pattern with the Class I adjectives: when the component noun does not occur with a case marker, the subject can only have nominative marking, as shown in (10).

(10) Mary{-**ga**/-*-**ni**} tawai-na-i
Mary{-NOM/-DAT} solidity-null-PRES
'Mary is childish.' (Class II)

In these cases, whether or not the dative-nominative alternation is allowed is determined by the general requirement that at least one nominative argument is necessary in a clause (Shibatani 1978).

To illustrate how the alternation takes place, let us take a brief look at the (nominal) adjectives *hituyoo-da* 'necessary' and *taikutu-da* 'bored'. First, *hituyoo-da* 'necessary' is a transitive adjective, taking two arguments, so the experiencer subject is allowed to bear dative as well as nominative case, as there is another argument marked with nominative case.

(11) Ken{**-ga/-ni**} okane-ga hituyoo-da.
Ken-{NOM/-DAT} money-NOM necessary-PRES
'Ken needs money.'

This type of alternation is allowed only when an adjective is transitive. Thus, the nominal adjective *taikutu-da* 'bored' cannot have dative marking on its sole argument, as shown in (12).

(12) Ken{**-ga/***-**ni**} taikutu-da.
Ken-{NOM/-DAT} bored-PRES
'Ken is bored.'

The important fact is that in Japanese, intransitive adjectives do not allow the dative case marking on their subjects, since they do not take any other nominative arguments.

Now, given that the nominative noun of the Class I adjective is identified as an independent argument, whereas the nominative noun appearing in the Class II adjective is not, due to the incorporation of the noun, the difference between the two types of adjectives in regard to the nominative-dative alternation noted in (7) and (8) follows naturally.

(13) a. *abunage-ga nai*: [$_{TP}$ … abunage-ga [$_{ADJ}$ na]-i] (Class I)
b. *tawai-ga nai*: [$_{TP}$ … ~~tawai-ga~~ [$_{ADJ}$ tawai-ga-na]-i] (Class II)

In (8), unlike (7), the subject is the sole independent argument, as illustrated in (13), and hence, can only be marked with the nominative case, excluding the dative case.

When the noun part is combined with the adjective *nai* without case marking, both types of complex predicates have their component noun incorporated into *nai*.

(14) a. *abunage-nai*: [$_{TP}$ … [$_{ADJ}$ abunage-na]-i] (Class I)
b. *tawai-nai:* [$_{TP}$ … [$_{ADJ}$ tawai-na]-i] (Class II)

The examples in (9) and (10) suggest that the forms without case marking constitute single words by incorporation, i.e. the noun part is not an independent argument of the clause by virtue of the noun incorporation to the adjective base in both classes of complex adjectives.

The suppression of a case marker via incorporation is different from a case-marker drop, which often takes place in colloquial speech. This is evidenced by the fact that in (15), the dative marking on the subject can be retained even if the nominative marking on the other argument is dropped.

(15) John{**-ga/-ni**} sonna yuuki(-ga) na-i desyo!
John{-NOM/-DAT} such courage(-NOM) NEG-PRES EMPH
'John does not have such courage!'

The availability of dative marking on the experiencer in the possessive clause (15) shows that the nominative argument is not incorporated even if its case marking is dropped. (In the case of a case-marker drop, unlike the case of noun incorporation, an intonation break is imposed between the noun and *nai.*) The same case alternation pattern is found in the positive counterpart, as (16) illustrates.

(16) John{-**ga**/-**ni**} sonna yuuki(-ga) ar-i masu ka?
John{-NOM/-DAT} such courage(-NOM) have-PRES POLITE Q
'Does John have such courage?'

It is clear then that the presence or absence of overt case marking does not play a crucial role in determining the possibility of the dative-nominative alternation. The data illustrate that the nominative-marked noun in the Class II adjective is incorporated into *nai*, and does not serve as an argument syntactically, despite its case marking.[7] The important fact is that in the Class II adjectives, the dative marking is not available for the subject even when the component noun occurs with nominative case, because it does not count as an argument syntactically separate from the adjectival base.

Another fact indicating the difference between the two classes of complex adjectives concerns adverb insertion. The Class I adjectives allow an adverb to appear after the component noun if marked with nominative case, but this possibility is excluded when no case marking appears, as shown in (17).

(17) a. Kare-wa abunage-ga (mattaku) na-i.
he-TOP danger-NOM entirely null-PRES
'He is (entirely) without danger.'
b. Kare-wa abunage (*mattaku) na-i.
he-TOP danger entirely null-PRES
'He is (entirely) without danger.'

Since an adverb can, in general, intervene between an independent argument and a predicate, the fact suggests that the case-marked component noun is syntactically separate from the adjectival base in the Class I adjectives. In contrast, the Class II adjectives do not allow an adverb to intervene between the two elements even if the component noun appears with case marking, as illustrated in (18).

(18) a. Kare-wa tawai-ga (*mattaku) na-i.
he-TOP solidity-NOM entirely null-PRES
'He is (entirely) childish.'
b. Kare-wa tawai (*mattaku) na-i.
he-TOP solidity entirely null-PRES
'He is (entirely) childish.'

[7] There is one nominal adjective that comprises case marking in it (i.e. *wagama-da* 'selfish'), which is morphologically decomposable as *wa-ga-mama* (1.sg-NOM-will) 'do as one's will' (the nominative being originated as genitive historically). Needless to say, the case marking inside it is not visible to the syntax, so that no dative-nominative alternation is allowed for the subject of the adjective. The complex expression [*ki-o-tuke*]-*o suru* (mind-ACC-attach-ACC do) 'pay attention' provides another case. Here, the complex head *ki-o-tuke* 'attention' includes accusative case in it, but it can occur with another accusative marker, apparently running afoul of the so-called 'double-accusative' constraint. Obviously, this is due to the fact that the accusative case marking inside the complex head is not visible to the syntax.

The facts of adverb insertion suggests that the case-marked component noun is syntactically separate from the adjectival base in the Class I adjectives, but not in the Class II adjectives.[8]

An asymmetry in projecting negative scope provides another argument for the distinction between the two types of complex adjectival clauses. The example in (19) illustrates that an adverb like *sukosimo* 'at all' is an NPI, which needs to be licensed under negation.

(19) John-ga sukosimo hon-o {yoma-nakat-ta/*yon-da}.
John-NOM at.all book-ACC {read-NEG-PAST/read-PAST}
'John {did not read/*read} books at all.'

Sentence (20) shows that in the Class I adjectives, the adjectival base *nai* licenses an NPI adverb *sukosimo.*

(20) Kare-no unten-wa sukosimo abunage-ga nakat-ta.
he-GEN driving-TOP at.all danger-NOM null-PAST
'His driving was not dangerous at all.' (Class I)

This suggests that the adjective base *nai* acts as an operator that projects negative scope, presumably due to the fact that it carries a logical meaning virtually identical to the one expressed by a grammatical negator. Nevertheless, the same *nai* does not license the NPI when it is combined with the noun.

(21) *Kare-no unten-wa sukosimo abunage-nakat-ta.
he-GEN driving-TOP at.all danger-null-PAST
'His driving was not dangerous at all.' (Class I)

A comparison of (20) and (21) shows that the adjectival predicate *nai* can license an NPI by projecting negative scope when it occurs in isolation (with no noun being incorporated).

Note that a negative sentence like (22), which involves a case-marker drop, does not change the possibility of NPI licensing.

(22) Kare-ni-wa sukosimo sonna yuuki(-ga) na-i desyo!
he-DAT-TOP at.all such courage-NOM NEG-PRES EMPH
'He does not have such courage at all!'

The fact of NPI licensing in (22) illustrates that the noun whose case marker is dropped (in colloquial speech) is not incorporated even if it appears contiguous with *nai*.

Interestingly, the Class II adjectives do not license NPIs, even when their component nouns are case-marked.

(23) a. *Kare-wa sukosimo tawai-ga nakat-ta.
he-TOP at.all solidity-NOM null-PAST
'He was childish at all.' (Class II)

[8] In colloquial speech, the case marker on the noun part of the complex adjectives may be dropped, and in such a case, a prosodic break is placed between the noun and the adjective. If (17b) is taken to involve a case-marker drop in a colloquial register (with an intonation break between the noun and *nai*), the example might be acceptable even if an adverb intervenes between the noun and the adjectival base. In contrast, (18b) is simply rendered unacceptable with the intervening adverb, for no such possibility exists.

b. *Kare-wa sukosimo tawai-nakat-ta.
he-TOP at.all solidity-null-PAST
'He was childish at all.' (Class II)

The examples in (23) show that *nai* in the Class II adjective does not project negative scope regardless of whether or not the noun part is case-marked. This fact follows straightforwardly, given that the noun part of the Class II adjective is incorporated to *nai* even if it appears with nominative case marking.

A final fact indicating the difference between the two classes of complex adjectives concerns premodifiers used for the purpose of emphasis.[9] The addition of a prenominal modifier is possible with the Class I adjectives, as given in (24a), but this is not possible with the Class II adjectives, as shown in (24b).

(24) a. John-no unten-wa [{**korepotti-no/nan-no**} abunage]-mo nakat-ta.
John-GEN driving-TOP {slightest-GEN/any-GEN} danger-also null-PAST
'John's driving was least dangerous.' (Class I)
b. *Mary-wa [{**korepotti-no/nan-no**} tawai]-mo nakat-ta.
Mary-TOP {slightest-GEN/any-GEN} solidity-also null-PAST
'Mary was highly childish' (Class II)

The fact suggests that while the noun appearing in the Class I adjective has status as an independent argument, which allows for prenominal modification, the noun in the Class II adjective does not.[10]

Summarizing the facts observed so far, the complex adjectives belonging to the Classes I and II behave differently, in regard to the dative-nominative alternation on the subject, adverb insertion, and the addition of a prenominal modifier. The impossibility of these operations on the Class II adjectives would be naturally expected if the component nouns are incorporated into the adjective bases, even when marked with nominative case. When the nouns are incorporated to *nai*, the adjective does not project negative scope and hence fails to license NPIs. On the other hand, the nouns constituting part of the Class I adjectives should count as syntactically independent arguments that do not undergo incorporation when marked with nominative case.

2.2. Uniform behavior

The two types of complex adjectives (i.e. the Class I and II adjectives) show some distinct behavior when their component nouns are case-marked. Even so, there are cases where two classes of complex adjectives pattern together, indicating that some syntactic operations are applicable to the case-marked nouns regardless of whether or not they are incorporated to the adjectival base *nai*. There are a number of such manifestations. In this section, I suggest that in the Class II adjectives, syntactic operations can affect the case-marked component nouns in the pre-incorporation position, as long as the morpho-

[9] Even with the Class I adjectives, prenominal modification is fairly restricted, and only a limited set of prenominal modifiers can be used, if possible at all. These are minimizing expressions, and need to co-occur with the particle *mo*. Note that *mo* is allowed to attach to the noun in the two types of complex adjectives, without affecting their acceptability.

[10] The nouns can be coordinated in some Class I adjectives, which gives us another indication that the noun is syntactically independent of the adjectival base.

(i) [azi-mo sokke-mo] na-i
taste-also interest-also null-PRES
'neither tasteful nor interesting'

syntactic conditions constraining their later incorporation to the adjectival base are not violated.

A first argument in support of the present view may be adduced from the applicability of the nominative-genitive conversion to the noun parts of the complex adjectives. In Japanese, nominative case marking on a noun can be changed to genitive case when the expression is embedded under a larger nominal (as in a relative clause or a noun-complement clause) (see Harada 1971, Watanabe 1996, and others).

(25) a. [John{-**ga**/-**no**} hasit-ta] koto
John{-NOM/-GEN} run-PAST fact
'the fact that John's driving is safe'
b. [John{-**ga**/-**no**} yon-da] hon
John{-NOM/-GEN} read-PAST book
'the book which John read'

The nominative-genitive alternation is optional, as seen in (25). This alternation is possible with the Class I and II adjectives. The examples in (26) represent a case of the Class I adjectives, and show that when the noun part of the complex adjective appears with case marking, the nominative-genitive conversion is allowed.

(26) a. [John-no unten-ni abunage{-**ga**/-**no**} na-i] koto
John-GEN driving-DAT danger{-NOM/-GEN} null-PRES fact
'the fact that John's driving is safe'
b. [abunage{-**ga**/-**no**} na-i] unten
danger{-NOM/-GEN} null-PRES driving
'safe driving'

The examples in (27) are cases involving the Class II adjective *tawai-ga nai*.

(27) a. [tawai{-**ga**/-**no**} na-i] koto
solidity{-NOM/-GEN} null-PRES fact
'the fact that it is childish'
b. [tawai{-**ga**/-**no**} na-i] hito
solidity{-GEN/-NOM} null-PRES man
'a childish man'

The fact indicates that the case marking appearing in the incorporated noun may be susceptible to the nominative-genitive case conversion.

Second, both types of adjectives (optionally) allow the occurrence of an emphatic particle *mo* 'also' to the right of the noun part. (28) shows that the Class I adjective allows nominative marking on the component noun to be replaced with *mo*.[11]

(28) Mary-no unten-wa abunage{-**ga**/-**mo**} nakat-ta-si….
Mary-GEN driving-TOP danger{-NOM/-also} null-PAST-and
'Mary's driving was without danger, and …'

When *mo* is added, the original nominative case marking is suppressed. The Class II adjective behaves in the same way here, since the particle *mo* can appear with the component noun, as seen in (29).

[11] For one reason or another, when *mo* is simply added, an adverbial use of the adjective is often preferred (see Nishio 1972).

(29) Mary-no kotoba-wa tawai{-**ga/-mo**} nakat-ta.
Mary-GEN speech-TOP solidity{-NOM/-also} null-PAST
'Mary's talk was childish.'

Both classes of complex adjectives allow an alternation between the particle *mo* and nominative case marking.

There is yet another indication that the incorporated noun in the Class II adjectives can undergo syntactic operations. This can be seen in the case conversion that applies under nominalization (Sugioka 1992, Kageyama 1993, Kishimoto 2006). In Japanese, when a clause is nominalized by attaching the suffix *-sa* to an adjective, an obligatory change of nominative to genitive case marking takes place, and no adverbial particle is admitted, as seen in (30) (see Martin 1975, Kishimoto 2005).

(30) a. Mary{-**ga/-mo**} kawai-i.
Mary{-NOM/-also} cute-PRES
'Mary is (also) cute.'
b. Mary{-**no/***-**ga/*-mo**} kawai-sa
Mary{-GEN/-NOM/-also} cute-NOML
'Mary's cuteness'

This case-marking change takes place on the Class I and II adjectives. Example (31) shows that under nominalization, the Class I adjectives allow only the genitive marking on the component noun, which replaces nominative marking.

(31) abunage{-**no/***-**ga/*-mo**} na-sa
danger{-GEN/-NOM/-also} null-NOML
'no danger'

Further, (32) shows that the Class II adjectives, in which the noun is incorporated to the adjective *nai*, invoke an obligatory case-marking change under nominalization.

(32) tawai{-**no/***-**ga/***-mo**} na-sa
solidity{-GEN/-NOM/-also} null-NOML
'childishness'

In the Class II adjectives, the nominative noun is incorporated to *nai*, but still, must undergo the case-marking change under nominalization, and the presence of the particle *mo* is excluded in the nominalized form.

Now, the question that arises with regard to the Class II adjectives is why they behave in the way they do. While maintaining the analysis taking the two classes of complex adjectives to differ as to whether the case-marked noun is incorporated, I suggest that the mixed behavior of the Class II adjectives arises due to the syntactic visibility of the nouns in pre-incorporation structures, as indicated in (33b).

(33) a. Class I: $[_{TP}$ …. Noun-NOM $[_{ADJ}$ na]-i]
b. Class II: $[_{TP}$ …. ~~Noun-NOM~~ $[_{ADJ}$ Noun-NOM na]-i]

Here, syntactic operations are assumed to access the component nouns outside the adjectival base in both Class I and Class II adjectives.

It is easy to see that some syntactic operations are applicable to the component nouns of the Class II adjectives (as well as the Class I adjectives), provided their pre-incorporation structures are visible. In fact, syntactic operations are allowed on the Class

II adjectives, insofar as they do not affect the morpho-syntactic well-formdeness of the post-incorporation structures: the genitive-nominative conversion and the addition of the particle *mo* are both possible, because these operations only affect the head elements. The same holds for the obligatory change on case marking under nominalization. Even when these operations are applied, the nominal elements retain their syntactic status as heads; hence they can undergo noun incorporation and the resultant post-incorporation structures are well-formed.[12]

Nevertheless, other syntactic operations fail to operate on the Class II adjectives by virtue of the noun incorporation: in the Class II adjectives in which the noun is case-marked, the dative-nominative alternation on the subject and adverb insertion are not possible, because the case-marked component noun loses the status of an independent argument by virtue of its incorporation to the adjectival base. Prenominal modification is prevented from applying to the complex adjectives, since a phrasal element cannot be incorporated. Further, *nai* does not project negative scope, because it is combined with the noun; when combined with a noun, *nai* no longer serves an operator that licenses an NPI outside.

In the present proposal, syntactic operations may apply to the pre-incorporated nouns in the Class II adjectives. This analysis would be reasonable if, as often discussed, syntactic operations are banned on parts of words by lexical integrity (see Lapointe 1980, Selkirk 1982, Di Sciullo and Williams 1987, Bresnan and Mchombo 1995). Arguably, however, exceptions to this constraint are occasionally observed (see Haspelmath 2002). Given this fact, one might ask whether there is the possibility that the first group of operations applies to the word-internal elements inside the incorporated bases. If these operations can apply to the word-internal elements, the facts might be accounted for without reference to the pre-incorporation structure. This cannot be the case, however. I will turn to this discussion in the next section.

3. The Class III adjectives

In the present perspective, syntactic operations are not accessible to word-internal elements. Under the analysis taking syntactic operations not to apply parts of words, we would expect that all the alternations discussed in the preceding section should be unavailable for the adjectival expressions whose pre-incorporation structure is not accessible. On the other hand, if there are any syntactic operations that can look into word-internal elements, no such possibility should exist. The adequacy of the proposal that the noun appearing inside the adjectival base is not targeted by any syntactic operations due to lexical integrity can be readily confirmed. The Class III adjectives provide a case in point.

(34) a. syoo-ga-nai
doing.way-NOM-null
'cannot help'

b. syoo-mo-nai
doing.way-also-null
'trivial'

c. doo-siyoo-mo-nai
how-doing.way-also-null
'no way of doing'

What is remarkable about the Class III adjectives is that they do not participate in any operations discussed above, even if they carry a case marker or a particle.[13]

[12] Needless to say, the presence of the pre-incorporation structures can be confirmed only when the nouns are case-marked, i.e. when the grammatical marking is available for the incorporated nouns. When no such accompanying case-markers appear in the complex adjectives, it is not possible to check whether or not a pre-incorporation structure is visible to the syntax.

[13] *Syoo-ga nai* is categorized as a Class II adjective when it is taken to describe the property of a human individual, meaning 'worthless'. This can be seen by the fact that the adjective, when used in this sense, can participate in the nominative-genitive conversion, as in (i).

To be concrete, let us illustrate how the Class III adjectives behave with regard to the diagnostics discussed thus far. First, (35) shows that the dative-nominative alternation on the subject is not possible with this class of adjectives.

(35) a. Hanasi{-**ga**/*-**ni**} syoo-mo-na-i.
story{-NOM/-DAT} doing.way-also-null-PRES
The story is boring.'
b. Sore{-**ga**/*-**ni**} syoo-ga-na-i.
that{-NOM/-DAT} doing.way-NOM-null-PRES
'That cannot be helped.'

The noun part of the adjective *syoo-ga-nai* in (35b) bears nominative case marking, but the nominative marking appearing on its subject cannot be changed to dative marking. This suggests that the nominative-marked noun in the complex adjective is not identified as a syntactically independent argument.

Second, (36) shows that an adverb cannot intervene between the component noun and the adjectival base, illustrating that the nominative noun constitutes part of the adjective base.

(36) Sore-wa syoo-ga (*mattaku) na-i.
that-TOP doing.way-NOM entirely null-PRES
'That cannot be helped (entirely)'

Third, the adjectival base *nai* does not license an NPI even if the component noun is case-marked.

(37) *Sore-wa sukosimo syoo-ga-na-i.
that-TOP at.all doing.way-NOM-null-PRES
'That cannot be helped.'

(37) suggests that that *nai* does not stand alone, i.e. the component noun is combined with the base syntactically, despite the presence of nominative case marking on the noun. In addition, no prenominal modifier can be added to this adjective, as shown in (39).

(38) *[{**korepotti-no/nan-no**} syoo]-mo-na-i
{slightest-GEN/any-GEN} doing.way-also-null-PRES
'the least trivial'

These are the properties that the Class III adjectives share with the Class II adjectives, which indicate that the noun part of the complex adjective is incorporated to the adjectival base, rather than serves as an independent argument separate from the base.

Furthermore, the case-marked nouns of the Class III adjectives do not undergo syntactic operations that are allowed for the case-marked nouns included in the Class II adjectives. Thus, the nominative-genitive conversion, which is available for the other classes of adjectives, is *not* possible with the Class III adjectives.

(39) a. [Sono-ziko-ga syoo{-**ga/*-no**}-nai] koto
that-accident-NOM doing.way{-NOM/-GEN}-null thing

(i) Ken-wa [syoo{-ga/-no} na-i] yatu da.
Ken-TOP doing.way{-NOM/-GEN} null-PRES fellow COP
'Ken is not a worthless fellow.'
The Class II adjective *syoo-ga nai* is a distinct lexical item from the Class III *syoo-ga-nai*.

'the fact that that accident is unavoidable'

b. [syoo{-**ga/*-no**}-nai] ziko
doing.way{-NOM/-GEN}-null accident
'the unavoidable accident'

The genitive case marking is not possible with the adjective *syoo-ga-nai*, showing that no alternation can be implemented with this adjective, despite the presence of nominative case marking on the component noun.

Next, the Class III adjectives do not allow an operation substituting an adverbial particle for their nominative case marker.

(40) Sore-wa doo-siyoo{-**mo**/*-**ga**}-na-i.
that-TOP how-doing.way{-also/-NOM}-null-PRES
'There is no way of doing.'

(40) shows that an alternation between nominative case and *mo* is not allowed in the Class III adjectives.

Incidentally, *syoo-mo-nai* (doing-way-also-null) 'trivial' comprises the adverbial particle *mo* inside, but it does not result from an optional operation substituting *mo* for the nominative case on *syoo-ga-nai* (doing-way-NOM-null) 'cannot help'. Although, historically, both expressions have been originated from the same form, as often mentioned in Japanese dictionaries (e.g. *Nihon Kokugo Daiziten* [A Grand of Dictionary of the Japanese Language]), these two expressions constitute distinct lexical items now, i.e. *syoo-mo-nai* is not construed as an alternant of *syoo-ga-nai*, since they carry distinct lexical meanings. If these two expressions are related by the particle replacement operation that can be instantiated in other classes of complex adjectives, they should carry the same basic meaning, but this is not the case. Given this fact, it is safe to state that no alternation replacing nominative case with *mo* is available for the Class III adjectives.

Finally, in the Class III adjectives, no genitive marking appears even under nominalization, and also, the forms with the particle *mo* are acceptable despite nominalization, as shown in (41).

(41) a. syoo{-**ga/*-no}**-na-sa
doing.way{-NOM/-GEN}-null-NOML
'unavoidability'

b. syoo*(-**mo**)-na-sa
doing.way-also-null-NOML
'triviality'

c. doo-siyoo*(**-mo**)-na-sa
how-doing.way-null-NOML
'no way of doing'

In the complex adjectives belonging to the Classes I and II, the morphological change of nominative to genitive case on the noun part of the adjective is obligatory. By contrast, in the Class III adjectives, this obligatory change of case marking is not implemented even under nominalization, and in fact, if the change is instantiated, ungrammaticality results. Moreover, with the adjectives comprising *mo* inside, the particle must be retained even under nominalization, as shown in (41b-c), which is unacceptable for the other classes of complex adjectives. The fact can be taken as a sign that there is no visible component noun outside the adjectival head for the Class III adjectives.

To summarize, (42) shows the observed patterns for the three classes of complex adjectives with regard to (A) the nominative-dative alternation (on the subject), (B)

adverb insertion, (C) NPI licensing, (D) the addition of a prenominal modifier, (E) the nominative-genitive conversion (on the noun part of the adjective), (F) *mo*-replacement, and (G) obligatory change from nominative to genitive case (and exclusion of the particle *mo*) under nominalization.

(42)		A	B	C	D	E	F	G
Class I	*abunage-ga nai*	√	√	√	√	√	√	√
Class II	*tawai-ga nai*	*	*	*	*	√	√	√
Class III	*syoo-ga-nai*	*	*	*	*	*	*	*

As seen in (42), the syntactic operations applicable to the complex adjectives are divided into two types. (A), (B), (C) and (D) may operate on the Class I adjectives, but not on the Class II adjectives. On the other hand, (E), (F) and (G) are applicable to both the Class I and II adjectives. Importantly, the Class III adjectives are not susceptible to any of them. The Class III adjectives behave differently from the Class II, as well as the Class I adjectives.

If (A), (B), (C) and (D) are possible, the component nouns of complex adjectives should be syntactically separate from *nai*, as we have seen in section 2.1. Since these properties are not found in the Class III adjectives, the component noun of the Class III adjectives can be assumed to form a complex head unit with the adjective *nai*, as depicted in (43).

(43) a. *syoo-ga-na-i*: $[_{TP}$ .... $[_{Adj}$ syoo-ga-na]-i]
b. *syoo-mo-na-i*: $[_{TP}$ .... $[_{Adj}$ syoo-mo-na]-i]

If lexical integrity holds for the complex head comprising the noun and *nai*, the nouns included in the adjective *nai* fall outside the domain where the syntactic operations are applicable. In fact, since (E), (F) and (G) are not allowed for the Class III adjectives, besides (A), (B), (C) and (D), it is reasonable to conclude that word-internal elements are not targeted by any syntactic operations, i.e. lexical integrity is respected in the Class III adjectives, with no component noun being found outside the complex adjectival base.

Meanwhile, the Class I adjectives show a sign that the component nouns are not included within the adjectival heads. (A), (B), (C) and (D) are possible with this class of adjectives, showing that their component nouns are syntactically separate from *nai*. The same Class I adjectives, whose component nouns should not be included in the adjectival bases, allow (E), (F) and (G). In contrast, the Class II adjectives do not allow (A), (B), (C) and (D), suggesting that their component nouns are included in the adjectival base. Nevertheless, the Class II adjectives allow (E), (F) and (G), which suggests that their component nouns should be found outside the complex heads.

The impossibility of (E), (F) and (G) on the Class III adjectives would not be expected if they operated on parts of complex heads. Thus, the facts of the Class II adjectives lead to the conclusion that their component nouns in pre-incorporation position can be accessed by syntactic operations, while they later form complex heads with the adjective *nai*, via noun incorporation, which can affect only heads (Baker 1988 and others). In essence, the Class II adjectives resist (A), (B), (C) and (D), which do not square with the process of noun incorporation. Nevertheless, (E), (F) and (G) are applicable to the case-marked component nouns of the Class II adjectives, thanks to the visibility of the pre-incorporation configurations: even if the component nouns found in pre-incorporation position undergo (E), (F) and (G), they can be incorporated to the adjectival bases, without violating the well-formedness conditions constraining the post-incorporation structures.

4. Conclusion

Japanese complex adjectives where nouns are combined with the adjectival base *nai* can be classified into three classes (Classes I, II, and III), distinguished according to the degree of the tightness in the noun+*nai* sequence. When the noun is case-marked, the noun part superficially looks like an element that is syntactically independent of the adjectival base. This is not necessarily the case, however. The noun parts of the Class II and III adjectives are incorporated to the adjective base, even when they appear with case marking.

The Class II adjectives behave as if their case-marked component nouns are transparent to the syntax in one respect, but they are not in another respect. I propose that in the Class II adjectives, the pre-incorporation as well as the post-incorporation structure is relevant for the determination of their well-formedness. In the Class II adjectives, syntactic operations can apply to the component nouns in pre-incorporation structures as long as well-formed post-incorporation structures can be derived. Since, in the Class II adjectives, the component nouns in both pre- and post-incorporated positions are visible, they behave as if they are sometimes included in the adjectival base, and sometimes they are not.

The data from the Japanese complex adjectives illustrate that syntactic operations do not directly target incorporated heads (i.e. the elements that have undergone noun incorporation), but that some operations may affect the incorporated elements, despite lexical integrity, when their pre-incorporation configurations are visible (or accessible) syntactically.

References

Baker, Mark 1988. *Incorporation: A Theory of Grammatical Function Changing.* Chicago: University of Chicago Press.

Baker, Mark 1996. *The Polysynthesis Parameter*. Oxford: Oxford University Press.

Bresnan, Joan, and Sam Mchombo 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13: 181-254.

Di Sciullo, Anna Maria, and Edwin Williams 1987. *On the Definition of Words*. Cambridge, MA: MIT Press.

Fukushima, Kazuhiko. 1998. Negation in the Lexicon, not in NegP. *CLS* 34-1, 171-184.

Harada, Shin-Ichi. 1971. Ga-no conversion and idiolectal variations in Japanese. *Gengo Kenkyuu* 60: 25-38.

Hathpelmath, Martin 2002. *Understanding Morphology*. London: Arnold.

Kageyama, Taro 1993. *Bunpoo-to Gokeisei* [Grammar and Word Formation]. Tokyo: Hituzi Syobo.

Kato, Yasuhiko 1985. *Negative Sentences in Japanese*. Sophia Linguistica XIX.

Kishimoto, Hideki 2005. *Toogokoozoo-to Bunpookankei* [Syntactic Structure and Grammatical Relations]. Tokyo: Kurosio Publishers.

Kishimoto, Hideki 2006. Japanese syntactic nominalization and VP-internal syntax. *Lingua* 116: 771-810.

Kishimoto, Hideki 2007. Negative scope and head raising in Japanese. *Lingua* 117: 247-288.

Kudo, Mayumi 2000. Hitei-no hyoogen [Negative expressions]. In Yoshio Nitta and Takashi Masuoka (eds.) *Toki, Hitei-to Toritate* [Time, Negation, and Emphasis], 94-150. Tokyo: Iwanami.

Kuno, Susumu 1973. *The Structure of the Japanese Language*. Cambridge, Mass.: MIT Press.

Lapointe, Steven 1980. *The Theory of Grammatical Agreement.* PhD. dissertation. University of Massachusetts, Amherst.

Martin, Samuel 1975. *A Reference Grammar of Japanese*. New Haven: Yale University Press.

Mithun, Marianne 1984. The evolution of noun incorporation. *Language* 60: 847-895.

Namiki, Takayasu 1988. Fukugoogo-no Nichi-ei-taisho: Fukugoo-meisi/Fukugoo-keiyooshi [Complex words in contrast between English and Japanese: Compound nouns and compound adjectives]. *Nihongogaku* 7-5: 68-78.

Nihon Daiziten Kankokai 2000-2002. *Nihon Kokugo Daiziten* [A Grand Dictionary of the Japanese Language], 2nd ed. Tokyo: Shogakkan.

Nishio, Toraya 1972. *Keiyooshi-no Imi Yohoo-no Kijutsuteki Kenkyuu* [Descriptive Studies on the Meaning and Use of Adjectives]. Tokyo: Shuuei Shuppan.

Nishio, Toraya 1985. Keiyooshi-kanyooku [Adjective idioms]. *Nihongogaku* 4-1: 45-53.

Selkirk, Elizabeth 1982. *The Syntax of Words*. Cambridge, MA: MIT Press.

Shibatani, Masayoshi 1978. *Nihongo-no Bunseki* [An Analysis of Japanese]. Tokyo: Taishukan.

Spencer, Andrew 1991. *Morphological Theory.* Oxford: Blackwell.

Sugioka, Yoko 1992. On the role of argument structure in nominalization. *Language and Communication* 10: 53-80. Keio University.

Watanabe, Akira 1996. Nominative-genitive conversion and agreement in Japanese: A cross-linguistic perspective. *Journal of East Asian Linguistics* 5: 373-410.

Yumoto, Yoko 1990. Nitieitaishoo fukugookeiyooshi-no koozoo: Meishi+keiyooshi/keyoodooshi-no kata-nituite [A Japanese-English contrastive study of the structure of complex adjectives: On the patterns of noun plus adjective/nominal adjective]. *Gengobunka Kenkyuu* 16: 353-372.

Yumoto, Yoko 2008. On the modularity of word formation seen from complex adjective formation. In Yoko Yumoto and Hideki Kishimoto (eds.) *Goi-no Imi-to Bunpoo* [Lexical Meanings and Grammar], 209-229. Tokyo: Kurosio Publishers.

Analogical modelling and paradigmatic word formation as attention-seeking devices

Ewa Konieczna
Instytut Filologii Angielskiej
eakon@poczta.onet.pl

1. Introduction

This paper is an attempt at showing that analogical modelling (Booij 2007; Plag 2006; Krott 2001) and paradigmatic word formation (Booij 2007, 2010) are successfully used as the so called *attention-seeking devices* (Lipka 1987, 2000) in the journalistic register in modern English. As Lipka (2000) and Hohenhaus (2007) maintain, naming and syntactic recategorisation are not the only functions of WFs, which, among other things, can also serve as attention seeking devices (ASDs). ASDs are often employed in the texts aimed at attracting the reader's attention, such as press articles and advertisements. The purpose for which they are used is the formation of creative coinages that tend to be *noteworthy*, or in other words, *foregrounded*, which means that they draw attention to themselves (Hohenhaus 2007: 16) by being not, or at least not completely rule-governed, or by being a play on words well-established in the lexicon.

Thus, the aim to be pursued here is to provide some partial evidence that the outputs of both analogical modelling and paradigmatic word-formation are successful in seeking the reader's attention. This is achieved by the fact that, as Crystal (1998: 104) has put it, they [...] *appeal directly to our ludic sensibilities.* Their playfulness is accomplished by manipulating the language, understood as bending and breaking its rules: *the unexpected language attracts our attention, making us read a piece which we might otherwise have passed over* (Crystal 1998: 105). The analysis to follow is going to be carried out within the framework of Construction Morphology (Booij 2007, 2010) in which the notion of analogy as well as paradigmatic word-formation occupy an important place.

2. Analogy in Construction Morphology

The debate as to whether innovative complex words are created through analogy, or by means of word-formation rules seems far from being solved (see, e.g. Bauer 2001; Becker 1990; Derwing and Skousen 1989). Construction Morphology seems to offer reconciliation between these two divergent attitudes, as it assumes a hierarchical lexicon with different levels of abstraction, which is tantamount to co-existence of analogical word-formation and word formation based on abstract schemas, which differ in terms of the degree of their abstractness (Booij 2007, 2008). In the light of this theory, there are some clear cases of analogical word-formation, where an individual word serves as a model, which can be exemplified by the set of analogically coined compounds in Dutch (Booij 2010: 89):

(1)

existing compound with an idiosyncractic meaning	*analogically coined compound*
a. *angst-haas* 'lit. fear-hare, terrified person'	*paniek-haas* 'lit. panic-hare, panicky person'
b. *moeder-taal* 'lit. mother language, native language'	*vader-taal* 'lit. father-language, native language of father'

With these compounds it is indeed possible to point to one particular compound used as a model for the formation of a novel compound and, what is more, one can retrieve the meaning of this novel compound only when being acquainted with the idiosyncratic meaning of the model compound.

However, an analogical pattern may become subject to generalization and turn into a constructional idiom, that is a schema which generalises across a set of existing complex words. An example of that may be the word *Watergate*, which functioned as a model for a number of words, denoting a political scandal. Since with the passage of time a significant amount of words, ending in *-gate* has been formed, it cannot be now stated that the word *Watergate* was always used as a model. Instead, it seems quite plausible that given a certain number of compounds, containing a combining form *-gate,* language users abstracted from them the following schema:

(2)
$[[x]_{Ni}\ [gate]_{N}]_{Nj} \leftrightarrow [\text{political scandal pertaining to } SEM_i]_j$

Still, the formation of a symbolic schema, containing *-gate* does not mean that there is no link between the novel word ending in *-gate* and the model word *Watergate.* According to Booij (2010: 91), *this word is still linked to the subschema and reinforces its entrenchment.*

As Booij (2010) has stated, the extent to which individual language users rely on abstract schemas, or on analogy would need to be tested empirically. Yet, it seems that language speakers may differ quite considerably in the degree to which they develop abstract subschemas for creating complex words, as they differ in their lexical knowledge, that is the amount of lexical items stored in their memory, which is directly proportional to the ability to abstract a pattern through generalising across sets of words they have internalised so far.

3. Paradigmatic word formation

Analogy is also relied on in paradigmatic word formation which is deriving a word from another complex lexeme in a paradigm (Booij 2010). It means that a new word is formed not through a pure morpheme concatenation but by substituting one constituent (an affix, or a compound constituent) for another, with the simultaneous preservation of the idiosyncratic meaning of a complex lexeme. Bauer (2001) refers to this phenomenon as paradigm pressure.

For example, the English compound *dot bomb* 'a failed dot com' (see example nr 3) was derived from the compound *dot com* by replacing the constituent *com* with *bomb* rather than by combining *dot* and *bomb* into a compound, taking into account the fact the two compounds share the idiosyncratic meaning, as they both denote the Internet company. As Booij (2005: 13) proposes, *such a case of word-formation cannot be accounted for in a purely syntagmatic approach to morphology, neither a morpheme-based, nor a rule-based one. It is based on specific words, and therefore a typical case of analogy.*

Paradigmatic word-formation leads to the creation of paradigmatic relations between novel co-derivatives, which, in turn, may lead to the establishment of a new subchema (Booij 2005). Paradigmatic relations can be built around one specific model word, as is the case in the following examples:

(3)

model compound	paradigmatically formed compounds
dot com 'Internet company'	*dot bomb* 'failed Internet company'
	dot snot 'arrogant owner of Internet company'

dot bam 'high street retailer who offers online shopping'
dot org 'Internet address for a non-commercial site'

These compounds are instantiations of the following subschema:

(4)
$[[dot]_{Ni}[x]_{Nj}]_{Nk}$ ↔ $[SEM_j$ with relation R to the Internet based company$]_k$

Note that *dot* has undergone metonymic sense extension (Booij 2010), understood as the extension of the range of word meanings through the conceptual mechanism of metonymy. In this particular case *dot*, an element of the address of each Internet company, metonymically represents an online enterprise.

In other cases it is impossible to point to one specific word, serving as a model. Instead, there is a set of words that form a family, that is a paradigm with a given morpheme as a building block in the same position, referred to as a constituent family (Plag and Kunter 2009). Consider the following examples (Booij 2007: 37):

(5)
bliksem- 'lightning' *bliksem-bezoek* 'lit. lightning visit, fast and short visit'
bliksem-actie 'lit. lightning action, fast and short action'

4. Database

Innovative coinages constituting the core of analysis in this paper come from the following sources:

- Internet websites aimed at collecting new words in the media: http://www.americandialect.org, http://blog.oup.com/category/word-of-the-year-reference/feed/' http://www.wordspy.com/index.asp, http://www.worldwidewords.org/, http://www.wordspy.com
- Online newspapers: http://www.guardian.co.uk/ , http://www.thetimes.co.uk/tto/news/ , http://www.independent.co.uk/
- The collection of novel lexemes in the mass media compiled by Kerry Maxwell *From al desco to zorbing. New words for the 21st century* (2006). For details see the References section.

5. Analogical modelling

The analysis of the corpus of nonce-formations and neologisms has shown that analogical modelling manifests itself through: paronymic word-formation, analogical rule breaking and analogical extension resulting from morphological reinterpretation.

5.1. Paronymic word formation

As Hartmann and James (1998: 28) propose, paronymy refers to *any 'lookalike', or 'meanalike' confusable words.* According to Attardo (1994: 120) *two words are paronyms when their phonemic representations are similar but not identical.* Thus, paronymy can be understood as phonetic similarity, or phonetic analogy. The examples of paronymic word formation below are clear cases of *morphology making use of phonology* (Booij 2005: 177), as both analogically modelled blending and phonetic analogy apply phonemic

contrast which is responsible for the creation of a novel coinage (see the discussion below).

5.1.1. Analogically modelled blending

Analogically modelled blending should be understood as the type of blending which yields outputs paronymic to one of the source words (SW)[1]:

(6)

analogically modelled blend	source word 1 (SW1)	source word 2 (SW2)
a. *shuicide*	*shoe*	*suicide*
b. *intexticated*[2]	*text*	*intoxicated*

shuicide 'suicide committed in the terrorist attack by means of a shoe bomb'
intexticated 'distracted by texting while driving a car'

Out of 48 analogically modelled blends as many as 27 constitute a minimal pair with the source word 2, which means that they stand in the relation of contrastive distribution, or, in other words, commutation, understood as the paradigmatic relationship between two segments such that the replacement of one segment by the other generates a different lexeme (Crystal 2008):

(7)

	analogically modelled blend minimal pair member 1	source word 1	source word 2 minimal pair member 2
a.	*sexting*	*sexual*	*texting*
b.	*sofalise*	*sofa*	*socialise*

sexting 'sending sexual messages'
sofalise 'socialise with friends from your home through electronic devices

In the above examples, as well as in other blends discussed in this section (apart from four examples in (8) below), the splinter of source word 1 consists of one, or two segments and its use in the blend formation produces both phonemic and lexical contrast between an analogically modelled blend and the source word 2. Consequently, it can be stated that the morphological process of blending employs phonological contrast to create a novel blend. This novel blend is a play on the word well-established in the lexicon by being phonetically similar to it.

As far as the phonological structure of an analogically modelled blend is concerned, it usually follows the typical pattern in which the number of syllables in the blend tends to be the same as the number of syllables of the source word 2 if it is longer than the source word 1 (Plag 2003; Lehrer 2008). Consequently, there are just four blends which constitute a minimal pair with the source word 1:

(8)

analogically modelled blend minimal pair member 1	source word 1 minimal pair member 2	source word 2
a. *proto**h**ype*	*proto**t**ype*	*hype*
b. *bot**a**x*	*bot**o**x*	*tax*

[1] Source words should be understood as lexemes fused in the blend formation. Source word 1 provides the phonological string that appears first in the blend and source word 2 enters the blend second.
[2] Note that Kemmer (2003) refers to such blends as intercalative. See also 8 (c) and 8 (d).

c.	*sp**i**m*	*sp**a**m*	*instant (messaging)*
d.	*s**c**am*	*sp**a**m*	*confidence (trick)*

*proto**h**ype* 'device promoted by a website before it is ready to go'
*bot**a**x* 'tax imposed on plastic surgeries'
*sp**i**m* 'spam targeting users of Instant Messaging service'
*s**c**am* 'e-mail from a fraudster aimed at getting financial benefit by getting the addressee's confidence'

As regards *spim* and *scam*, see (8c) and (8d), they are interesting cases. First of all, these blends can be labelled intercalative (Kemmer 2003: 72), which means that *two words involved in the blend are so tightly integrated in the blended word that it the sounds of one source lexeme are interspersed between the sounds of the other.* Secondly, what has been incorporated into the source word 1 is the initial letter of the source word 2 which itself constitutes part of a noun phrase, which is not a typical example of a blend.

Note that 16 blends differ from source word 2 with respect to the onset of the word initial syllable. Consider some other examples belonging to this category (bold indicates the contrasting syllable onset):

(9)

analogically modelled blend	source word 1	source word 2
minimal pair member 1		minimal pair member 2
a.	***sh**owmance*	*show*	***r**omance*
b.	***c**arbage*	*car*	***g**arbage*
c.	***m**ockumentary*	*mock*	***d**ocumentary*
d.	***n**ouse*	*nose*	***m**ouse*

***sh**owmance* 'romance that actors engage in for the run of the show'
***c**arbage* 'distastefully modified car'
***m**ockumentary* 'programme in which actors pretend to be ordinary people'
***n**ouse* 'computer mouse controlled by the nose'

The remaining blends form minimal pairs, differing in the left-branch (Rogers 2000) of a branching syllable onset, see (10a), or the onset of the second, or third syllable, see (10b) and (10c). There are four examples of analogically modelled blends which differ from the source word 2 in a syllable peak, see (10d) and (10e), and just one example which differs with respect to the coda of its initial syllable, see (10f).

(10)

analogically modelled blend	source word 1	source word 2
minimal pair member 1		minimal pair member 2
a.	***f**reemium*	*free*	***p**remium*
b.	*so**f**alise*	*sofa*	*so**ci**alise*
c.	*proto**h**ype*	*proto**t**ype*	*hype*
d.	*bot**a**x*	*botox*	*t**a**x*
e.	*n**o**tworking*	*not*	*n**e**tworking*
f.	*we**d**site*	*wedding*	*we**b**site*

***f**reemium* 'programme which is free but which offers extra features for money'
*n**o**tworking* 'surfing social networking instead of working'
*we**d**site* 'website dedicated to an impending wedding'

Paronymic blending can be expressed by means of paradigmatically related schemas containing phonemes which stand in relation of contrastive distribution. Following the idea of hierarchical lexicon with more abstract schemas dominating specific subschemas (Booij 2010), the following schema dominating all the paronymic blends (with the exclusion of those form (8)) can be proposed:

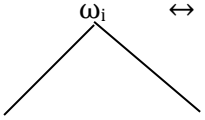
(11)
$[[S1]_i[S2]_j]_{N\ [PB]k}$ ↔ [ENTITY WHICH IS A HYPONYM OF SW2 AND RELATED TO SEM $_j]_{N\ [PB]k}$ (where S1 and S2 are splinters of SW1 and SW2, respectively). Note that PB stands for the paronymic blend.

Paronymic blends, coined by means of this schema, are a play on the source word 2, at the same time naming a subcategory of the concept that SW2 denotes, with features of this category specified and represented by S1.

As far as the specific types of paronymic blends discussed above are concerned, their formation can be represented by the following subschemas:

(12)

ω_i ↔ $N_{[PB]i}$ ↔ [ENTITY $_k$ CHARACTERISED BY THE PROPERTY$_j]_i$



$OSW1_j$ $S2_k$

(where OSW1 is the onset of the initial syllable of SW1)

An instantiation of this subschema are the blends from (9) above.

Note another schema which represents intercalative blends, see e.g. (10b), (10c), (10d), (10e), or (10f):

(13)

ω_i ↔ $N_{[PB]i}$ ↔ [ENTITY $_j$ CHARACTERISED BY THE PROPERTY$_k]_i$

$S2_j$ $S1_k$ $S2_j$

Other blends (21 in number) either differ in more than one phoneme from a source word 2, see (11a) and (11b), or they have been created through blending of the onset of the source word 1 with the source word 2, if the latter begins with the vowel, see (11c) and (11d):

(14)

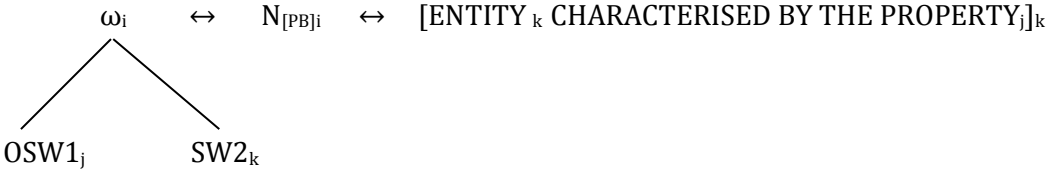
analogically modelled blend	source word 1	source word 2
a. *bragabond*	*brag*	*vagabond*
b. *bankster*	*bank*	*gangster*
c. *slacktivism*	*slack*	*activism*
d. *globesity*	*globe*	*obesity*

bragabond 'person who travels aimlessly and brags about it'
bankster 'banker whose aim is to swindle you out of money'
slacktivism 'doing projects that require the least amount of effort'

globesity 'worldwide epidemic of obesity'

Paronymic blends created through fusing the onset of the source word 1 with the source word 2, beginning with the vowel are an instantiation of the following subschema:

(15)



Note that all the blends analysed in this section are characterised by the presence of overlapping, which means that they contain a phonological string that is present in both source words (Kemmer 2003). This contributes to their intelligibility and goes in line with the results of experiments conducted by Gries (2004), who claims that the formation of blends is governed, among other things, by the tendency to preserve as much as possible from both source words. They also constitute a piece of evidence in favour of Kubozono's (1990) and Gries's (2004) observation that blends tend to preserve the syllable structure of source word 2. Consequently, analogically modelled blends discussed in this paper turn out to be paronymic to the source word 2 (apart from four cases discussed above). They constitute an example of *unexpected language* (Crystal 1998), as they can be regarded as the phonological and graphemic distortion of words well-established in the lexicon. For this reason, they evoke the effect of playfulness, which is tantamount to winning the reader's attention.

As regards the relevance of paronymic blending to the theory of Construction Morphology, this process can be presented by paradigmatically related schemas with contrasting phonological segments which perform a morphological function. Assuming that blends can be treated as abbreviated compounds (Lehrer 2003), a phonological segment of SW1 usually functions as a modifier, and the whole SW2, or the splinter of SW2 is both a semantic and syntactic head of the shortened compound, i.e. blend. Consequently, paronymic blends can be treated as morphological constructions which are instantiations of various subschemas presented in this section.

5.1.2. Phonetic analogy

The theory of Construction Morphology assigns considerable importance to the phenomenon of analogy in the formation of novel lexemes (see section 2). Evidently, analogy is seen as the word formation process operating on the morphemic level, while phonetic analogy to be discussed here operates on the phonemic (segmental) level. Yet the starting point is the same: it is the existing word which spurs the formation of a novel coinage that draws on a model lexeme.

The phenomenon of phonetic analogy should be understood as the formation of words which is motivated by paronymy, that is phonetic similarity to their bases, without semantic fusion taking place, which must occur in blending (Gries 2004). In the corpus I have collected there are just seven cases of nonce-formations created through phonetic analogy:

(16)

paronymic coinage	model lexeme
a. *naughties*	*nineties*
b. *freegan*	*vegan*

c.	*shopgrifting*	*shoplifting*
d.	*retrosexual*	*metrosexual*
e.	*Baracknophobia*	*arachnophobia*
f.	*wombanisation*	*womanisation*
g.	*hobby bobby*	*bobby*

Just one of these coinages, namely *retrosexual* 'classically male' (16d) constitutes a minimal pair with the model lexeme. An interesting example is the compound *hobby bobby* 'volunteer part-time policeman' (16g) which is very much different from any other paronymic items discussed in this section, as it contains both the model lexeme, i.e. *bobby* and the word that has been juxtaposed with it through analogy not analogically created, as all the lexemes analysed here, i.e. *hobby.* Thus, for this particular word analogical modelling does not consist in creating a paronymic novel lexeme, but in combining two already existing lexemes, which are phonetically similar.[3] As regards the semantics of the above paronymic coinages, they can belong to the same semantic field, e.g. *Baracknophobia* and *arachnophobia* are both a kind of phobia, *freegan* and *vegan* denote people with particular eating habits, *shopgrifting* 'buying and item, using it and then returning it for a full refund' and *shoplifting* stand for illegal activities that one may engage in while shopping. Yet another coinage, i.e. *naughties* 'the decade 2000-2009' is not semantically related to the model lexeme, it merely demonstrates the graphemic and phonetic similarity. One of these coinages is the synonym to the model lexeme, namely *wombanisation,* while yet another one is the antonym, i. e. *retrosexual.*

It is worth noting that there are also two examples of novel lexemes in the formation of which not the phonetic similarity to the model lexeme has been employed but the phonetic identity with it. At the same time, no semantic proximity can be observed as in some of the examples from (16); on the contrary, the resultant meaning turns out to go against the reader's expectations, which is attention-grabbing.

(17)

model lexeme	homonymic coinage
skier	*SKI-er*
pardon	*Par-Don*

SKI-er is an acronym, standing for 'spend the kids' inheritance' and it denotes a retired person who decides to spend his/her money instead of leaving it to his/her children. This acronym has been coined according to a prop-lexeme (Ungerer 1991) *skier* with which it is homonymous. As regards *Par-Don* 'person who splits his or her life between Paris and London', it has been modelled on *pardon* of Anglo-Norman (*pardoun*) origin (OED), so there is some historical justification behind it.

The attention-seeking function of the above lexemes is fulfilled by the contrast between the form and meaning of the model lexeme and paronymic coinage. The opposition between the established word and the new lexeme turns out to be surprising, as the lexical innovation of this kind stands the chance of being perceived as misrepresentation of another word, well-known to the reader, which is tantamount to excellent potential for *noteworthiness.*

5.1.3. Analogical extension resulting from morphological reinterpretation

[3] Compounds can be quite frequently motivated by ablaut, or the consonant which is similar, e.g. *hurly-burly*

Morphological reinterpretation, or, in other words folk etymology, is a common linguistic phenomenon, both of synchronic and diachronic importance. It stems from people's striving to make sense of opaque words and it results in the creation of a wide array of novel derivations by analogy, not only in the journalistic register but also, for example in the child language (see, e.g. Chmura-Klekotowa 1971; Slobin 1985; Tomasello 2003) and in everyday language, e.g. *shopaholic, fishburger, walkathon,* etc. It appears that in the journalistic register morphological reanalysis is aimed not so much at coming up with semantically transparent words but *at making the language do things it does not normally do* (Crystal 1998) through violating and transgressing morpheme boundaries. In effect, ingenious lexemes are formed that do make an impact on readers.

The journalistic register has bred numerous examples of novel schemas which emerged from reanalysing lexemes that are monomorphemic from a synchronic point of view, or were created through morpheme secretion brought about by blending. While the former category proves to be far less productive than the latter, sometimes a clear-cut boundary between the two cannot be established.

First of all, let us look into schemas established as a result of a morphological reinterpretation of monomorphemic lexemes:

(18)

morphological reinterpretation	subsequent coinages
a. *typo > typ-o*	b. *speak-o* 'oral mistake'
	c. *thumb-o* 'texting error'

As a result, two subschemas have been formed:

(19)

$[[x]_{Vj}\text{-o}]_{Ni} \leftrightarrow [\text{MISTAKE RELATED TO SEM}_j]_i$

$[[x]_{Nj}\text{-o}]_{Ni} \leftrightarrow [\text{MISTAKE RELATED TO SEM}_j]_i$

It means that the ending *-o* has acquired the status of a morpheme and the meaning 'a kind of an error' when attached to nouns or verbs. As stated by, e.g. Adams (2001: 58), the suffix *-o* is used in English to evoke *various attitudes according to the context*, as in *boyo, cheapo,* etc. Lopez-Rua (2007) notices that the last letter from the clipping *typo* can be used in the online jargon in words such as *scanno* 'error due to a scanning problem', or *mouso* 'error because of the malfunction of a mouse' but considers it merely the meaning extension of the low frequency suffix *-o.* It seems however, that *-o* as used in *boyo* and *typo* can be regarded as two different homonymic suffixes, as the former is used in the formation of diminutives to express affection, or contempt, which is not the case with the latter.

As regards the analogical extension through abstraction of a new subschema that emerged from blending, the constituent families from the corpus under discussion are not modelled on one particular blend. Instead, we can talk about the set of words that share a constituent, being here a splinter which emerged in the process of blending, in the same position. Even though it is theoretically possible to point to the earliest established blend in each constituent family, none of these is sufficiently well entrenched to be recognised as a model. As has already been remarked (see section 2), individual language users differ a lot as to their reliance on analogy-based account, or schema-based account, which are not mutually exclusive as they can coexist (Tuggy 2007). Consider the following examples:

(20)

source word 1	source word 2	blend
nay	*vacation*	*naycation*
stay	*vacation*	*staycation*
day	*vacation*	*daycation*
hay	*vacation*	*haycation*
man	*vacation*	*mancation*

naycation 'vacation spent without travelling and spending money on leisure activities'
haycation 'vacation on a farm'
mancation 'vacation in which the participants are all men'

For these words it seems indeed impossible to indicate the model blend, as although *staycation* seems to be the most popular, *daycation* was first attested over fifteen years earlier, that is in 1986. What appears fairly unquestionable, however, is the fact that the analogical creation of blends has led to morphological reinterpretation, abstraction of a new morpheme, being a splinter of a source word 2, and consequently the formation of new subschemas:

(21)
$[[x]_{Ni}\ [cation]_N]_{Nj} \leftrightarrow [\text{VACATION RELATED TO SEM}_i]_j$
$[[x]_{ADVi}\ [cation]_N]_{Nj} \leftrightarrow [\text{VACATION RELATED TO SEM}_i]_j$

Note some other constituent families created in a similar manner:

(22)

abstracted splinter	source word 2	constituent family
b. *–(a)logue*	*dialogue*	*halfalogue, trialogue, civilogue*
c. *–uppie*	*yuppie*	*scuppie, duppie, luppie, huppie*
d. *–tarian*	*vegetarian*	*locatarian, flexitarian, pescatarian*
e. *–vore*	*herbivore*	*locavore, proxivore, vegivore, informavore, opportunivore, lolcavore*
f. *–moir*	*memoir*	*me-moir, momoir, foodoir*
g. *–fiti*	*graffiti*	*shoeffiti, giraffiti, scratchiti*
h. *–flation*	*inflation*	*slowflation, stagflation, agflation*

a. *halfalogue* 'one side of a dialogue', *trialogue* 'conversation between three people', *civilogue* 'civil dialogue in which participants avoid insults';
b.*scuppie* 'socially conscious yuppie', *duppie* 'depressed yuppie', *luppie* 'Latino yuppie', *huppie* 'Hispanic yuppie';
c. *locatarian* 'person eating locally grown food', *flexitarian* 'vegetarian who sometimes eats meat or fish', *pescatarian* 'vegetarian eating fish';
d. *locavore, proxivore* the same as *locatarian, vegivore* 'person with special fondness for vegetables', *informavore* 'somebody who consumes information', *opportunivore* 'person consuming whatever he/she can find', *lolcavore* 'somebody spending a lot of time watching lolcats'
e. *me-moir* 'memoir that is exceptionally seld-centered', *momoir* 'memoir about motherhood', *foodoir* 'memoir focused on food and cooking'
f. *shoeffiti* 'shoes tied together by the laces and dangling from a powerline', *giraffiti* 'graffiti up high', *scratchiti* 'graffiti in which marks are etched into hard surfaces'

g. *slowflation/ stagflation* 'slow growth and high inflation', *agflation* 'inflation driven by rising prices of agricultural products'

as well as some schemas abstracted from them:

(23)
$[[x]_{DETi}\ [(a)logue]_N]_{Nj} \leftrightarrow [\text{DIALOGUE RELATED TO SEM}_i]_j$
$[[x]_{ADJi}\ [(a)logue]_N]_{Nj} \leftrightarrow [\text{DIALOGUE RELATED TO SEM}_i]_j$

(24)
$[[x]_{ADJi}\ [uppie]_N]_{Nj} \leftrightarrow [\text{YUPPIE RELATED TO SEM}_i]_j$
$[[x]_{ADJPi}\ [uppie]_N]_{Nj} \leftrightarrow [\text{YUPPIE RELATED TO SEM}_i]_j$,
etc.

As can be observed, it is the source word 2 that generates a splinter which is then subsequently used in analogical formations as the syntactic head. As for (22e), one of constituent family members, namely *locavore* has spurred some other paronymic coinages, such as *locapour* 'somebody who drinks locally produced wine or beer' and *lolcavore*. Thus, an analogically coined word can become a model lexeme itself.

Let me now point to one more instance of analogical modelling through morphological reinterpretation, which, however, does not end in the isolation of a new morpheme. Consider the following examples from my corpus:

(25)
a. *inherit-ance* > *in-heritance* → *pre-heritance* 'passing the capital to your children before you die'
b. *postpone* > *post-pone* → *pre-pone* 'arrange something at an earlier time'
c. *retire* > *re-tire* → *pro-tire* 'give up your present career to take up a hobby'
d. *obituary* > *o-bituary* → *pre-bituary* 'obituary prepared prior to person's death'
e. *revenge* > *re-venge* → *pre-venge* 'revenge taken in advance of the expected harm'
f. *recrimin-ation* > *re-crimination* → *precrimination* 'recrimination made in advance'
g. *rehab* > *re-hab* → *pre-hab* 'preemptive enrollment in a rehab facility'
h. *surviv-or* > *sur-vivor* → *pre-vivor* 'a person with some genetic mutation causing cancer'
i. *rebut-al* > *re-buttal* → *pre-buttal* 'preemptive rebuttal'

While in (18), (20) and (22) above morphological reinterpretation rests on the secretion of a new form, then used in analogical word formation, which subsequently develops into an abstract schema, the reanalysis from (25) does not end in resegmentation of a word by creating morpheme boundaries, as in (25b), (25c), (25d), (25e) and (25g), or shifting them, as in (25a), (25f) and (25h), but it goes one step further, which is the addition of the prefix *pre-* to the newly established 'base'. Whatever the kind of morphological reanalysis we are dealing with it is certainly a kind of linguistic manipulation aimed at showing off the writer's wit and imagination through rearranging morpheme boundaries, or creating them anew, which ends in creating novel morphemes, or establishing new patterns of morphological parsing. As the violation of morpheme boundaries does not take place on a regular basis in the language, it definitely enjoys the capacity for *foregrounding* lexical innovations.

5.1.4. Analogical rule bending and breaking

As noted in the introduction, Crystal (1998) regards breaking and bending the rules as the manifestation of the writer's creativity and ingenuity, which is tantamount to attracting the reader's attention. The analysis of rule breaking in the journalistic register has shown that a particular morphological rule can be violated on a one-off basis, as the creation of a multiply complex noun *togethering* 'spending time together will many of your relatives' by combining two morphological processes, i.e. conversion of an adverb to a verb: *together* (adv) → *together* (v) and suffixation with *–ing.* This process in known in the theory of Construction Morphology as schema unification defined as the simultaneous application of two or more morphological patterns (Booij 2010):

(26)
$[[[\text{together}]_{ADV}]_{V}\ \text{ing}]_{N}$

Apart from that, there are many instances of analogical rule breaking understood as multiple rule violation, that is a non-canonical use of a particular affix in the formation of a number of derivatives. This phenomenon can manifest itself through, e.g. attaching the agentive suffix *-er,* i.e. to nouns, N + N compounds, or even numbers:

(27)
a. *birther* 'a person questioning whether Obama was born in the USA'
b. *99er* 'somebody who is unemployed for a long time'
c. *truther* 'a person who believes that the US government allowed the 9/11 attacks'
d. *deather* 'a person who believes that the US health reform will lead to more deaths'
e. *tea-bagger* 'Tea Party movement participant'
f. *griefer* 'a person who intentionally harasses others online'
g. *domainer* 'a person who makes a living from domain name speculation'
h. *binner* 'a person who collects and sells used bottles and cans'

Although it is maintained by Plag (2003) and Adams (2001) that the agentive suffix *-er* is used not only in deverbal derivatives but also in denominal, e.g. *sealer* 'a person hunting seals', or even with numerals in some lexicalised derivatives, e.g. *fiver,* it is definitely the most productive with verbal bases. Plag (2003) even admits that *-er* is traditionally described as a deverbal suffix. For this reason, any coinages in which this prefix is attached to a numeral, a compound, or even a noun may and do strike as an instance of rule bending, as they run counter the prevalent linguistic trend.

Apart from the unconventional use of the suffix *–er*, the corpus also yields instances of rule breaking through the unusual use of two prefixes, that is *un-* and *de-.* As regards the prefix

un-, it is quite extensively used in online communication:

(28)
a. *unfriend* 'remove a person from a list of friend on a social networking site'
b. *unfollow* 'stop following a Twitter account that you were previously following'
c. *unlike* 'take your approval of something said online'
d. *unschooling* 'learning not at school but by experience'
e. *unbreed* 'a dog of uncertain pedigree'

As can be seen, *-un* attaches either to verbs, see (28a), (28b) and (28c), or gerunds, see (28d) and (28e). In English the prefix *un-* is used with verbs to denotes reversal. The verb

base needs to be transitive and resultative, which means no combinations with durative verbs, such as *play, walk,* or *sing* are allowed (Szymanek 1989). Additionally, *the condition for a verb to be formed depends on whether the physical possibility to undo the result of an action is conceivable. This explains why such verbs as unbeat, unhit, unkill, unpoison are not found* (Marchand 1969). In view of these conditions, the rules for the formation of reversative verbs have been broken for *unfollow* (28b) and *unlike* (28c). Although both of them are transitive neither of them can be considered resultative, as the idea of duration is inherent in their semantics. However, what is the most important is the fact that that the prefix *un-* as used with these verbs does not denote reversal, but indicates that a certain activity stopped. In contrast, the meaning of *unfriend* (28a) is typically reversative, as *friend* (v) 'put somebody onto the list of friends' (as used in the context of social networking services) has been created from *friend* (n) through conversion, while *unfriend* has got exactly the opposite meaning. As for *unschooling* (28d) and *unbreed* (28e), these are typical instances of rule breaking as the prefix *un-* may attach to nouns but only in the meaning 'lack of', as in *unbelief, unease, untruth* (Plag, 2003; Marchand, 1969), while with these two nouns the meaning of the prefix is 'not', which is typical for adjectives, as in *unavailable, unbroken,* etc.

Yet another instance of the attention-catching derivation can be represented by the non-canonical use of the prefix *de-*:

(29)

a.	*de-Scottishify*	'rebrand a product to remove its Scottish connotations'
b.	*depolicing*	'ignoring by the police petty crimes committed by minorities'
c.	*deshopping*	'buying something and returning it to get the money back'
d.	*deconflict*	'to avoid conflict while planning a military strategy'
e.	*dealert*	'to separate a nuclear weapon's warhead from its delivery system'
f.	*de-elect*	'remove an elected official from office'
g.	*deproliferate*	'reduce in number'

According to Szymanek (1998), Adams (2001) and Plag (2003), the prefix *de-* is used to form reversative and privative verbs. Marchand (1969) notes just a few examples of nouns formed by means of *de-* with the meaning 'loss, lack, removal of __ ', such as *deactivation, decontrol, decompression, deemphasis* and *demerger.* What is more, two semantic classes of privative verbs have been distinguished (Marchand 1973): verbs that can be paraphrased as 'deprive of ___' 'free from ___ ', e.g. *unmask, delouse,* etc. and verbs known as ablatives, which can be rephrased as 'remove from____', e.g. *unhook, deplane,* etc.

As for the above examples, the rules for forming reversative and privative verbs by means of the prefix *de-* have been broken, or at least bent. Sometimes they have been followed but by attaching this prefix to bases to which it attaches very rarely, they can be said to perform the function of attention-seeking devices. The verb *de-Scottishify* (29a) can be considered to be privative, as its meaning is 'to remove the signs of Scottish origin from the product'. As Szymanek (1996) maintains, the prefix *de-* only occasionally is used as part of a parasynthetic, prefixal-suffixal complex to form privative verbs, as in *decaffeinate, deacidify, deodorize*. Moreover, I have not found a single example of a privative verb created from the proper name. The rarity and oddity of this pattern is emphasised here graphically through hyphenation and subsequent capitalisation of *Scottishify* which has not been attested as a verb.

Depolicing (29b) and *deshopping* (29c) are two nouns and as it has been observed above, there are just a few cases of substantives in English formed with *de-*. Moreover, it can be postulated that the meaning of these two nouns is only partially privative since *depolicing* does not stand for 'the complete lack of policing' only for 'the lack of policing

with regard to petty crimes committed by certain social and ethnic groups in order to avoid accusations of racial profiling'. Along similar lines, *deshopping* cannot be interpreted as 'lack of shopping', as the element of buying is involved in the activity, being its starting point. Thus, *deshopping* is 'buying a thing and then returning it for a refund', because of that, from the point of view of the outcome of the activity *deshopping* could be understood as 'not shopping', so its meaning can also regarded as reversative.[4]

Deconflict (29d) and *dealert* (29e) are privative because their meaning can be paraphrased as 'free from _____', that is 'to free the craft or weapon from accidental collision by changing its flightpath' and 'free nuclear weapons from the state of alert', respectively. According to Szymanek (1996), the prefix *de-* derives privative verbs predominantly from monosyllabic nouns, e.g. *deflea, dehorn, dewool,* etc. Besides, in a vast majority of cases privative verbs are formed from concrete nouns not abstract ones. Both *conflict* and *alert* are abstract, besides they are disyllabic, so they are not 'typical candidates' for forming privative verbs, which makes them noteworthy.

De-elect (29f) and *deproliferate* (29g) are reversative, even though the possibility of undoing the action of electing is inconceivable, so here the semantic restriction on the reversative verb formation has been violated. As for *proliferate*, it belongs to the group of verbs which express the idea of reversal suppletively, i.e. *proliferate - plummet*. However, since *plummet* is an institutionalised word, it does not have the potential of attracting the reader's attention, the task that *deproliferate* performs quite easily.

5.2. Paradigmatic relations

As stated in section 3 above, analogy is also employed in paradigmatic word formation which is deriving a word from another complex lexeme in a paradigm (Booij 2010). As a result, paradigmatic relations between novel co-derivatives are formed, which, in turn, may lead to the creation of a new subchema. The analysed corpus has yielded the following instances of paradigmatic word formation: replacement of a compound constituent, replacement of an affix by a compound constituent, replacement of a compound constituent by an affix, or particle and paradigmatic acronimisation.

5.2.1. Replacement of a compound constituent

Replacement of a compound constituent should be understood as the formation of a novel compound not through concatenation of two independent lexemes but from the compound already existing in the language by replacing one of its constituents with another one (usually it is the modifier that gets replaced). This type of paradigmatic word formation is by far the most productive in my corpus. The majority of compounds created through paradigmatic word formation are modelled around one specific complex lexeme, and usually a new subschema is abstracted form them, as has been shown in section 3 above. What is essential in paradigmatic compound formation is the fact that novel compounds are characterised by the preservation of the idiosyncratic meaning. Apart from that, the extension of meaning of the head and modifier through conceptual mechanisms of metaphor and metonymy as well as the phenomenon of semantic concentration (to be explained below) is taking place. Note the following cases of paradigmatically created compounds:

(30)

model compound	paradigmatically formed compound
sitcom	*bitcom* 'short sitcom available on the Internet'
	zitcom 'sitcom featuring teenagers'
	slackcom 'sitcom featuring slackers'

[4] An interesting coinage is also *deshopper,* that is a person who engages in *deshopping.*

The above paradigmatically formed compounds exhibit semantic concentration (Meesters 2004), which means that the meaning of the whole compound *sitcom* is projected, or in other words 'concentrated' on one of its constituents, in this case a head, that is *com.* Consequently, *com* does not stand for any type of a comedy but for its particular type, namely situation comedy. As for *zitcom* and *bitcom*, they are phonetically motivated, differing from the model compound only in the initial consonant. It is interesting to remark that *zit* has been used here metonymically, as it stands for 'pimple' and here represents a teenager, being an example PART for WHOLE metonymy.

(31)

bookworm	*muckworm* 'miser'
	muskworm 'perfume dealer'
	ringworm 'person regularly attending boxing matches'
	red tapeworm 'person who adheres excessively to official rules'

This subchema has been motivated through the metaphorical extension of meaning of the lexeme *bookworm*. According to OED (online version), *bookworm* is 'a kind of maggot which destroys books by eating its way through the leaves' in a literal sense, while figuratively it denotes a person who is very fond of reading. Thus, the meaning of *bookworm* has been extended to apply not only to a worm's fondness of books but also that of a person. As regards paradigmatically formed compounds with *worm* as a head, its meaning has been reinterpreted as 'a person very fond of x', where x is the premodifier, forming the following subschema:

(32)

$[[X]_{Ni}[worm]_{Nj}]_{Nk}$ ↔ [PERSON VERY FOND OF $SEM_i]_k$ (where SEM_i can be used metonymically)

Consequently, *muckworm* is a person very fond of money (*muck* 'worldly wealth, money' (OED)), *muskworm*, where *musk* metonymically stands for any kind of perfume, denotes a person with a special liking of perfumes, in this case a perfume dealer, *ringworm* stands for the fan of the boxing matches, where *ring* is again the metonymic representation of boxing, etc. Thus, in the above compounds the head is metaphorical, while the modifier is metonymical with the exception of *muck.*

Consider some other paradigmatically formed compounds in which metaphorical extension has also taken place:

(33)

a. *couch potato*	*cot potato* 'baby watching a lot of TV'
	mouse potato 'person spending a lot of time over the Internet'
	baked potato 'drug user watching TV while intoxicated'
b. *blue chip*	*red chip*
	1. 'stock less reliable than blue chip but still a good investement'
	2. 'stock of a company based in a communist country'
	new chip 'stock from a young Chinese company'
c. *whitewash*	*greenwash* 'mislead the public by pretending to be environmentally responsible'
	blackwash 'to blacken the character of/caluminate'

As for the first model compound, that is *couch potato* (33a) it is the head which is metaphorical. According to OED, *potato* may have been coined punningly after *(boob-) tuber* = television addict and vegetable tuber, but perhaps simply by association with *vegetable* (slang orig. *U.S.*) 'a person who spends leisure time passively or idly sitting around, especially watching television, or videotapes'. *Cot potato* and *mouse potato* are two other examples of endocentric compounds with a metaphorical head, while in *baked potato* both the modifier and the head are metaphorical as *baked* stands for 'intoxicated' in the US slang. *Couch, mouse* and *cot* represent the case of metonymic sense extension as they constitute an example of THING for PERSON metonymy.

When it comes to the model compound in (33b), *blue chip* stands for 'stock thought to be safe and likely to make a profit' and in this case the whole compound has undergone a metaphorical transfer of meaning because in a literal sense it is 'blue counter used in poker, usually of high value'. Thus, the stock of a potentially high value is conceptualized as a blue counter which is valuable in poker, probably due to the fact that sometimes the situation at the Stock Exchange is as unpredictable as the outcome of a game of poker. In the subsequent coinages, that is *red chip* (2) and *new chip* the idiosyncratic meaning of *chip,* that is 'share' has been preserved, while the adjective *red* has been used metaphorically as it represents the communist system. As for *red chip* (1), the whole compound has again undergone a metaphorical transfer because in poker a red chip is fairly valuable although not as valuable as a blue one.

Lastly, *whitewash* (33c) which means 'free from blame', or 'try to hide unpleasant facts about somebody or something' is a metaphorical extension of the verb *whitewash* in its literal sense, i.e. 'cover the walls with the mixture of chalk, lime and water to make them white'. In this particular case hiding unpleasant facts about a person, or a thing is conceptualised as making them purer than they really are. Thus, *white* represents purity and innocence, *green* conceptualises being ecological, while *black* stands for guilt and wickedness. The idiosyncratic meaning of *wash* has been preserved in all these compounds and it can be paraphrased as 'falsely make somebody believe in something', namely 'falsely make somebody believe that people or things are better than they really are' for *blackwash,* 'falsely make somebody believe that a company leads an environmentally friendly policy' for *greenwash,* and 'falsely make somebody believe that people are worse than they really are'.

Consider yet another group of paradigmatically formed compounds:

(34)

software	*adware* 'software used in advertising
	censorware 'filtering software'
	firmware 'permanent form of software built into certain computers'
	groupware 'software aimed at facilitating collaborative working'
	cloudware 'software running on the Internet server'
	malware 'programmes written with the intent of being damaging'
	ransomware 'computer malware demanding ransom for the restoration of the system'
	spyware 'malware spying on computer users'

First of all, it should be said that *software* was modelled on *hardware,* after which also *peopleware* and *liveware* 'computer personnel' were created. It means that the phenomenon of semantic specialisation has taken place (Booij 2010), that is the word *ware* has developed more specialised meanings when embedded as the head of compounds.[5] In this particular case it denotes computer equipment, programmes, or personnel. Moreover, we can talk here about double semantic concentration as first the

[5] Compare the meaning of *ware* in *tableware, stoneware,* etc.

meaning of *software* has been projected on the head of the compound in the following coinages: *adware, censorware, firmware, groupware, cloudware* and *malware,* as all them denote various kinds of software. Next, the meaning of *malware* has been concentrated on the head of *spyware* and *ransomware* since both of them stand for types of *malware.*

Semantic concentration also takes place in the following subschema:

(35)

broadcasting	*mindcasting* 'posting messages reflecting one's thoughts'
	lifecasting 'broadcasting one's activities on the Internet for 24hs'
	silvercasting 'delivering programmes aimed at a small audience'
	vodcasting 'podcasting video content'
	godcasting 'podcasting religious messages'

All the above coinages stand for a kind of broadcasting, and in the last two examples a double semantic concentration can be observed as well because *vodcasting* and *godcasting* are instances of podcasting, a subtype of broadcasting. There is one coinage *egocasting*, the structure of which fits into the above subchema: N/ADJ + *casting*, however, semantically it is different as it means 'reading, watching or listening only to the media which reflect one's opinions', so *casting* does not stand here for a kind of broadcasting but it denotes the opposite process, that is reception of what is broadcast. It may be argued that the paradigm pressure seems to be so strong that the head of the compound has undergone semantic reinterpretation from 'broadcasting' to 'watching or listening'.

Yet another constituent family constitutes an example of semantic specialisation:

(36)

watershed	*walkshed* 'the area that a person can comfortably cover on foot'
	foodshed 'region that produces food used by a person, family, or town'
	fibreshed 'region producing fibres for the clothing used by a person, or a family'

Watershed (36) in North American usage is the 'drainage basin' or 'the catchment area', that is the area of land *that drains into a large water source, such as a river, lake, or ocean. Thus, it can be said that the noun shed has acquired a more specialized meaning when used the head of compounds, namely 'area covered by something, e.g. walk, food production, or fibre production.'*

Another model compound, *glass ceiling* meaning 'unofficial barrier to personal advancement in employment'(37) is an interesting case of a model compound as it has generated two constituent families: in the first set (37a) the head has been replaced, while in the other (32b) the head has been retained with the modifier being substituted:

(37)

a. *glass cliff* 'an important job (usually given to a woman) with a high risk of failure'
   glass wall 'social prejudice that prevents one from moving within departments of an organisation'
b. *grass ceiling* 'a set of barriers that prevent women from using golf to conduct business'
   silver ceiling 'prejudices preventing older employees from personal advancement'

Both (37a) and (37b) rely on the metaphorical and metonymical sense extension. *Glass* stands for something which is unseen yet breakable, while *ceiling* represents the barrier.

In paradigmatically formed compounds the metaphor is present in (37a) since *cliff* represents failure as you can easily fall down from the cliff and *wall* stands for a barrier preventing lateral movement in a company. Apart from that, (37b) displays instances of metonymy, as *grass* is used to denote golf since it is played on it, and *silver* embodies older people who often have got silver hair.

Apart from compounds created around one particular model lexeme there are also constituent families in which it is impossible to point to one particular word serving as a model. Consider the following examples:

(38)

insourcing	'obtaining goods or services in-house'
outsourcing	'obtaining something from the outer source'
friendsourcing	'gathering information from online friends'
crowdsourcing	'obtaining services from people outside the company, usually customers, or amateurs'
multisourcing	'relying on the services of external and internal firms'

According to OED, *insourcing* was first attested in 1979 and it was created in the process of prefixation: *in + sourcing*, while the first recorded use of *outsourcing* dates back to 1981 and it was derived by suffixation: *outsource + ing.* However, it be cannot be explicitly stated which of these two spurred subsequent coinages. The chronological criterion speaks in favour of *insourcing*, however, the criterion of frequency definitely opts for *outsourcing*, being a far more popular term, as the number of Google hits from August 16th 2011 amounts to 98, 500, 000 in comparison with 2, 900, 000 hits for *insourcing.* This constituent family has yielded the following subschemas:

(39)

$[[X]_{Prepi}[sourcing]_{Nj}]_{Nk} \leftrightarrow$ [OBTAINING SOMETHING IN THE WAY RELATED TO $SEM_i]_k$
$[[X]_{Ni}[sourcing]_{Nj}]_{Nk} \leftrightarrow$ [OBTAINING SOMETHING IN THE WAY RELATED TO $SEM_i]_k$
$[[X]_{Prei}[sourcing]_{Nj}]_{Nk} \leftrightarrow$ [OBTAINING SOMETHING IN THE WAY RELATED TO $SEM_i]_k$

The constituent family to be analysed below constitutes yet another example of semantic specialisation. The verb *mine* 'to dig holes in the ground in order to find and obtain coal, diamonds, etc' has developed a more specialised meaning, i.e. 'extract data, knowledge, or facts', suggesting some difficulty involved like in the process of getting out raw materials. For that reason, this particular case of semantic specialisation can again be regarded as metaphorical extension.

(40)

crowd mining	'extracting knowledge from large databases of social information'
audio mining	'extracting words from an audio file'
data mining	'discovering new patterns from large data sets'

Examples (41a) and (41b) are different from all the constituent families analysed so far because of the phenomenon of synonymy.

(41)

model compound	paradigmatically formed compounds
a. *freedom fries*	*freedom pat/ freedom grope/ freedom fondle/ freedom frisk* 'pat-down procedure at U.S. airports'

b. *top kill* *static kil /bottom kill* 'pumping mud or cement into the well to stop the flow of oil'

As for (41a), the model compound are well known *freedom fries*, a political euphemism for *French fries.*[6] Consequently, *freedom* underwent semantic reinterpretation (Booij 2010), which means that *freedom,* commonly associated with American values, acquired the meaning 'to be found/taking place in America'. In (41b) synonymy holds both for the model compound *top kill* and paradigmatically formed compounds: *static kill* and *bottom kill* as they all mean the same.

Lastly, the corpus under discussion yielded several constituent families, consisting merely of two elements, namely the model compound and just one paradigmatically formed compound:

(42)

model compound	**paradigmatically formed compound**
doppelgänger	*Googlegänger*
daughter track	*mommy track*
facelift	*bodylift*
Bluetooth	*bluejack*

Doppelgänger 'evil twin' gave rise to *Googlegänger* 'a person with the same name as you whose records and/or stories are mixed in with your own when you enter your name in the Google search engine that is while self-googling'. *Gänger* is a borrowing from German for 'walker', however here in the course of semantic reinterpretation it acquired a new meaning, namely 'your other identity'. *Daughter track* 'career path that allows a woman to work flexitime in order to take care of aging parents' spurred the coinage of *mommy track* 'career path that allows a woman to work flexitime in order to take care of aging parents', which means that the noun *track* underwent semantic specialisation (through metaphorical extension) from 'narrow road with uneven surface' to 'career path'. *Facelift* triggered the formation of *bodylift* with *lift* having acquired a new meaning 'medical operation aimed at making you look younger'. *Bluetooth* (an anglicised version of Danish *Blatand*), according to OED, is a byname of Harald I of Denmark credited with unifying the country during his reign, while *bluejack* means 'send anonymous messages to strangers around you using Bluetooth'. This is a case of semantic concentration as the meaning of *Bluetooth* has been projected onto the modifier, that is *blue.*

5.2.2. Replacement of an affix by a compound constituent

(43)

model compound	paradigmatically formed compounds
re-cycling	*up-cycling* 'improving the quality when recycling'
	down-cycling 'downgrading the quality when recycling'
	free-cycling 'giving away unwanted items for recycling'

The structure of the above compound nouns can be represented by the following subschemas:

(44)

$[[X]_{Pi}[cycling]_{Nj}]_{Nk} \leftrightarrow [SEM_j \text{ with relation R to } SEM_i]_k$

[6] This compound dates back to 2003 when it was coined as a result of the anti-French attitude over the France's opposition to the intervention of United Nations in Iraq.

$[[X]_{Adji}[cycling]_{Nj}]_{Nk} \leftrightarrow [SEM_j \text{ with relation R to } SEM_i]_k$

In the above constituent family semantic concentration is at work as well, as the meaning of *recycling* has been projected on all the compounds which denote the process of recycling while the preposition, or the adjective specify its type.

5.2.3. Replacement of a compound constituent by an affix or preposition

(45)

model compound	paradigmatically formed compounds
offshoring	*inshoring* 'bringing foreign workers to one's country'
	onshoring 'establishing companies in one's country'
	nearshoring 'moving jobs to a nearby country'
	homeshoring 'locating companies back at home'
	rightshoring 'restructuring the company so as to achieve balance between domestic and foreign jobs'

The model compound, that is *offshoring* stands for 'moving employment bases abroad'. It was presumably coined from the phrase *off shore*, however, *shoring* has acquired here a new meaning due to semantic reinterpretation, that is 'company's policy of development with regard to foreign or domestic employment and business'. Thus, the following subschemas have been created:

(46)

$[[X]_{Pi}[shoring]_{Nj}]_{Nk} \leftrightarrow [COMPANY'S\ POLICY_j \text{ with relation R to } SEM_i]_k$
$[[X]_{Ni}[shoring]_{Nj}]_{Nk} \leftrightarrow [COMPANY'S\ POLICY_j \text{ with relation R to } SEM_i]_k$
$[[X]_{Adji}[shoring]_{Nj}]_{Nk} \leftrightarrow [COMPANY'S\ POLICY_j \text{ with relation R to } SEM_i]_k$

The constituent family below (47) exhibits semantic concentration as all the paradigmatically formed compounds denote downloading the type of which has been specified by the premodifier, expressed by a preposition:

(47)

model compound	paradigmatically formed compounds
download	*preload* 'download software before it is required'
	upload 'transfer data from one computer to another'

Consider now an interesting case of the preposition replacement on a one-off basis:

(48)

model compound	paradigmatically formed compounds
bystander	*upstander* 'a person who decides to take an action'

In English there is no such a noun as *stander,* so *upstander* cannot have been coined from a phrasal verb *stand up* even though its meaning is inherent in it, as *upstander* is somebody who stands up for a person or principle.

5.2.4. Affix replacement

Affix replacement is yet another type of paradigmatic word-formation which is represented in my corpus just by one set of examples:

(49)

model derivative	paradigmatically formed derivatives
aforestation	*deforestation* 'destroying the forest'
	inforestation 'converting an area into a forest'
	reforestation 'reforesting an area'

The formation of *deforestation* from *aforestation* 'converting an area into a forest' has already been discussed by Bauer (2001) who is of the opinion that this is the case of paradigmatic word formation, maintaining that if it were created by suffixation, the choice of *-ation* instead of *-ing* might be not be the first one to make. Along similar lines, *inforestation* and *reforestation* are also instances of prefix replacement.

5.2.5. Paradigmatic acronimisation

(50)

model acronym	paradigmatically formed acronyms
NIMBY	*NUMBY* 'not under my back yard'
	GOOMBY 'get out of my back yard'
	YIMBY 'yes in my back yard'
	IMBY 'in my back yard'

As shown above, *NIMBY* 'not in my back yard' has spurred the formation of three other acronyms, denoting either objection to (*NUMBY, GOOMBY*), or acceptance of (*YIMBY*, *IMBY*) things happening in a person's neighbourhood. The idiosyncratic meaning of the model lexeme, that is 'the attitude towards things happening in one's neighbourhood' has been preserved in all the acronyms, so there is no doubt that we are dealing here with paradigmatic word formation.

6. Conclusions

The analysis of the corpus under discussion has shown that analogical modelling and paradigmatic word formation produce *unexpected language* either through rule bending and breaking, including phonetic and graphic distortion (analogical modelling), or by means of wordplay (paradigmatic word formation). Most of novel words formed in this way are nonce formations, created to make an impact on listeners, or *to capture the mood of the moment* (Crystal 1998: 30). As for analogical modelling, the most productive turned out to be paronymic blending and analogical modelling resulting from morphological reinterpretation. The outputs of the former seem to break the spelling rules, e.g. *murketing* 'murky marketing' instead of *marketing*, whereas the outputs of the latter represent an example of violating morpheme boundaries, e.g. *surviv-or* > *sur-vivor* → *pre-vivor*, often leading to the abstraction of a new morpheme, e.g. *-uppie* from *yuppie*. Apart from that, the cases of analogical rule breaking were attested, such as for example the use of the prefix *un-* + V not to denote reversal but to indicate that a certain activity stopped.

It seems that it may be proposed that paronymic blending can be treated as a subtype of paradigmatic word formation as it also consists in deriving a novel word from another word well-established in the lexicon. While in 'classical' paradigmatic word formation it is a morpheme that is substituted, in paronymic blending it is a phonological segment of the source word (usually SW2) that is replaced with a splinter (in a vast majority cases of SW1). A splinter can be regarded as a phonological segment that has a morphological function of its own (Booij 2005), as it is used to contribute to blend formation in which it acts as a modifier (see, e.g. (11), (12), or (15)). The blend created

in this way forms a paradigmatic relationship with one of the source words (usually SW2). This is yet another exemplification of a'tripartite parallel architecture' of the grammar, as proposed by Culicover and Jackendoff (2006), Jackendoff (2007) and also advocated by Booij (2005), according to which morphology deals with the relationships between three types of information: phonological, morphological and syntactic, assuming that word is a complex unit, functioning in these three dimensions.

With regard to paradigmatic word formation, as understood by Booij (2005, 2010), paradigmatically formed compounds constitute the most numerous category, presumably due to a high productivity of compounding in English. All the compounds that the present corpus has yielded are endocentric ones and it is usually the modifier that is replaced with the head preserving its idiosyncratic meaning, e.g. *lifecasting/mindcasting,* etc. from *broadcasting*. Some subschemas produced as a result of paradigmatic compounding display semantic concentration, e.g. *adware* modelled on *software,* or/and semantic reinterpretation, e.g. *mommy track* and *daughter track,* while others have been formed through sense extension by means of conceptual mechanisms of metaphor and metonymy, e.g. *ringworm*. Paradigmatically formed compounds are a play on the compounds well established in the lexicon (in this case model compounds) and as such they are successfully used to attract attention. Paradigmatic affixation and paradigmatic acronimisation turned out to be marginal cases.

On the whole, analogical modelling and paradigmatic word formation are quite pervasive in mass media because they can be successfully used as attention-seeking devices by generating creative coinages, whose formation is not always rule-governed and highly context dependent since they are analogically modelled on the already existing words. Moreover, ludic motivation is of paramount importance here as the reliance on the above mentioned processes creates ample scope for expressing humour and wit. On top of that, their use is pragmatically motivated as they perform the function of naming and are employed for the information condensation effect.

The author wishes to express her deepest gratitude to Professor Geert Booij and Professor Pavol Štekauer for their constructive feedback and suggestions.

References

Adams, V. (2001). *Complex words in English.* Harlow: Longman.
Attardo, S. (1994). *Linguistic theories of humour.* Berlin; New York: Walter de Gruyter.
Bauer, L. (2001). *Morphological Productivity.* Cambridge: Cambridge University Press.
Becker, T. (1990). *Analogie und morphologische Teorie.* München: Wilhelm Fink Verlag.
Booij, G. (2005). *The grammar of words.* Oxford: Oxford University Press.
Booij, G. (2007). Construction Morphology and the Lexicon. In Montermini, F., Boye, G. & Hathout, N.R. (Eds) *Selected Proceedings of the 5th Decembrettes: Morphology in Tolouse.* (pp. 33-44).
Booij, G. (2010). *Construction Morphology.* Oxford: Oxford University Press.
Chmura-Klekotowa, M. (1971). Neologizmy słowotwórcze w mowie dzieci. *Prace Filologiczne* 21, 99-235.
Croft, W. (2000). *Explaining language change.* Harlow: Longman.
Crystal, D. (1998). *Language Play.* London: Penguin Books.
Derwing, B. L. & Skousen, R. (1989). Morphology in the Mental Lexicon: A New Look at Analogy. *Morphology Yearbook* 2, 55-71.
Gries, S. (2004). Shouldn't it be *breakfunch?* A quantitative analysis of blend structure in English. *Linguistics* 42(3), 639-667.
Hohenhaus, P. (2007). How to do (even more) things with nonce words (other than naming). In
J. Munat (Ed.) *Lexical creativity. Texts and contexts* (pp. 15-39). Amsterdam/Philadelphia: John Benjamins.
Kemmer, S. (2003). Schemas and lexical blends. In T. Berg, et al. (Eds) *Motivation in language: from case grammar to cognitive linguistics* (pp. 69-97). Amsterdam: Benjamins.
Krott, A. (2001). *Analogy in morphology: the selection of linking elements in Dutch compounds.* Nijmegen: Max Planck Institut für Psycholinguistik.
Kubozono, H. (1990). Phonological constraints on blending in English as a case for phonology-morphology interface. *Yearbook of Morphology* 4, 1-21.
Lehrer, A. (2003). Understanding trendy neologisms. *Rivista di Linguistica* 15 (2), 369-382.
Lehrer, A. (2008). Blendalicious. In J. Munat (Ed.) *Lexical creativity, text and context* (pp. 115-133). Amsterdam: Benjamins.
Lieber, R. & Stekauer, P. (Eds) (2009). *The Oxford handbook of compounding.* Oxford: Oxford University Press.
Lipka, L. (1987). Word-formation and text in English and German. In B. Asbach –Schnittker & J. Roggenhofer (Eds) *Neuere Forschungen zur Wortbildung und Historiographie der Linguistik.* (pp. 59-67). Tübingen: Gunter Narr.
Lipka, L. (2000). English (and general) word-formation – the state of the art in 1999. In B. Reitz & S. Rieuwerts (Eds) *Anglistentag 1999 Mainz proceedings*. (pp. 5-20). Trier: Wissenschaftlicher Verlag.
Lopez-Rua, P. (2007). Keeping up with the times: lexical creativity in electronic communication. In J. Munat (Ed.) *Lexical creativity, texts and contexts* (pp. 137-163). Amsterdam/Philadelphia: John Benjamnins Company.
Marchand, H. (1969). *The categories and types of present-day English word formation.* Munich: C.H Beck.
Marchand, H. (1973). Reversative, privative and ablative verbs in English [in:] *Issues in Linguistics.* 636-643.
Maxwell, K. (2006). *From Al desko to Zorbing. New words for the 21st century.* London: Macmillan
Meesters, G. (2004). *Marginal morfologie in het Nederlands. Paradigmatische samenstellingen, neo-klassieke composite en splintercomposita.* Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.
Plag, I. (2003). *Word-formation in English.* Cambridge: Cambridge University Press.
Plag, I. (2006). The variability of compound stress in English: structural, semantic and analogical factors. *English Language and Linguistics* 10, 143-172.
I. Plag & G. Kunter (2009). Constituent family size and stress assignement in English. *Linguistische Berichte,* Sonderheft 17.
Slobin, D. (Ed.) (1985). *The crosslinguistic study of language acquisition.* Hillsdale, NJ: Erlbaum.
Szymanek, B. (1998). *Introduction to morphological analysis.* Warszawa: PWN.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of child language acquisition.* Harvard: Harvard University Press.
Tuggy, D. (2007). Schematicity. In D. Geeraerts and H. Cuyckens (Eds) *The Oxford handbook of cognitive linguistics* (pp. 82-116). Oxford: Oxford University Press.
Ungerer, F. (1991). Acronyms, Trade Names and Motivation. *AAA* 16.2. 131-158.

Dictionaries:

Crystal, D. (2008). *A dictionary of linguistics and phonetics.* Oxford: Blackwell Publishing.
Hartman, R. R. K. & G. James, (1998). *Dictionary of Lexicography.* Oxford: Routledge.
Oxford English Dictionary available at http:// www.oed.com/
Maxwell, K. (2006). *From al desco to zorbing*. London: Macmillan.

Online sources of the corpus data:

http://www.americandialect.org/index.php/amerdial/categories/C178// [consulted December 2010 and January 2011]
http://blog.oup.com/category/word-of-the-year-reference/feed/ [consulted November and December 2010]
http://www.wordspy.com/index.asp [consulted January 2011]
http://www.worldwidewords.org/ [consulted January 2011]
http://www.guardian.co.uk/ [consulted December 2010]
http://www.thetimes.co.uk/tto/news/ [consulted December 2010]
http://www.independent.co.uk/ [Consulted December 2010]

Rival suffixes:
synonymy, competition, and the emergence of productivity

Mark Lindsay
Stony Brook University
mark.lindsay@stonybrook.edu

1. Introduction

Complex systems can arise (and change) from the sum of numerous smaller interactions through natural selection. The term and idea are Darwin's but have broad application beyond biology. Three essential elements, when present in a continuous system, lead to an emergent process similar to natural selection:

(a) random perturbation of the system,
(b) a means of propagating some patterns, and
(c) the extinction of other patterns from the system.

All of these elements are present in derivational suffix systems. Speech errors, neologisms, diachronic phonological changes, and borrowings introduce random changes into the system. Productive derivation and exemplars ensure that successful patterns are more likely in the future. Finally, an intolerance of synonymy means that a suffix that does not continually attach to new words will ultimately cease to be productive and will therefore be eliminated from the system. (For an extended discussion of this theory, see Lindsay and Aronoff 2012.)

Although one affix will tend to dominate a broad domain, a language can settle into a stable system that includes the less competitive affixes as productive elements. This is achieved if the less-productive affix happens to find a niche: a clearly defined subdomain within its potential domain — a subsystem that is therefore distinct and predictable to a speaker in spite of a general trend towards another affix. Furthermore, in order for an affix to remain productive, this subdomain must also be a large enough subset of all eligible words that speakers can generalize its usage and that the affix will have an ongoing inflow of new words to combine with. A subdomain whose members share some linguistically salient characteristic should have the highest likelihood of establishing itself in the grammar of speakers; this characteristic (or combination of characteristics) is what is used to establish a pattern for forming new words. These characteristics could, then, be phonological, morphological, semantic, pragmatic, and so on.

The investigations in this paper build upon work from Lindsay and Aronoff (2012). I will investigate evidence for emergence in suffix systems: when new potential suffixes enter the system, the system ultimately organizes these elements; as the language evolves, the organization of elements adapts to these changes. These emergent processes in the English suffix system are a part of glossogenetic evolution (Hurford 1990, also discussed in Steels 1997 and Fitch 2010, among others), a concept that is distinct from phylogenetic evolution, i.e. the evolution of language ability in humans. In particular, I will examine rival suffix pairs *-ic*/*-ical* and *-ity*/*-ness* in English, showing that each suffix thrives because of adaptation to the system. I will then look at the pair of *-ize*/*-ify* and their equivalents in three Romance languages: French (the primary source of the English suffix), Spanish, and Portuguese. We will see how a coincidence in Latin ultimately led to the organizational system we see in all of these languages today.

2. Rival English suffixes *-ic* and *-ical*: comparison with a traditional corpus

The suffixes *-ic* and *-ical* are borrowed suffixes originating in Greek and Latin. While *-ic* directly corresponds to Greek *-ikós*, *-ical* is an etymologically redundant affix created from a combination of factors: Greek *-ikós*, Old French *-al*, and the many scientific nouns ending in *-ic* or *-ics* borrowed from Latin (Marchand 1969:236, 241). Both *-ic* and *-ical* are healthy, productive suffixes in the present day, although they are in direct competition. Lindsay and Aronoff (2012) found that these suffixes could both succeed because each suffix had developed a productive niche.

To evaluate each of these competing suffixes, productivity was measured in a novel way, by incorporating Google search results[1]: the exact literal string for words is queried, and the Estimated Total Hits (ETM) result from Google is recorded for each word; one can then look for numerical patterns in these numbers to learn about productivity.[2] By comparing ETM values for each form for a given stem (e.g. *biolog-ic* and *biolog-ical*), the more productive suffix will tend, over a large number of comparisons, to have a higher ETM value more often than the less productive suffix.

When the ETM values for each *-ic*/*-ical* pair were compared, *-ic* was found to be the "winner" in 10613 out of 11966 pairs.

	Total Stems	**Ratio**	***olog* Stems**	**Ratio**
Favoring *-ic*	**10613**	**7.84**	74	1
Favoring *-ical*	1353	1	**401**	**6.42**
Total	11966		475	

Table 1: *-ical* is productive in stems ending in *olog* (from Lindsay and Aronoff 2012)

While *-ic* was preferred overall, it was not preferred in all domains. After systematically examining all possible neighborhoods, it was revealed that stems ending in *olog* overwhelmingly favored *-ical* over *-ic*. This represents a morphologically bound niche for *-ical*.

Although the results from Google deal with data that is untagged and often noisy, we find similar results in a more traditional corpus, such as the Corpus of Contemporary American English (COCA), a balanced corpus containing 425 million words of text that are tagged for part of speech. In this follow-up study, a similar comparison of *-ic*/*-ical* rivals was conducted using word frequencies from among the 60,000 most frequent words in COCA.

[1] Other measures of morphological productivity exist, such as those discussed in Baayen (1993), Plag (1999) and Bauer (2001). The approach used in this article is not meant to replace currently existing methods; rather, it is an additional means of measuring productivity that exploits the vast amount of linguistic information contained within the World Wide Web.

[2] One must be cautious when incorporating Google ETM values into a measurement of usage. While Google is a vast and freely-available resource, it is also "noisy"; that is, individual results contain false positives due to typos, non-native speech, spam, the lack of part-of-speech tagging, and so on. Furthermore, ETM results represent the number of pages a string is estimated to appear in, not the number of occurrences. (Other discussion of such considerations can be found in Hathout and Tanguy 2002, among others.) For these reasons, it is important that little weight is placed upon the actual raw numbers themselves (only relative differences should be considered) or upon any individual word pairs. For the time being, it is also important to restrict investigations to single words, rather than phrases, due to the algorithm by which Google estimates phrasal results. A broad investigation of suffixes mitigates many of these concerns, as we are dealing with single words, regular inflection patterns, and a large number of stems.

All adjectives ending in *-ic* and *-ical* were queried in COCA; there were 1465 *-ic*/*-ical* pairs in all — a much smaller sample than the 11966 queried through Google. The number of tokens for each form was compared, and the form with a greater number of tokens was the "winner" for that pair. In total, 1197 pairs favored *-ic* over *-ical* (a ratio of 4.5 to 1). As in the previous study, in the *olog* subset, *-ical* was favored over *-ic* in 73 of 86 pairs (a ratio of 5.6 to 1).

	Total Stems	**Ratio**	***olog* Stems**	**Ratio**
Favoring *-ic*	**1197**	**4.5**	13	1
Favoring *-ical*	268	1	**73**	**5.6**
Total	1465		86	

Table 2: COCA results for *-ic* and *-ical*

As COCA is much smaller than what is indexed by Google, often only one form in each pair was present in the corpus; only 6% of pairs had both *-ic* and *-ical* forms with non-zero frequencies in COCA. Nonetheless, the results from COCA are very similar to those from Google, and lead to the same conclusions about the productivity of *-ic* and *-ical*. That is, we see clear evidence of a morphological niche in *olog* stems where *-ical* prospers.

3. Morphologically constrained *-ity* vs. *-ness*

In contrast to *-ic* and *-ical*, the origins of the suffixes *-ity* and *-ness* differ greatly from one another. Native suffix *-ness* had existed well before any *-ity* words had entered into English and was productive with both native and borrowed words; indeed, derivation from French adjectives was common by 1300 (Marchand 1969:335). The earliest *-ity* words were whole-word loans from French starting around the 14th century. As *-ity* words began entering the language, the edges of the domain of *-ness* were gradually eroded as *-ity* established a niche in which it could be productive. The establishment of a productive *-ity* follows a typical pattern for borrowed suffixes, as illustrated in Figure 1. First, whole words were borrowed that happened to end in *-ity*; over time, as these *-ity* borrowings accumulated, derived forms containing *-ity* began to emerge with increasing frequency. After productivity reached a critical mass, it continued to increase in spite of a decrease in whole-word borrowings.

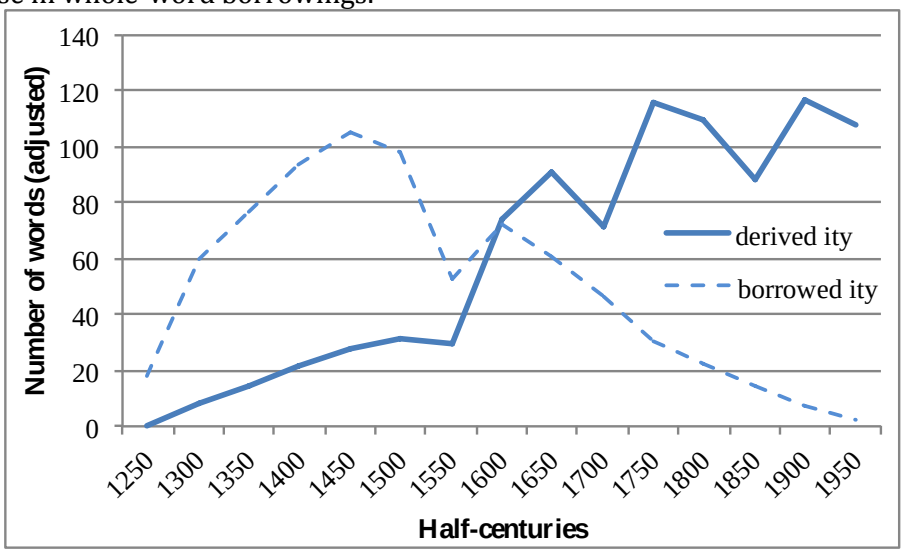


Figure 1: Whole-word borrowings of *-ity* precede its gradual development of productivity (data from first citations in the OED).

The synchronic productivity of these suffixes was compared using the methodology discussed in Section 2, which leverages Google indexing. In this case, the Estimated Total Matches were queried for 3256 rival pairs. Overall, *-ness* was found to be preferred in 2381 pairs, versus 875 for *-ity*, for a ratio of 2.72 in favor of *-ness*.
While *-ness* is more generally successful, there are a number of subdomains in which *-ity* dominates. In Table 3 below, we see the dominant subdomains for *-ity* (with at least 50 members):

	-ity	***-ness***	***ratio***
able	**1025**	379	2.7
al	**500**	183	2.7
- ial	***113***	*18*	*6.3*
- ual	***55***	*6*	*9.2*
- ral	***50***	*14*	*3.6*
ic	**181**	37	4.9
ar	**111**	18	6.2

Table 3: *-ity*/*-ness* neighborhoods favoring *-ity*

In contrast to *-ical* (which dominated only stems ending in *olog*), there are multiple large domains in which *-ity* is highly successful, each corresponding to a Latinate suffix: *-able*, *-al*, *-ic*, or *-ar*. However, its dominance is far from total: there are hundreds of forms among these sets that prefer *-ness*, a reflection of *-ness*'s overall productivity in English.

One might initially be inclined to attribute *-ity*'s preference for exclusively Latinate suffixes to some type of feature or grammatical constraint in English (e.g. Aronoff 1976, Booij 1977, Plag 1999, and others). But if we consider how *-ity* entered into English and how its place in the language evolved, such an explicit stipulation is not necessary. Because *-ity* came into English via whole-word borrowings from French and Latin, it is natural that a large number of cases would exist in which *-ity* is part of a word that already has a Latinate suffix embedded in it. In fact, it is likely because of these co-occurrences that *-ity* was able to achieve the level of productivity that it enjoys today. With native suffix *-ness* already highly productive in English, *-ity* needed to occupy a niche in order to thrive. Being part of an emergent system, *-ity* capitalized on what it was given: dominance among words containing Latinate suffixes like *-able*, *-al*, *-ic*, and *-ar*. Its productivity does not extend to any Germanic suffixes because *-ness* is already dominant there (as well as being generally dominant), thus leaving no opportunity for *-ity* to "get its foot in the door", as it were. In addition, the fact that *-ity* is also coupled with a shift in stress created an additional phonological hurdle to creeping into Germanic domains.

Nonetheless, when a suffix that hosts *-ity* has itself become productive with all bases, as in the case of *-able*, it does precisely that (Marchand 1969:313):

> With **-able/-ability** […] the synchronic relevancy of the pattern has gone far beyond its original morphological basis. While other adjectives derive substantives in **-ity** only when the adjective is Latin coined […], the derivative range of **-able/-ability** […] today comprises practically any adjective in **-able**, including adjectives derived from native roots (*lovable*/*lovability*).

Those neighborhoods where *-ness* is preferred tend to have a very different character from those that prefer *-ity*. The following are particularly noteworthy neighborhoods where *-ness* is preferred:

	-ity	*-ness*	*ratio*
ed	2[3]	**804**	
ing	0	**347**	
ess	4[3]	**281**	
ish	0	**251**	
ful	0	**250**	
ent	6[3]	**84**	
ant	3[3]	**56**	
ous/os	187	**646**	3.5
ive	128	**333**	2.6

Table 4: *-ity*/*-ness* neighborhoods favoring *-ness*

In the first seven neighborhoods above, the domination of *-ness* is total; there are effectively no exceptions (see description of false positives in footnote 3). These neighborhoods also largely correspond to Germanic suffixes: *-ed*, *-ing*, *-ess*, *-ish*, and *-ful*. The exceptions to that are *-ent* and *-ant*, which only seem to prefer *-ness*; most *-ent* and *-ant* words prefer neither suffix, but rather take *-ence/-ency* and *-ance/-ancy*, respectively. As there would not be any French borrowings ending in *-entity* or *-antity*, there would be no analogical basis for deriving such forms.

The two neighborhoods in which there is some amount of contention are *ous/os* and *ive*. In both cases, these are Latinate suffixes, coming from Old French and French/Latin, respectively (Marchand 1969:315, 339), so we might expect *-ity* to dominate. Instead, it is *-ness* that is dominant. This further underscores how a natural system evolves: just because *-ity* was able to take advantage of exemplars favoring *-ability,* and other Latinate suffixes, this does not guarantee that it will have the same level of success in all such cases. The fact that *-ity* does provide healthy competition in both *ous*/*os* and *ive* neighborhoods (in contrast to the Germanic neighborhoods) shows that its Latinate origins did have some impact on the productivity of the rivals in those domains; that is, there were some whole-word borrowings ending in *ivity* (e.g. *festivity*, *captivity*, *activity*) and *osity* (e.g. *curiosity*, *virtuosity*, *generosity*) trickling into English, providing some basis for extending the pattern, but not enough to dominate these niches.

4. Phonologically defined niche: *-ize* versus *-ify*

The English suffix *-ize* originated as Greek *-izō*, entering into English through Late and Ecclesiastical Latin *-izare* and French *-iser*. Its rival, *-ify*, comes ultimately from Latin *-ificare*, though many words came into English via French *-ifier* (Marchand 1969). Lindsay and Aronoff (2012) examined *-ize* and *-ify* in present-day English. They queried 2636 *-ize*/*-ify* suffix pairs and found that 2217 favored *-ize* and 419 favored *-ify*: a 5.0 ratio. Therefore, it would seem that *-ize* has been more productive overall. However, like *-ical*, *-ify* is productive in a certain subdomain of stems. As we see in Figure 2, the productivity of each affix correlates strongly with the number of syllables in the stem.

[3] Forms favoring *-ity* in these neighborhoods were as follows:

ed: *rubedity*, *heredity*

ess: *necessity*, *nonnecessity*, *unnecessity*, and *supernecessity*

ent: *entity*, *nonentity*, *identity*, *nonidentity*, *coidentity*, *stringentity* (which owes its high frequency to an object in the Java programming language called *StringEntity*)

ant: *quantity*, *overquantity*, *disquantity*

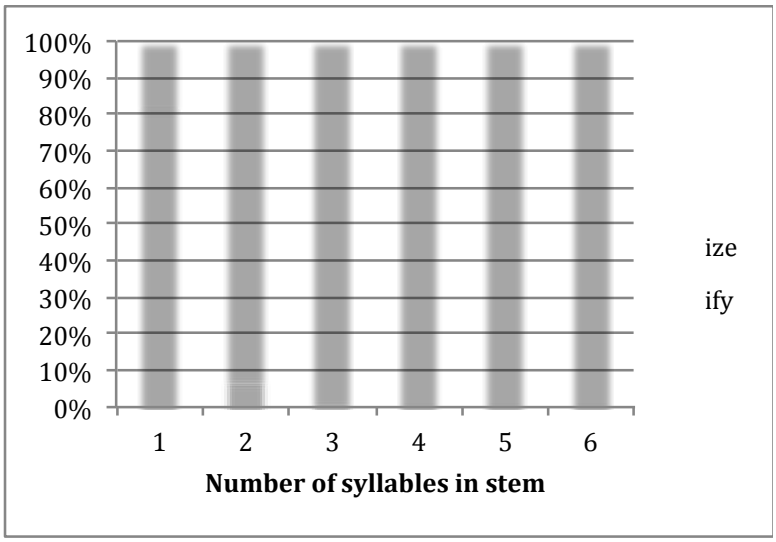


Figure 2: Strong preference for *-ify* in monosyllabic stems, sharply dropping off as stem length increases (from Lindsay and Aronoff 2012, Figure 11).

Monosyllabic stems favored *-ify* over *-ize* in 82% of comparisons, dropping sharply to just 4% for disyllabic stems. Thus, we see a strong tendency but not a strict rule: the shorter the stem, the more likely *-ify* is preferred — but neither suffix is totally restricted.

Both *-ize* and *-ify* are borrowed suffixes, which emerged out of little more than whole-word Romance borrowings that were continually entering into the language at the time. As we see in Figure 3 below, both *-ize* and *-ify* were first borrowed beginning in the late 13th century. The borrowing continued and, over time, this growing collection of borrowings began to develop organization. Both emerged as productive suffixes at the same time as well; namely, in the 16th century. From the beginning, *-ize* was the more productive of the two.

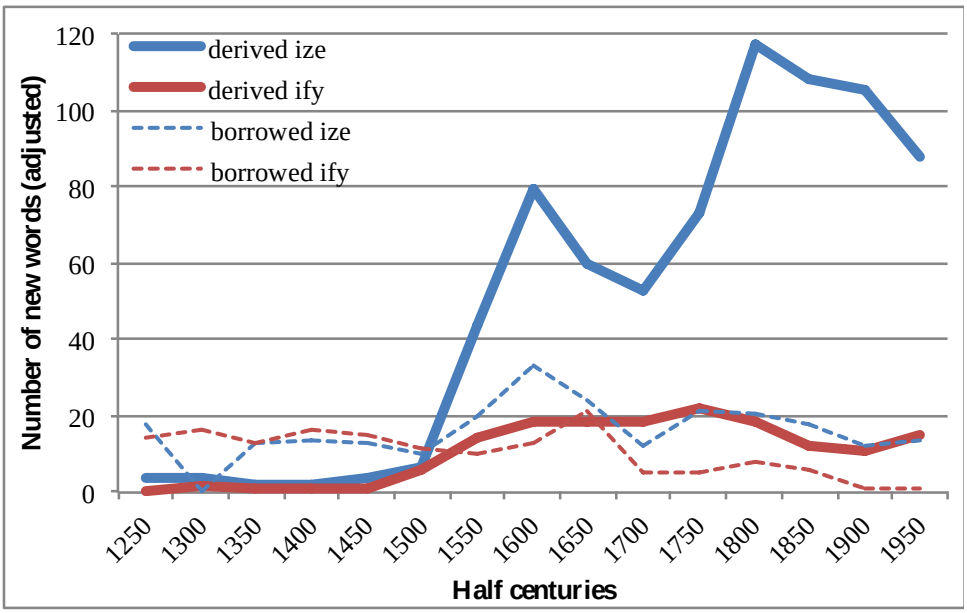


Figure 3: Derived and borrowed forms of -ize and -ify

An interesting question arises about the relative productivity of these rival suffixes in the Romance languages themselves: are the productive domains in these languages similar to English? In this investigation, we will look at the productive domains of *-ize*/*-ify* suffix

rivals in French, Spanish, and Portuguese[4], using the same method of investigation that was used for English.

A total of 365 pairs were queried in French. As with English, the full inflectional paradigm was taken into account, and the query results for all inflectional forms for a given word were summed together (e.g. 365 pairs, 37 inflections, leading to 27010 Google queries). Fortunately, the morphological paradigm for these suffixes is regular, thus making the process of generating forms much less complicated.

The overall results for French showed a similarity to those for English: *-iser* was heavily favored over *-ifier*, where it was the winner of 292 pairs: a 4.0 ratio.
In Table 5, we can see a breakdown of the number of winners grouped by syllables in the stem:

	1	**2**	**3**	**4**	**5**	**6**	**Total**
-ifier	50	22	1	0	0	0	*73*
-iser	25	144	85	32	3	3	*292*
	75	*166*	*86*	*32*	*3*	*3*	*365*

Table 5: count of types preferring each suffix (1-6 syllables in stem) for French

Like English, there is a tendency for *-ifier* to have a monosyllabic stem, as 68% of all *-ifier* winners have monosyllabic stems, while *-iser* is preferred in polysyllabic stems, where 83% of all *-iser* words occur. Indeed, we see a similar pattern to English in Figure 4:

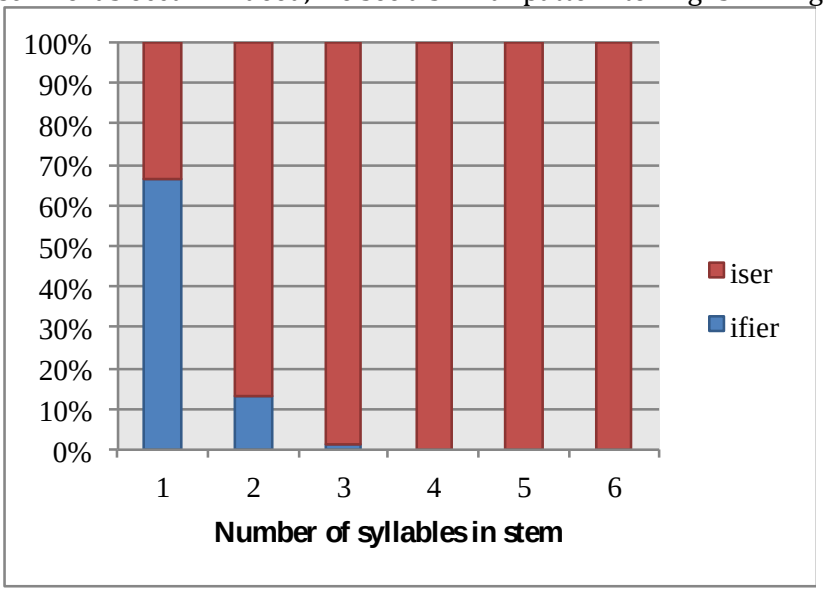


Figure 4: *-ifier* maps to monosyllabic stems, but more weakly than English

Like English, we see an attraction towards shorter stems in *-ifier*; however, *-ifier* is the winner only about two-thirds of the time with monosyllables. That is, the pattern is somewhat weaker in French than in English, but clearly present.
Next, looking at Spanish, we see that *-ificar* is clearly strongest among monosyllabic stems, with just over 50% preferring *-ificar*, while *-izar* peaks at disyllabic stems.

[4] Dictionary lists of query words were compiled largely from free dictionaries, particularly those found at *WinEdt.org*. An initial investigation into Italian was conducted as well, but was not included due to a small sample size used (100 words). However, the results did match with the other Romance languages investigated here.

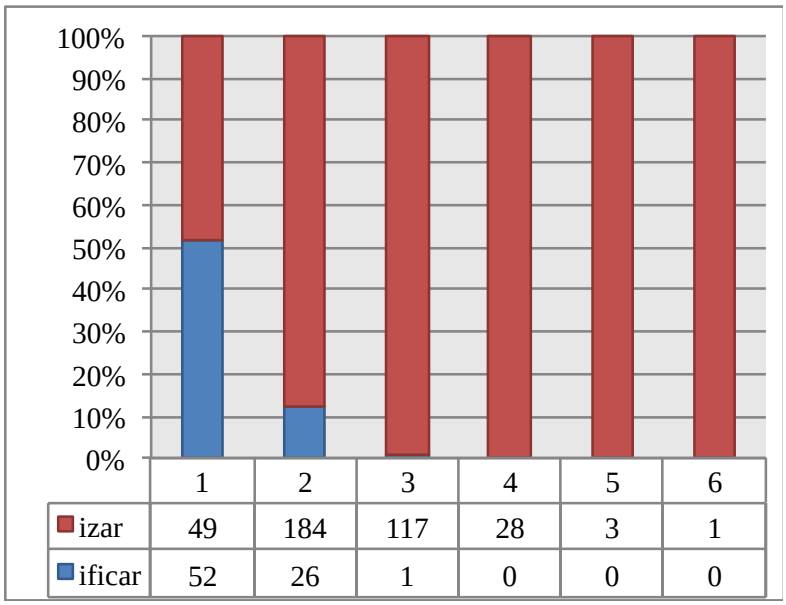


Figure 5: Spanish *-ificar* also tends toward monosyllabic stems

Portuguese also follows the pattern of its sister languages, with a monosyllabic *-ificar* tendency.

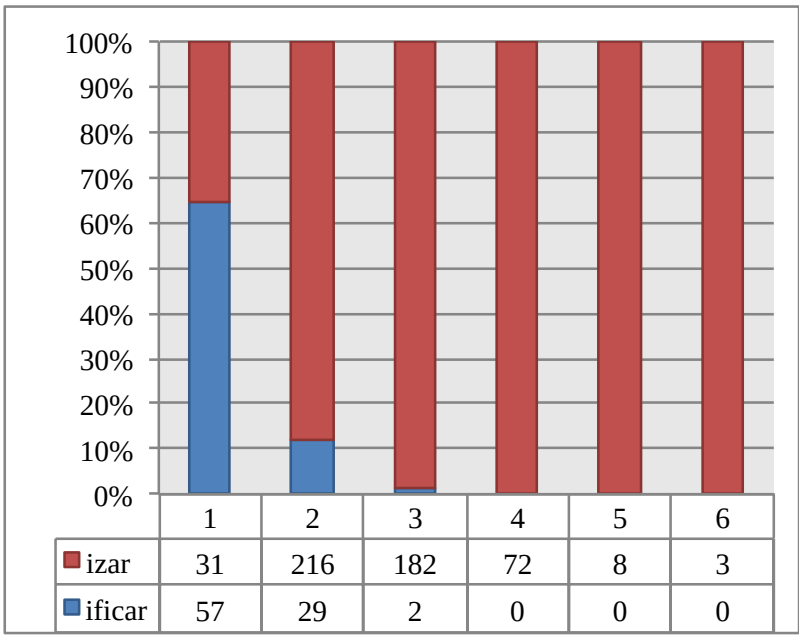


Figure 6: Portuguese shows an even stronger *-ificar* monosyllable correlation

We see a striking similarity in the relative distributions of *-ize* and *-ify* forms across all four languages. In Figure 7 below, we see the distributions of *-ize* for all four languages.

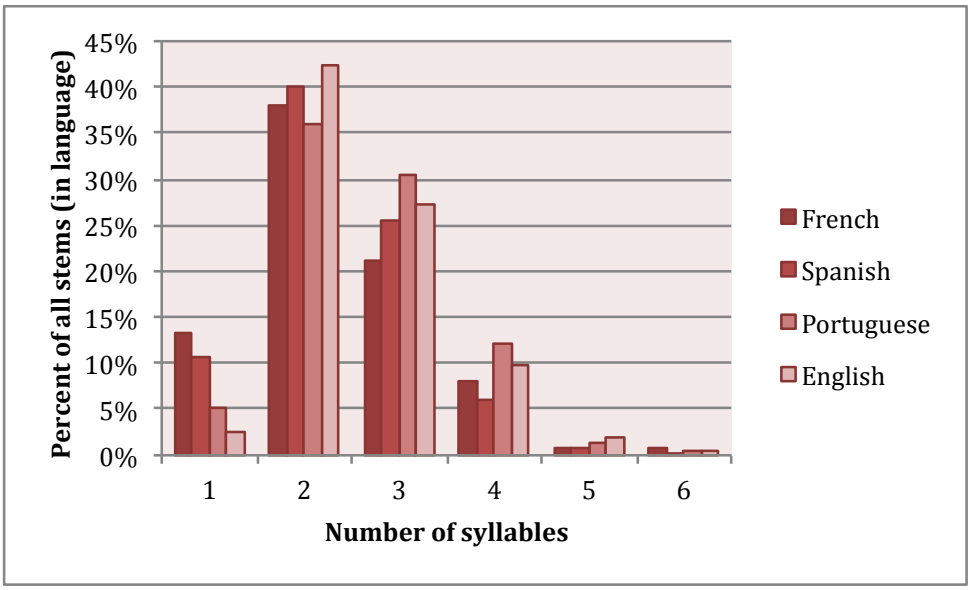


Figure 7: Relative distributions of *-ize* across all languages

In all cases, *-ize* peaks at two syllables and slowly tapers off, whereas *-ify* peaks at one syllable and drops off more rapidly, as we see in Figure 8.

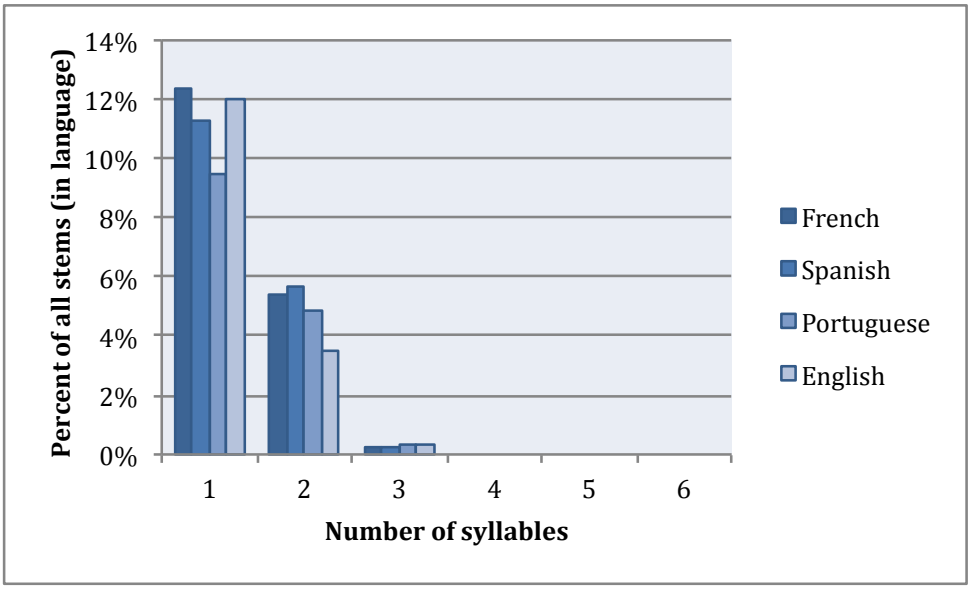


Figure 8: Relative distributions of *-ify* across all languages

Despite other phonological differences among these languages, this relationship between syllables and these suffixes is universal in the languages examined. In the case of French, Spanish, and Portuguese, this relationship was inherited directly from Latin; in the case of English, the relationship was actually recreated due to the exemplars provided by the whole-word borrowings.

We might be able to explain the source of this split, then, by looking at Latin, where both of these suffixes began[5]. Both *-izare* and *-ificare* were latecomers to Latin, with *-izare* being part of whole-word borrowings from Greek and *-ificare* being a grammaticalization of the Latin verb *facere* ('to make/do'). The latter formed a few verbs

[5] Thanks go to Mark Aronoff for suggesting this route of investigation.

in Classical Latin, more in Late Latin, and grew exceedingly frequent in Medieval Latin (Marchand 1969:300). Words containing *-izare* finally entered Latin from Greek around the 3rd or 4th century (Marchard 1969:318) with the conversion of the Roman Empire to Christianity. As such, most Latin *-izare* words have Greek or Semitic stems (91.9%) and are largely ecclesiastical in nature (e.g. *euangelizare*, *hymnizare*, *prophetizare*). In contrast, *-ificare* words almost exclusively contain Latin stems (98.8%). This divide between Latin and Greek/other turns out to be crucial for the patterning of these suffixes in all of the modern languages that have been discussed in this section. Because Latin naturally tends toward monosyllabic stems, while Greek stems tend to be polysyllabic, the number of syllables in the stems averages to 2.03 for the *-izare* words and 1.04 for *-ificare*. These results are summarized below:

	-izare	*-ificare*
Words	35	82
Latin stems	3	81
Non-Latin stems	32	1
Percent Latin stems	8.1%	98.8%
Average syllables in stem	2.03	1.04

Table 6: Summary of Latin -izare and -ificare words

We know that *-ize* and *-ify* came into English through whole-word borrowings from French. The *-iser*/*-ifier* data show that French was not only the source of two new productive English suffixes; in addition, English speakers' grammars were also influenced by the way that the forms were distributed in French. This led to a recreation in English of the productive niches for each of these suffixes (based on number of syllables) in addition to the suffixes themselves.

Furthermore, if the distribution of *-iser* and *-ifier* in present day French is similar to the distribution several centuries ago (i.e. representing a mono/bisyllabic split that is somewhat weaker than English), this would suggest that English took this pattern from French and "ran with it", not just respecting the pattern but strengthening it.

Thus, the patterning of *-izare* and *-ificare*, the equivalents in the Romance languages, and ultimately English, was the result of a coincidence: two rival verbalizing suffix patterns happened to enter into Latin around the same time; Latin natively preferred monosyllabic stems while the Greek words had longer stems. This superficial correlation between stem length (or overall word length) and suffix was recognized by Latin speakers and was salient enough to propagate into (and sustain itself in) all of the other Romance languages we have examined. Finally, and most interestingly, this pattern was even recreated in the English language when the suffixes were borrowed, based on the words English speakers were encountering.

5. Conclusion

Rival suffixes *-ic* and *-ical*, made of borrowed elements, are both productive today because *-ical* has carved out a morphologically constrained productive niche: it is preferred in stems ending in *olog*. This pattern is clear when comparative productivity is measured using data from Google, as well as a comparative sample from COCA, a more traditional corpus.

We saw the emergence of productivity in the suffix *-ity*, where whole-word French borrowings over several centuries ultimately led to a productive suffix. Because native rival *-ness* was already productive in English from the start, *-ity* was only able to emerge because of a high co-occurrence in several common Latinate neighborhoods (including *able*, *al*, *ar*, and *ic*), allowing a niche pattern to enter the system. This pattern of co-

occurrence continued to be the basis for deriving new *-ity* forms, even as suffix *-able* began to extend its productivity beyond Latin stems.

Lastly, while *-ize* is more productive than *-ify* overall, *-ify* is preferred overwhelmingly in words with monosyllabic stems. English, as well as other Romance languages, owes this particular pattern to Latin. Its native suffix *-ificare*, grandfather of *-ify*, developed through grammaticalization, and as such, conformed to Latin patterns of affixation that preferred monosyllabic stems. When Latin later began borrowing *-izare* words from Greek, these words contained naturally longer Greek stems. Though accidental, this distribution was recognized as a pattern by speakers and formed the basis of the productive niches of these two suffixes, which propagated into Spanish, Portuguese, and French (and probably others). Later, like *-ity*, *-ic*, and other borrowed suffixes, *-ize* and *-ify* entered English via whole-word borrowings, in this case from French. Since both *-ize* and *-ify* words were being borrowed into English at the same time, the same pattern was recognized by English speakers and recreated as these suffixes emerged simultaneously as productive affixes several centuries later.

In each of these rivalries, we see the organization of new elements entering the language. Suffix *-ical* emerged from the conflation of two other productive suffixes, and remains productive despite its redundancy because it found a niche in one large subset of stems. Meanwhile, *-ity* became productive in a system that already had a highly productive native suffix, only due to the influx of a preponderance of superficially similar words ending in *-ity* that eventually could not be overlooked. The productive pattern of *-ize* and *-ify* reflects a superficial prosodic difference between Latin and Greek that was many languages (and many centuries) removed by the time the very same derivational pattern emerged anew in English. Since language is an evolving system, this organization emerges gradually out of whatever bits and pieces the system happens to have in front of it.

References

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Baayen, R. Harald. 1993. On frequency, transparency, and productivity. In Booij, G. and J. van Marle (eds.), *Yearbook of morphology 1992*. Dordrecht: Kluwer. 181-208.

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Booij, Geert. 1977. *Dutch morphology: A study of word formation in generative grammar.* Holland: Foris.

Fitch, W. Tecumseh. 2010. *The Evolution of Language.* New York: Cambridge University Press.

Hurford, J. 1990. Nativist and functional explanations in language acquisition. In I. M. Roca (ed.), *Logical Issues in Language Acquisition*. Dordrecht: Foris Publications, 85-136.

Lindsay, Mark and Mark Aronoff. 2012. Natural Selection in Self-Organizing Morphological Systems. To appear in Fabio Montermini, Gilles Boyé, and Jesse Tseng (ed.) *Selected Proceedings of the 7th Décembrettes*. Germany: Lincom Europa.

Marchand, Hans. 1969. *The categories and types of present-day English word-formation. A synchronic-diachronic approach*. Munich, Germany: Beck.

Plag, Ingo. 1999. *Structural constraints in English derivation*. Berlin: Mouton de Gruyter.

Steels, Luc. 1997. The Synthetic Modeling of Language Origins. In Gouzoules, H. (ed.), *Evolution of Communication, vol. 1, nr. 1*. Amsterdam: John Benjamins Publishing Company, 1-34.

Italian -ismo/-ista and Ancient Greek -ismós/-istḗs formations. Morphological processes and diachronic relationships*

Heike Necker
Universität Zürich
hnecker@rom.uzh.ch

Liana Tronci
Università per Stranieri di Siena
tronci@unistrasi.it

1. Preliminaries, corpus and literature

The aim of this study is to examine Italian and Ancient Greek noun formations which are characterized by the highly productive suffixes It. *-ismo* and *-ista* and AGr. *-ismós* and *-istḗs*, such as for instance It. *bonapartismo* 'Bonapartism, i.e. behaviour like a Bonaparte, a personal and authoritarian regime based on direct popular consensus with demagogic and plebiscitary tendencies; support for the restoration of the Napoleonic Dynasty', *bonapartista* 'Bonapartist' and AGr. *lakōnismós* 'imitation of the Lacedaemonian manner; acting in the Lacedaemonians' interest', *lakōnistḗs* 'one who imitates the Lacedaemonians; one who sides with the Lacedaemonians'. Traditionally, grammarians and scholars underline the high productiveness of our suffixes in both languages; suffice it to take a look at Schwarze (1995), Grossmann & Rainer (2004), Dardano (2009) for Italian, and Chantraine (1933), Schwyzer (1953) for AGr. The same holds, however, for the etymologically corresponding suffixes in many modern European languages: nouns such as e.g. German *Marxismus*, *Marxist*, English *Marxism*, *Marxist*, French *marxisme*, *marxiste* are considered to form so-called *internationalisms* (see e.g. Braun, Schaeder & Volmert 2003) and indeed often the English translation of our examples also presents *-ism*.

Our interest in this topic originates from the fact that, in both modern languages and Ancient Greek, noun formations characterized by these suffixes appear to share not only morphological but also "syntactic", that is, combinatorial properties. By means of a simple morphological procedure, it may be observed that lexical bases are variable as far as the lexical category is concerned (see § 2). Besides, it may be noticed that the relationship between lexical base and derived noun yields different derivational processes, which are manifested by different lexical meanings (see § 1.3.).

This paper tries to illustrate these different derivational processes, considering both Ancient Greek and Italian (see § 1.1). As regards Ancient Greek, a synchronic survey will be furnished, in order to clarify what kind of derivational processes are involved in the *-ismós*/*-istḗs* noun formations (§ 4). As regards Italian, the synchronic survey will be paralleled by a diachronic research on the more ancient texts of Italian (see § 6). Similarities and differences between derivational processes of the two languages will emerge from our research.

* All the conclusions expressed in this work come from close collaboration between the two authors. Italian regulations on authorship require the authors to specify which sections they are responsible for: sections 1.1, 1.2, 2 and 6 are the work of H. N., whilst 1, 1.3, 1.4, 3, 4, 5 and 7 were contributed by L.T. The research of L.T. has been developed within the PRIN 2008 project *Contact and change in the history of Mediterranean languages* (PRIN 2008). We would like to thank Nunzio La Fauci, for comments and advice on various stages of the work and Ronald Packham for his help with the English. We are most grateful to the organizers of the 8th Mediterranean Morphology Meeting for the possibility of contributing to the present publication.

1.1. Languages and corpora

Italian and Ancient Greek are the two languages this research deals with. Italian has been chosen among other modern languages showing similar derivational processes, because of our knowledge and specific interest in this language. However, this research will subsequently be extended to other modern languages. As for Ancient Greek, it is important to underline that the noun forms in question are the etymological origin of the suffixes of all modern languages, at least from the point of view of form. Some AGr. *-ismós/-istḗs* nouns become Latin loanwords, with *-ismus/-ista*: for instance, AGr. *barbarismós* 'barbarism' or *exorkistḗs* 'exorcist' become Lt. *barbarismus*, *exorcista*; then, starting from Latin, these formations spread through modern languages: It. *-ismo/-ista*, French *-isme/-iste*, English *-ism/-ist* and German *-ismus/-ist*.

As we will show below, the high productiveness of both derivational suffixes in Italian (and in other modern languages) cannot be simply explained by recourse to diachronic relationships, that is, lexical borrowing (from Ancient Greek to Latin) and transmission (from Latin to modern languages). With the exclusion of forms such as It. *idiotismo* 'idiotism', *solecismo* 'solecism', *sillogismo* 'syllogism', *eufemismo* 'euphemism', which derive from the corresponding AGr. formations *via* Latin[1], the derivational processes by means of which new It. *-ismo/-ista* forms are created have to be explained in terms of synchronic word-formation rules. For this reason, we examined all *-ismo* and *-ista* formations within the dictionary Zingarelli 2003 (together with Zingarelli 2008 on Cd-rom): we found 1570 *-ismo* forms and 1380 *-ista* forms. Scholars observe that the majority of Italian *-ismo* and *-ista* formations are systematically related to each other (Schwarze 1995: 502) and are created starting from extremely different lexical bases, such as adjectives (e.g. *totalitarismo* 'totalitarianism'), proper names (*bonapartismo* 'Bonapartism'), but also whole phrases as in *menefreghismo* 'couldn't-care-less attitude, so-what attitude' (see § 2). This is an important point, not only for Italian, but also for Ancient Greek noun formations. As regards Ancient Greek, our lexical collection was made within the *Greek-English Lexicon* by Liddell, Scott & Jones: 860 *-ismós* forms and 530 *-istḗs* forms were found. From the comparative point of view, it is interesting to notice that:

(a) *-ismós* and *-istḗs* suffixes show a morphological relationship such as in *kordakismós* 'the dancing of the Kordax' and *kordakistḗs* 'dancer of the Kordax';

(b) categorially different lexical bases may be recognized in Ancient Greek derivational processes as well: adjectives such as in *psellismós* 'stammering' (from *psellós* 'stammering'), proper names such as in *philippismós* 'Philippism, to be on Philip's side' (from *Phílippos* 'Philip'), prepositional phrases such as in *skorakismós* 'lit. going to the crows, going to hell!' (from *es kórakas* 'lit. to the crows, get lost! go to hell!'). For a more detailed description on this point, see § 2., as well as Chantraine (1933: 138f.) and Schwyzer (1953: 493).

1.2. Literature

We will take a short look at two approaches in relevant literature. To our knowledge, there are not many morphological studies on this topic, neither for Italian nor for Ancient Greek. The first study we will take into consideration is one on word formation in Italian by Grossmann & Rainer (2004). Here the primary criterion for classification is the semantic (or rather interpretative) category. The authors classify It. *-ismo* formations into four groups, based on the meaning of the derivate:

[1] They are the only four *-ismo* forms which occur, for instance, in the novel by Alessandro Manzoni *I promessi sposi* (1840).

1) approaches/conceptions of every type (political, philosophical, scientific, religious, artistic and also individual): e.g. *mitterandismo* (the base is a proper name, 'Mitterandism'), *assolutismo* (the base is an adjective, 'Absolutism'), *trasformismo* (the base is a verb, in biology 'evolutionism', in politics 'transformism'), *altruismo* ('altruism'), *egoismo* ('egoism');

2) social phenomena: e.g. *abusivismo* 'tendency to consider wrongful acts a behavioural norm', *analfabetismo* 'analphabetism', *dantismo* 'study or cult of Dante, word or locution coined by Dante' and the subset of sporting activities, *alpinismo* 'alpinism';

3) medical terms (diseases, anomalies): e.g. *alcolismo* 'alcoholism', *reumatismo* 'rheumatism';

4) linguistic characteristic, peculiarity: e.g. *anglicismo* 'Anglicism', *arabismo* 'Arabism'.

First, this approach surely does not account for the polysemy of some *-ismo* forms (e.g. *dantismo* can be interpreted as 'study of Dante or imitation of Dante') clearly identifiable within a context of occurrence, or the relation to the base. Furthermore, a category where *dantismo* and *alpinismo* are put together may seem somehow rough-and-ready.

We will take a short look at the study of Roché (2007) where the French *-isme* forms are analysed. Roché is confronting the notion of *modèle de construction des mots* with the notion of rules of formation, and adds restrictions and two frequently neglected parameters to his model: the influence of the existing lexicon, on the one hand, and phonological restrictions of well-formedness, on the other. He illustrates the possible interference of existing forms (in the lexicon) with morphological circumstances and phonological restrictions. In opposition to purely referential classifications, Roché subdivides the formations with *-isme* into three models capable also of capturing the formations with *-iste* respectively:

- Model 1) is a so-called axiological relation (see e.g. Greimas 1969) and the basis is referred to and evaluated *via* the three traditional axes (good, beautiful, true): *être favorable à l'esclavage* 'to be in favour of slavery' (*esclavagisme = le fait d'...* 'slaverism'*, esclavagiste = la personne qui est favorable...* 'defender of slavery'), *tout ramener à soi* (*egoisme, egoiste* 'egoism', 'egoist'). *Fédéralisme* 'federalism' which is not a quality noun but '*le fait de privilégier ce qui est fédéral*'. The bases are mostly nouns, adjectives or verbs.
- Model 2) instead forms process nouns (*nom processif*) in parallel with agent nouns (sometimes only one of both): *exorciser → exorcisme, exorciste* 'exorcism, exorcist', *parachute → parachutisme, parachutiste* 'parachuting, parachutist'. The base is a verb or a noun denoting a participant in the process (e.g. instrument).

Within models 1 and 2 the formations are doubly motivated, on the one hand by the relation to the base, on the other by the relation of the derivatives with *-isme* with those with *-iste*.

- Model 3) captures formations with *-isme* that are quality nouns formed on the base of adjectives or nouns referring to a person who does not have a respective formation with *-iste*: *parallélisme* 'parallelism', *dilettantisme* 'amateurism, dilettantism'.

Model 3) is similar to de-adjectival nouns denoting a characteristic/an attribute and quality nouns formed on the base of person denotations with the suffixes *-ie, -erie, -ise, -itude, -at*. Model 2) is similar to action nouns formed with *-age, -ement, -ion* as far as *-isme* is concerned, and *-eur, -ier* for *-iste*. Model 1) finally, has only equivalents with *pro-* and *anti-* derivatives.

In our opinion, the criteria used for classification are too much based on interpretation and do not pay enough attention to derivational processes, that is to the morphosyntactic relationships between *-ismo* nouns and their lexical bases. Furthermore, these criteria rely only on dictionary indications; needless to say, often the forms in question do not have only one meaning, as already mentioned above.

As far as Ancient Greek is concerned, grammarians notice the correlation between *-ismós*/*-istḗs* formations and verb forms in *-ízō*; such as in the case of *lakōnismós* 'imitation of the Lacedaemonian manner; acting in the Lacedaemonians' interest', *lakōnistḗs* 'one who imitates the Lacedaemonians; one who sides with the Lacedaemonians' and *lakōnízō* 'to imitate Lacedaemonian manners'. Leaving aside this correlation, grammarians do not mention other morpho-lexical properties of these classes: in particular, they do not underline the important point concerning the different lexical values which *-ismós* noun forms show with respect to their lexical bases. This point will be dealt with in what follows.

1.3. Italian and Ancient Greek: first comparative remarks

The first morphological comparative aspect to underline is that both languages show parallel derivations from the same lexical base, but also derivations from *-ismo* → *-ista* and vice versa cannot be excluded (for It., see Grossmann & Rainer 2004: 257f.). Interestingly, a linear co-occurrence of both suffixes is never found (neither X-*ismista* nor X-*istismo*). The same holds for the corresponding AGr. suffixes *-ismós* and *-istḗs* (neither X-*ismistḗs*, nor X-*istismós*).

From the morpho-lexical point of view, the most surprising aspect is the polysemy correlated to the derivational processes involving both Italian and Ancient Greek suffixes, particularly the suffixes It. *-ismo*, AGr. *-ismós*. This aspect is not stressed enough in literature, in our opinion. In order to show this, let us consider words such as It. *dantismo* and *petrarchismo* which are both characterized by proper names as their lexical base, designating two Italian poets. Nevertheless, their morpho-lexical values are not the same, and these different values spring from their syntactic combinations. In the dictionary Zingarelli 2008, these two words are glossed:

(1) *dantismo*: 1. Studio, imitazione di Dante; 2. Parola o locuzione coniata da Dante.[2]

(2) *petrarchismo*: 1. Imitazione dello stile di Petrarca; 2. Corrente poetica diffusa per secoli in Europa, ispirata alla lirica del Petrarca.[3]

According to the dictionary, these two words both have two different meanings, but the different meanings of the first word do not correspond to the two different meanings of the second word. If we join to the comparison the word *manzonismo*, the puzzle becomes even more complicated:

(3) *manzonismo*: 1. Imitazione dello stile letterario del Manzoni; 2. Teoria linguistica del Manzoni per cui il fiorentino parlato dalle persone colte era da scegliere come modello di lingua nazionale italiana.[4]

The same holds for the Ancient Greek *-ismós* formations. Let us compare the following lexemes and their glosses, taken from the *Greek-English Lexicon* by Liddell, Scott & Jones:

(a) *attikismós* 'siding with Athens, loyalty to her', from *Attikós* 'Attic, Athenian' (see also, for a different meaning, *attikistḗs* 'one who affects or collects Attic expressions');

(b) *dōrismós* 'speaking in the Doric dialect', maybe from *Dōrieús* 'Dorian, descendant of Dorus';

(c) *lakōnismós* 'imitation of the Lacedaemonian manner; acting in the Lacedaemonians' interest', from *Lákōn* 'Laconian or Lacedaemonian';

(d) *krētismós* 'Cretan behaviour, i.e. lying', from *krḗs*, *krētós* 'Cretan';

[2] 'Dantism: 1. study, imitation of Dante; 2. word or idiom coined by Dante.'

[3] 'Petrarchism: 1. imitation of Petrarch's style; 2. poetic movement widespread in Europe for centuries inspired by Petrarch's lyric poetry.'

[4] 'Manzonism: 1. imitation of Manzoni's literary style; 2. Manzoni's linguistic theory that implies that the Florentine spoken by educated people has to be chosen as a model for the national Italian language'.

(e) *kilikismós* 'Cilician behaviour, i.e. drunken butchery', from *Kílix*, *Kílikos* 'Cilician'.
All these *-ismós* forms are related to lexical bases which designate the name of a people, but their glosses differ from each other, involving a political party, a linguistic device, general behaviour.

So, faced with all these different meanings in both languages, the question is: how can these different meanings be explained? The answer we will try to give in this paper is, that the different meanings are effects of the different lexical-syntactic combinations, so the different meanings furnished by the dictionaries are only the most superficial effect of combinative processes which involve syntax, lexicon and morphology.

1.4. Aims and methodology of analysis

Our study aims to describe the morphological processes involved in the derivation of It. *-ismo*/*-ista* and AGr. *-ismós/-istḗs* formations, making a contribution to synchronic and intra-linguistic morphology on the one hand, and diachronic and contrastive phenomena on the other hand. In fact, our approach to the two languages under analysis will be twofold and diverse in this study. The analysis of Ancient Greek will be undertaken mainly from a synchronic point of view, while the part for Italian will adopt a diachronic perspective. These different approaches are motivated by various arguments. First of all, the differences can be found in language itself: in the case of Ancient Greek, we are faced with an ancient language with a closed corpus, whereas for Italian we are facing a modern living language with an open corpus. For Ancient Greek no neologisms, new formations or native speakers can be found, for Italian every day new formations can show up, but in contrast with Italian we have a close relation with verbs in *-ízō* in Ancient Greek as we will see (therefore not only the couple of *-ismós*/*-istḗs* but also *-ízō* has to be taken into consideration, whereas in Italian we find the etymological continuation of *-ízō*, the suffix *-eggiare*, but there is no comparable relationship to *-ismo* and *-ista* as in Ancient Greek). In order to deal with a huge number of modern forms (deriving from manifold sources, e.g. Ancient Greek borrowings, loanwords from other languages, Italian formations, erudite formations of e.g. scientific terminology) we need to disentangle first the Italian development – therefore our study of Italian will here be limited to Old Italian. Furthermore, Ancient Greek represents the origin of the Italian suffixes. But not only, for, as we will see, a great number of these forms are borrowed from Ancient Greek directly or *via* other languages (e.g. French, English, German). So Ancient Greek has to be analyzed as the model for new formations from the beginning until now. Ancient Greek is the starting point, but also remains a parallel means for borrowings until the present.

2. It. and AGr.: morphological similarities and differences

In both languages, *-ismo*/*-ista* and *-ismós*/*-istḗs* are combined with different types of lexical bases and are highly productive:

- ✓ Nouns: It. *abolizionismo* 'abolitionism', *abolizionista* 'abolitionist' (*abolizione* 'abolition'); AGr. *oiōnismós* 'omen from the flight or cries of birds', *oiōnistḗs* 'one who foretells from the flight and cries of birds'(*oiōnós* 'bird'), *tokismós* 'usury, the business of the usurer' (*tókos* 'interest, profit, fruits'), *ptualismós* 'ptyalism' (*ptúalon* 'saliva'), *melismós* 'dismemberment' (*mélos* 'member, part'), *andrismós* 'virility' (*anḗr, andrós* 'man');
- ✓ Proper Names: It. *bonapartismo* 'Bonapartism', *bonapartista* 'bonapartist' (*Bonaparte*); AGr. *philippismós* 'siding with Philip' (*Phílippos* 'Philip'), *akkismós* 'affectation' (*Akkṓ* 'Acco', name of a woman in Plutarch), *antisthenismós* 'life according to the doctrine of Antisthenes' (*Antisthénēs* 'Antisthenes');

- ✓ Adjectives: It. *ambientalismo* 'environmentalism', *ambientalista* 'environmentalist' (*ambientale* 'environmental'); AGr. *psellismós* 'stammering', *psellistḗs* 'stammerer' (*psellós* 'stammering'), *homalismós* 'levelling, smoothing, uniformity' (*hómalos* 'flat, level, equal');
- ✓ Adverbs: It. *pressappochismo* 'superficiality', *pressappochista* 'careless, inaccurate, sloppy person' (*pressappoco* 'roughly, about'); AGr. *badismós* 'walking, going', *badistḗs* 'goer' (*bádēn* 'step by step'), *opsismós* 'slowness, tardiness' (*opsé* 'late');
- ✓ Compound words: It. *terzomondismo* 'Third-worldism', *terzomondista* 'Third Worlder' (*terzo mondo* 'Third World'); AGr. *tetrapodismós* 'a going on all fours', *tetrapodistḗs* 'one who goes on all fours' (*tetrápous* 'quadruped'), *allophulismós* 'adoption of strange/foreign customs' (*allóphulos* 'stranger, foreign');
- ✓ Idioms: It. *cerchiobottismo* 'in journalism: attitude of expressing appreciation and criticism at the same time versus two contrasting positions', *cerchiobottista* 'someone showing *cerchiobottismo'* (based on the idiom *dare un colpo al cerchio e uno alla botte* 'to give a hit to the hoop and one to the barrel'), *panciafichismo* ('derogatory for neutralism during the First World War, based on the idiom *salvare la pancia per i fichi* 'to save the belly, stomach for the figs'); AGr. *skorakismós* 'contumely' (*es kórakas* 'go and be hanged', lit. [go] to the ravens'), *koinismós* 'mixture of various dialects' (*koinḗ diálektos* 'ordinary language').

Concerning It. only:

- ✓ Numerals: *diciannovismo* 'complex of phenomena characterizing the political phase immediately following the First World War', *diciannovista* 'participator in the Fascist Party from its beginning' (*diciannove* 'nineteen', i.e. 1919, year of the foundation of the Fascist Party);
- ✓ Prepositions: *dietrismo* 'tendency to sense hidden intrigues and manoeuvres everywhere', *dietrista* 'someone sensing hidden intrigues and manoeuvres everywhere' (*dietro* 'behind');
- ✓ Whole phrases: *menefreghismo* 'couldn't-care-less attitude, so-what attitude', *menefreghista* 'person who could not care less' (*me ne frego* 'I don't give a damn');
- ✓ Abbreviations and acronyms (only for *-ista* formations): *aclista* 'member of the ACLI' (*ACLI = Associazioni cristiane lavoratori italiani* 'Christian associations of Italian workers'), *ciellenista* 'supporter of the CLN' (*CLN = Comitato di liberazione nazionale* 'National committee for the liberation of Italy'), *piduista* 'member of the secret P2 Masonic lodge' (*P2 = Propaganda due* 'Propaganda two').

This variety of possible bases is in neat contradiction with the Unitary Base Hypothesis of Aronoff (1976), i.e. the idea that derivational suffixes are limited to bases of one lexical category. The suffixes analyzed here are not the only contradictions to the UBH, e.g. the Italian modifying suffixes (diminutive, augmentative and pejorative suffixes) combine with various lexical categories; whereas the presented richness of our suffixes is unique, exceeding even the modifying suffixes.

3. The role of Latin

It is well known that diachronic relationships between AGr. and It. have to be described *via* Latin. Is it true for It. *-ismo*/*-ista* and AGr. *-ismós*/*-istḗs* formations as well? So, Lt. *barbarismus*, *cynismus*, *exorcismus*, *iconismus*, *iudaismus*, *priapismus* – which are directly connected with It. *barbarismo*, *cinismo*, *esorcismo*, *iconismo, giudaismo*, *priapismo* – are to be considered as Lt. loanwords from AGr. *barbarismós* 'use of a foreign tongue, barbarism', *kunismós* 'Cynical philosophy or conduct', *exorkismós* 'administration of an oath', *eikonismós* 'delineation, description', *ioudaismós* 'Judaism', *priapismós* 'priapism'. It is the same for Lt. *-ista* formations such as *citharista* 'a player on the cithara', *evangelista* 'an evangelist', *exorcista* 'an exorcist', *panegyrista* 'a eulogist', *psalmista* 'a composer of

psalms', which are Lt. loanwords from AGr. *kitharistḗs* 'player on the cithara', *euangelistḗs* 'bringer of good tidings, evangelist', *exorkistḗs* 'exorcist', *panēguristḗs* 'one who attends a *panḗguris*, i.e. a national or general assembly', *psalmistḗs* 'psalmist', subsequently occurring also in It. *citarista*, *evangelista*, *esorcista*, *panegirista*, *salmista*. Unlike Ancient Greek and Italian, Lt. *-ismus/-ista* formations are not productive, as Werner (1980: 492) points out: "das Suffix […] im antiken Latein wohl nicht produktiv gewesen und wahrscheinlich ebensowenig im Mittellatein". Actually, only 42 *-ismus* and 48 *-ista* formations were found in the on-line Latin dictionary by Lewis & Short (http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0059). Mostly, they are loanwords from Ancient Greek, and Latin therefore did not have any productive word-formation rule involving *-ismus*/*-ista* suffixes. Consequently, the huge productivity of both It. *-ismo*/*-ista* and AGr. *-ismós*/*-istḗs* formations may not be explained *via* Latin, and has to be described first from a synchronic point of view, with respect to the two different linguistic systems (i.e. Ancient Greek and Italian), then, as far as Italian is concerned, from a diachronic point of view, which is important in order to account for the relationship between inherited forms and new formations. According to this diachronic point of view, the present research will be focused on Old Italian texts (see § 6). We start with a synchronic survey of Ancient Greek.

4. Ancient Greek

4.1. Paradigmatic relationships: *-ismós*, *-istḗs*, *-ízō*

The most interesting paradigmatic relationship concerning *-ismós*/*-istḗs* noun formations is the one with *-ízō* verb forms: with few exceptions (which might be due also to gaps in AGr. documentation) *-ismós*/*-istḗs* nouns are paralleled by *-ízō* verb forms. Morphologically, both formations appear to be composed starting from lexical bases which are variable from a categorial point of view (nouns, adjectives and so on: cf. § 2). As far as the semantic value is concerned, both noun and verb formations show similar relationships with their lexical bases. The case of the couple *xenismós* / *xenízō* is a well-suited example because both derived forms show two different meanings: *xenismós* (a) 'hospitality', (b) 'strangeness'; *xenízō* (a) 'to host', (b) 'to behave, to talk like a stranger'. These different meanings depend obviously on the different meanings of the lexical base *xénos* (a) 'host', (b) 'stranger', but it is important to underline that the two different meanings of the derivatives are created in the syntactic combinations in which these forms occur, therefore derivational processes are to be described from a morphosyntactic point of view. The following examples are extracted from AGr. literary texts and show occurrences of *xenismós* (a) 'hospitality' (ex. 1), (b) 'strangeness' (ex. 2) and *xe(i)nízō* (a) 'to host' (ex. 3), (b) 'to behave, to talk like a stranger' (ex. 4).[5]

(1) *tòn gàr toû Hērakléous* ***xenismòn*** *prṓiēn hēmîn en poiḗmatí tini diéiei, hōs dià tḕn toû Hērakléous suggéneian ho prógonos autôn hupodéxaito tòn Hērakléa* (Pl. *Ly.*205c7)
'Only two days ago, he was recounting to us in some poem of his the **entertainment** of Hercules, how on account of his kinship with Hercules, their forefather welcomed the hero'

(2) *Tà gàr huperaíronta tḕn koinḕn sunḗtheian tôn par'eníois ethismôn, hótan mèn autopathôs dóxēi gínesthai dià tò mégethos tôn sumptōmátōn, éleon ekkaleîtai parà*

[5] English translations of all Ancient Greek literary examples are drawn from the website: http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman

toîs horôsi kaì toîs akoúousi, kaì sugkineî pōs hékaston hēmôn ho ***xenismós*** (Plb. XV 17,1)
'Manifestations of emotion which go beyond what is customary among a particular people, if they are thought to be the result of genuine feeling evoked by extraordinary disasters, excite pity in the minds of those who see or hear them; and we are all in a manner moved by the **novelty** of the spectacle.'

(3) *Toûton kaì* ***xeinízete*** *kaì eû poieûntes phaínesthe* (Hdt. 9,89)
'It is for you to **entertain** him, and show that you do him good service'

(4) *diabeblḗkasi gár mou tòn patéra, hōs* ***exénizen*** *[...] hṓsper dè déon hēmâs di'ekeínas tàs atukhías apolésthai, tò xenízein autoû katēgorḗkasin* (Dem. *57*,18)
'They have maliciously asserted that my father **spoke with a foreign accent**. […] but just as though it were right that I should be brought to ruin on account of his misfortunes, they have made his foreign accent the basis of a charge against him'

4.2. Derivational processes: at the interface of morphology and syntax

These examples clearly show that (a) there is a close relationship between *-ismós* and *-ízō* forms, as far as the derivational processes are concerned, and (b) different lexical meanings of *-ismós* (and *-ízō*) forms are surface effects of different morphosyntactic processes. The above examples illustrate two different derivational processes involving *-ismós* and *-ízō*:

1. A sort of "transitivizing" derivational process: with regard to the lexical base, *-ízō* verb forms may be paraphrased as 'to cause to be X' and *-ismós* noun forms as 'the fact of causing to be X', where X designates the lexical base. This transitivizing derivational process correlates with the occurrence of a new subject concerning the argument structure of the lexical base: this subject may be interpreted as 'who gives hospitality, who causes to be a host'.
2. A derivational process which is, in contrast with the preceding, "non-transitivizing" and in which *-ízō* verb forms may be interpreted as 'to behave like X, to pretend to be X', and *-ismós* noun forms as 'the fact of behaving like X, of pretending to be X'. Also in this second kind of derivational process there is a new subject function, which is interpreted as 'who behaves like X, pretends to be X'. The difference between the two morphosyntactic processes concerns the syntactic combinations in which the lexical bases occur as nouns, adjectives and so on. We will not dwell on this point. Suffice it to say that AGr. *-ízō* verbs are correlated, from an etymological point of view, with the It. *-eggiare* verb forms (cf. Tronci 2010 for a survey on AGr. *-ízō* verbs). Beyond etymological correspondences, AGr. *-ízō* and It. *-eggiare* verb forms show interesting similarities as regards syntactic and textual values. Both verb formations (for It. *catoneggiare*, *americaneggiare* and so on, for AGr. *xenízō* in the main text) may show, among others, the semantic value 'to behave like X, to pretend to be X' (where X is the lexical base). In some studies of La Fauci (2006, 2007, 2008), firm evidence can be found to support the idea that in the derivational process with *-eggiare* antonomasia is involved. We can assume that antonomasia is at work also in the AGr. derivational process concerning the *-ízō* verbs described here.

Also other *-ismós* noun forms show the same paradigmatic correlation with *-ízō* verb forms, on the one hand, and the same morphosyntactic and lexical variability, on the other hand. Among others, we may mention the formations derived from *sṓphrōn* 'wise': *sōphronismós* (a) 'correction, admonition, warning, hint', (b) 'temperance, wisdom', which is paralleled by the verb form *sōphronízō* (in the active voice) 'to cause to be wise, to correct' and *sōphronízomai* (in the middle voice) 'to be getting wise, to be temperate'.

Other forms testify only the first derivational process: from the adjective *kátharos* 'pure', the noun *katharismós* 'purification' and the verb *katharízō* 'to purify', from the adjective *kainós* 'new', the noun *kainismós* 'renewal' and the verb *kainízō* 'to renew, to cause to be new', from the comparative adjective *neőteros* 'newer', the noun *neōterismós* 'innovation, political revolution' and the verb *neōterízō* 'to make innovations', from the noun *spádōn* 'eunuch', the noun *spadonismós* 'weakening (of a sound etc.)' and the verb *spadonízō* 'to cause to be weak and slack'. Most of these *-ismós* forms are paralleled also by *-istḗs* forms, such as *sōphronistḗs* 'corrector, one who moderates', *kainistḗs* 'renovator', *neōteristḗs* 'innovator' and so on. All *-istḗs* forms sharing these morphosyntactic relationships show the same lexical interpretation: they refer to the new subject involved in the derivational process, so they express 'one who V-*ízō* / one who makes N-*ismós*'.

The same kind of relationship between *-ismós* and *-istḗs* formations may be observed also in the second type of derivational process. In this case as well, the noun's formation is "mediated" by means of *-ízō* verb forms. So, *-ismós* nouns take the intransitive syntax of the verb and its antonomastic value 'the fact of behaving like X, of pretending to be X' (X = lexical base), and *-istḗs* nouns are correlated with the expression of the subject of the corresponding *-ízō* verb forms 'one who behaves like X, one who pretends to be X'. The morphological process of this second type applies preferably to lexical bases which may be involved in antonomasia: in particular, proper names, such as *Puthagóras*, which is the lexical base for the series *puthagorismós* 'Pythagorean doctrine', *puthagoristḗs* 'follower of Pythagoras', *puthagorízō* 'to be a disciple of Pythagoras', people names, such as *Lákōn* 'a Laconian, Lacedaemonian', which is the lexical base for the series *lakōnismós* 'imitation of Lacedaemonian manner', *lakōnistḗs* 'one who imitates the Lacedaemonians', *lakōnízō* 'to imitate Lacedaemonian manners', but also appellative nouns which syntactically became proper names, by means of antonomasia, such as the case of *xénos* 'stranger', for the series discussed above.

4.3. How may the relationship between *-ismós/-istḗs* and *-ízō* be explained?

The paradigmatic relationship between *-ismós/-istḗs* nouns and *-ízō* verb forms is a matter of fact, as we have shown above. The few gaps could be caused by lack of attestation or by loss in literary transmission, which are the usual risks we always run in linguistic research on closed corpora. Leaving aside these gaps, we might affirm that each *-ismós/-istḗs* noun is paralleled by an *-ízō* verb form, and this morphological parallelism is supported by the syntactic-semantic value of all derivatives, according to systematic derivational rules. It is true, however, that AGr. literary texts give evidence for the assumption that both *-ismós/-istḗs* and *-ízō* forms were created ad hoc and were often short-lived; sometimes they are documented in only one text. An adequate example is the couple *philippismós* 'siding with Philip' and *philippízō* 'to be on Philip's side, in his party'. These two words are presumably created on the occasion of a political dispute between Demosthenes and Aeschines, and are documented in their works *De corona* and *In Ctesiphontem*, respectively. Obviously, these two lexemes refer to the political attitude of the Macedonian king Philip, who tried to expand his own political power at the expense of Athens. Demosthenes was against Philip's political attitude, whereas Aeschines was in favour of it. So, from a linguistic point of view, the proper name Philip undergoes an antonomastic process and is thus ready to occur as a lexical base for *-ismós* and *-ízō* forms, belonging to the second derivative process described here, as the following examples show:

(5) *allà tí taût'epitimô, pollôi skhetliőter'álla katēgorēkótos autoû kaì katepseusménou? hòs gàr emoû* ***philippismón****, ô gê kaì theoí, katēgoreî, tí hoûtos ouk àn eípoi?*(Dem.*18*,294)

'But why reproach him for that imputation, when he has uttered calumnies of far greater audacity? A man who accuses me of **Philippism** – Heaven and Earth, of what lie is he not capable?'

(6) *ou perì toútōn Ameiniádēs mèn proúlegen eulabeîsthai kaì pémpein eis Delphoùs eperēsoménous tòn theòn hó ti khrḕ práttein, Dēmosthénēs dè antélege **philippízein** tḕn Puthían pháskōn, apaídeutos ôn kaì apolaúōn kaì empimplámenos tês dedoménēs huph'humôn autôi exousías?* (Aesch. *In Ctesiph*. 130)
'In view of this did not Ameiniades warn you to be on your guard, and to send messengers to Delphi to inquire of the god what was to be done? And did not Demosthenes oppose, and say that the Pythia **had gone over to Philip**? Boor that he was, gorged with his feast of indulgence from you!'

In the light of the situation illustrated above, we could challenge the status of the derivational suffix *-ismós*. According to a simple morphological procedure, such as that suggested by grammarians, *-ismós* noun forms appear to be created by means of the noun morpheme *-mó-* joined to the *-ízō* verbal lexical bases. So, from a strictly morphological point of view, *-ismós* forms are deverbal nouns, as Chantraine (1933: 138f.) underlines saying that "Sauf quelques exceptions notées au passage ces dérivés sont tirés de verbes en *-ízō*".[6] In the AGr. grammar by Schwyzer (1953: 493) the same morphological point of view is suggested. In this perspective, the affixes *-ismós* and *-istḗs* are therefore to be interpreted as a combination of the verbal affix *-íz-* and the nominal affix *-mo-*/*-tḗs-*. It is true, however, that these two morphological processes – the one which gives birth to *-ismós*/*-istḗs* noun forms and the one which creates *-ízō* verb forms – appear to be related to each other, but they do not depend mechanically on each other, from a synchronic point of view. First, there are some *-ismós*/*-istḗs* noun forms which do not show corresponding *-ízō* verb forms, so they appear to be created independently. Nouns such as *antisthenismós* 'way of life according to the teaching of Antisthenes', *koinismós* 'mixture of dialects', *allophulismós* 'adoption of foreign customs', *kuphōnismós* 'punishment by the *kúphōn*' and others do not correlate with *-ízō* verb forms, from the morphological point of view. Obviously, this may be caused by gaps in the texts' tradition, but it may also be evidence for an "autonomization" of the noun suffix *-ismós*. Further, it must be added that the nominal suffix *-mo-* is not productive in Ancient Greek (see Ciacci 1999/2000 and Ronzitti 2006). Both underline that the nominal suffix *-mo-* occurs in many nominal formations – in Ancient Greek and in Sanskrit as well – but these nouns are not morphologically analyzable because their meaning is no more compositional. Combined with the verbal suffix *-iz-*, the nominal suffix *-mo-* appears to be highly productive, as we have shown. It may thus be suggested that the high productivity of *-ismós* formations does not depend on the productivity of the nominal suffix *-mo-*, but, rather, on the combination of the two suffixes, that is on *-ismós*. This means that between *-ismós*/*-istḗs* noun forms and *-ízō* verb forms, there is a morphosyntactic relationship but this relationship does not yield a mechanic adjunction of affixes. It may be suggested that already in Ancient Greek (at least in some chronological phases) a process of morphological reanalysis has taken place: *-ismós* nouns are reanalyzed as nouns derived from the same lexical bases as *-ízō* verb forms. The two derivatives do not depend on each other, but they are in a paradigmatic relationship.

5. First conclusions

[6] It is true, however, that Chantraine (1933: 138ff.) is very ambiguous on this topic: on the one hand, he affirms that *-ismós* nouns spread together with, and are derived from *-ízō* verb forms, on the other hand, he points out that it is *-ismós* which gives birth to a great amount of derivatives in Attic and Hellenistic works.

As far as AGr. is concerned, the morphological processes which lead to *-ismós*/*-istḗs* forms are to be described as two derivational processes, applied subsequently. Nouns such as AGr. *barbarismós*, *kunismós*, *exorkismós, eikonismós, ioudaismós, priapismós* and the corresponding *-istḗs* forms (*exorkistḗs* 'exorcist'*, eikonistḗs* 'registrar') stand in a derivational relationship with *-ízō* verb forms: *barbarízō* 'to behave or speak like a barbarian', *kunízō* 'to live like a Cynic', *exorkízō* 'to administer an oath', *eikonízō* 'to copy from a pattern', *ioudaízō* 'to side with or imitate the Jews', *priapízō* 'to be lewd'. The affixes *-ismós* and *-istḗs* are therefore to be interpreted as a combination of the verbal affix *-íz-* and the nominal affix *-mo-*/*-tḗs-*, as Schwyzer (1953: 493) underlines. The two morpho-lexical processes, involving *-ízō* verbs and *-ismós*/*-istḗs* nouns, are both productive and related to each other. In It. there exist two morpho-lexical processes which are etymologically related to those of AGr., namely the one involving verbs ending in *-eggiare* / *-izzare* and the one involving nouns ending in *-ismo*/*-ista*, but there is no comparable systematic morpho-lexical relationship between them in Italian.

6. A closer look at Old Italian

6.1. Modern and Old Italian in view of the productivity of *-ismo* and *-ista* forms

In Modern It. we are faced with a huge amount of data to classify and systematize. A mere classification into semantically based classes is not satisfactory, as we have already pointed out discussing the approach of Grossmann & Rainer (2004). We therefore started the other way round by examining the origin and development of the variety and richness of formations in Italian.

In Modern It. *-ismo* and *-ista* show an extraordinary productivity. It was not the same in Old It., as confirmed by our research on texts ranging from Dante Alighieri (13th century) to Galileo Galilei (first half of the 17th century) available at http://www.liberliber.it/biblioteca/index.htm. In these texts, we searched for the strings *-smo, -sme, -smi, -esim-*, for *-ismo* forms, and *-ista, -iste, -isti*, for *-ista* forms. No productive use of the suffix *-ismo* is found before the 16th century, and couples of *-ismo*/*-ista* forms such as *ateismo* 'atheism' / *ateista* 'atheist' are not found previously. During the 13th-15th centuries, *-ismo* attested forms were related to Lt. or AGr., with a few exceptions, such as *incantesimo* 'spell, enchantment' (13th century) and *ruffianesimo* 'whore-mongering, bootlicking' (14th century). On the contrary, as regards *-ista* forms, the 13th century already presents *Arnaldisti*, *Leonisti*, *Speronisti* denoting supporters, disciples, followers of religious, heretical movements (a proper name is the morphological base in all three cases: *Arnaldo da Brescia, Lione, Ugo Speroni*).

Our research on the texts yields the following results: for *-ismo* roughly 40 types were found; for *-ista* about 90 types (both groups are variable as far as inflection and orthography are concerned, for *cristianesimo* we find *Christianesimo* / *Christianesmo* / *Cristianesimo* / *cristianismo* / *cristianesmo* or for *evangelista*: *Evangelista* / *Vangelista* / *guagnelisti* / *Evangelisti* / *Vangelisti* / *vangelisti* / *evangelisti*). From these types we excluded some forms for the following reasons: *battesimo* 'christening' and *battista* 'Baptist' since both are part of the vocabulary already before the analyzed period, they both have Latin/Ancient Greek origin and a possible base for derivation (the verb *battezzare* 'to christen') is documented much later; furthermore *battista* is mostly used as a proper name in the texts. In the case of *aforisma* 'aphorism', *cataclisma* 'cataclysm' and *sofisma* 'sophism' there is no possible base in Italian and the form is nowadays *-isma* not *-ismo*, whereas in our texts there are some variations with *-ismo*. In *algorisimo* 'algorithm' which seems to be a variant of *algoritmo*, there is no suffix since the base is

the Arabian proper name *al-Huwarizmi*. Finally, *anacronismo* 'anachronism', *parossismo* 'paroxysm', *anabattista* 'Anabaptist', *antagonista* 'antagonist', *gimnosofista* 'gymnosophist', *grammatista* 'grammatist', *lanista* 'lanista, owner and trainer of gladiators', *petaurista* 'flying phalanger, glider' and *sofista* 'sophist' have no possible base. With the only exception of *algorisimo*, all forms have Ancient Greek parallels.

The remaining types for *-ismo* are:
anatematismo 'anathematism', *antimacchiavelismo* 'anti-Machiavellianism', *arianismo* 'Arianism', *ateismo* 'atheism', *barbarismo* 'barbarism', *bardascismo* 'male homosexual prostitution', *Calvinismo* 'Calvinism', *catechismo* 'catechism', *Catolichismo* 'Catholicism', *cristianesimo* 'Christianity', *ebraismo* 'Hebraism', *epitetismi* 'epithetisms', *essorcismi* 'exorcism', *gentilesimo* 'gentilism', *Giudaismo* 'Judaism', *incantesimo* 'enchantment', *Iustinianismo* 'Justinianism', *luteranismo* 'Lutheranism', *Mahumetisimo* 'Muhammadanism', *Monachismo* 'monasticism', *ostracismo* 'ostracism', *paganesimo* 'paganism', *papismo* 'papism', *paralogismo* 'paralogism', *Peripatecismo* 'Peripateticism', *Piratismo* 'piratism', *profanismo* 'profanism', *puttanesimo* 'behaviour like a whore', *ruffianesimo* 'whore-mongering, bootlicking', *Schematismi* 'schematism', *sillogismo* 'syllogism', *soldanesimo* 'Sultanism', *solecismo* 'solecism', *stryanismi* 'wizardries', *trigonismo* 'trigonismi'.

And for *-ista*:
abachista 'book-keeper, from abacus', *Achitofellisti* 'wise counselors like Ahitophel', *albertista* 'follower of Albertus Magnus', *albichista* 'Albigense', *alchimista* 'alchemist', *algebrista* 'algebraist', *Arnaldiste* 'followers of Arnaldo da Brescia', *artista* 'artist', *ateista* 'atheist', *averroista* 'Averroist', *barzellettista* 'writer of a frottola', *cabalista* 'cabalist', *calvinista* 'Calvinist', *canonista* 'canonist', *casista* 'casuist', *chimista* 'chemist', *citarista* 'citharist', *computista* 'book-keeper', *conclavista* 'conclavist', *confessionisti* 'confessionists', *contista* 'book-keeper', *copernichista* 'follower of Copernicus' ideas', *copista* 'copyist', *cronichista* 'chronicler', *donatisti* 'Donatists', *duellisti* 'expert duellists', *essorcisti* 'exorcists', *etimologisti* 'etymologists', *evangelista* 'evangelist', *figurista* 'figurist', *fisionomista* 'physiognomist', *fisionotomista* '?physio-anatomist', *flautista* 'flutist', *galenisti* 'Galenists', *galileista* 'Galilean', *gesuitista* 'Jesuit', *giuristi* 'jurist', *humanista* 'humanist', *ipocratista* 'Hippocratist', *legista* 'jurist, legist', *Leoniste* 'followers of Peter Waldo, the Poor of Lyons', *macchiavellista* 'Machiavellian', *Maumettisti* 'Mohammedans', *meteorista* 'meteorologist', *mineralisti* 'mineralists', *montanista* 'Montanist', *naturalista* 'naturalist', *Navarrista* 'Navarrist', *nobilista* 'honoris cause', *notomista* 'anatomist', *novellisti* 'short story writers', *organista* 'organist', *palacista* (*palazzista*) 'lawyer', *papalista* 'papalist', *papista* 'papist', *paulianisti* 'Paulianists', *persianisti* 'persianists', *priorista* 'book where the priors are listed', *priscillianisti* 'Priscillianists', *Problemista* 'anonymous critic of Galilei and author of the problem *De lunarium montium altitudine*', *prologhista* 'prologist', *pseudotomisti* 'pseudo-Thomists', *regolisti* 'persons obsessed with rule observation', *ribichista* 'player of a rebec', *romanisti* 'Romanists', *salmista* 'psalmist', *scotista* 'Scotist', *scritturisti* 'Scripturists', *semplicista* 'simplist', *sommista* 'summist', *sorbonista* 'Sorbonist', *Speroniste* 'followers of Ugo Speroni', *spigolista* 'puritan, sanctimonious person, bigot', *Talmudista* 'Talmudist', *teoremisti* 'writers of a theorem', *Tomisti* 'Thomists', *umorista* 'follower of humourism in medicine', *vangelista* 'female evangelist', *viglefisti* 'Wyclifists, followers of John Wyclif'.

What are the lexical categories of bases that we find in these types? There are nouns (11 *-ismo*, 47 *-ista*, but 2 cases could also be deverbal), proper names (5 *-ismo*, 24 *-ista*), adjectives (3 *-ismo*, 6 *-ista*) and verbs (6 *-ismo*, 3 *-ista* , but 2 could be denominal, see above). In some cases (10 *-ismo*, 1 *-ista*) it is not easy to decide whether the base is nominal or adjectival, due to the easy shift from one category to another. So for the period observed, we have no examples with compounds, adverbs, idioms, numerals, prepositions, whole phrases or abbreviations as bases.

Some *-ismo*/*-ista* forms which are attested in texts dating back to the 13th-16th centuries are not found in Modern Italian dictionaries (Zingarelli, Devoto-Oli), as the following examples show (they occur in 16th century texts, excepting other indications): *Antimacchiavellismo* 'anti-Machiavellianism', *bardascismo* 'male homosexual prostitution', *epitetismi* 'epithetisms', *Iustinianismo* 'Justinianism', *Piratismo* 'piratism', *profanismo* 'profanity', *soldanesimo* (= *sultanismo* 'Sultanism'), *stryanismi* (= *stregonesimi*, *stregonismi*) 'wizardries', *trigonismo* 'trigonism'; *Arnaldiste*, *ipocratista* 'Hippocratist', *Leoniste*, *Speroniste* (1200); *Maumettisti* 'Mohammedans' (1400); *Achitofellisti* 'wise counselors like Ahitophel', *albertista* 'follower of Albertus Magnus', *albichista* 'Albigense', *barzellettista* 'writer of a frottola', *contista* 'book-keeper', *galileista* 'Galilean', *gesuitista* 'Jesuit', *Navarrista* 'Navarrist', *nobilista* 'honoris cause', *palacista* (*palazzista* 'lawyer'), *papalista* 'papist', *paulianisti* 'Paulianists', *priorista* 'book where the priors are listed', *Problemista* 'anonymous critic of Galilei and author of the problem *De lunarium montium altitudine'*, *prologhista* 'prologist', *regolisti* 'persons obsessed with rule observation', *sorbonista* 'Sorbonist', *spigolista* 'puritan, sanctimonious person, bigot', *teoremisti* 'writers of a theorem', *viglefisti* 'Wyclifists, followers of John Wyclif' (1500). Most of them denote disciples of political or religious trends (this does not seem to be the case with *barzellettista, contista, Problemista, regolisti, spigolista* and *teoremisti*). As far as the lexical categories of bases are concerned, we find here mostly nouns or proper names, a clear sign of the productivity of these types.

6.2. Productivity at work

These above-mentioned forms are a clear sign of the suffix's morphological productivity: each form may appear or disappear, but the morphological processes continue to be viable and productive and the forms are interpretable in the range of the productive processes – briefly, the possibility of forming *ad hoc* formations that are understandable. We will illustrate with two examples.

Galileista is found in a letter of Piero Dini to Galilei (1615):

(7) *ma ne parlerò ancora al medico Fabii Fiammingo, che spesso è in casa mia et è gran* ***Galileista*** *e da' dotti molto stimato;*
'but I will talk of it again to the doctor Fabii Fiammingo, who is often at my home and is a great **Galilean** and highly esteemed by the learned'
(Zingarelli 2003 does not have *galileista*, but *galileiano,* first documented in 1745)

The form *Galileista* is used here in a letter to Galilei himself.

Problemista is found in a letter of Federico Cesi to Galilei (1612):

(8) *et oltre il* ***Problemista****, scopro io qui in altri l'istesso affetto...*
'and apart from the **Problemist**, I find here in others the same affection'

The *Problemista* is an anonymous critic of Galilei and author of the problem *De lunarium montium altitudine*.

As far as the morpho-lexical relationship between *-ismo* and *-ista* is concerned, already OIt. texts show some couples of forms, e.g.: (*Antimacchiavellismo* / *macchiavellista* 'Machiavellian'), *ateismo* 'atheism' / *ateista* 'atheist', *Calvinismo* 'Calvinism' / *calvinista* 'Calvinist', *essorcismi* 'exorcisms' / *essorcista* 'exorcist', *Mahumetisimo* 'Mohammedanism' / *maumettisti* 'Mohammedans', *papismo* 'papism' /

papisti 'papists'. For example in the *Istoria del Concilio Tridentino* of Paolo Sarpi we find *papismo* and several times *papisti*:

(9) *Dicevano che fosse un stabilimento totale del **papismo**; biasimavano sopra tutto la dottrina della giustificazione e che fosse posta in dubio la communione del calice et il matrimonio de' preti.*
'they said that it was a complete establishment of the **papism**; they disapproved above all the doctrine of justification and that the communion of the chalice and the marriage of the priests have been doubted'

(10) *Onde fece il re un editto, proibendo li nomi d'ugonotti e **papisti**, ordinando che, sotto pretesto di scoprir le congregazioni proibite per causa di religione, nissun potesse entrar né con pochi, né con molti in casa d'altri*
'therefore the king made an edict prohibiting the names of the Huguenots and **papists**, and ordered that, with the pretext of uncovering/revealing the prohibited congregations based on religion, nobody could come into someone else's house either with few or with many'

Some types show either *-ismo* or *-ista* affixes only (but they are correlated with *-ista* and *-ismo* forms respectively, which are attested later in monolingual dictionaries indicated here according to Zingarelli) e.g.: *averroista* (*averroismo* 'Averroism' 1941), *catechismo* 'catechism' (*catechista* 'catechist' before 1603), *chimista* (*chimismo* 1863), *copernichista* (*copernicanesimo* 1956), *Cronichisti* (*cronachismo* 'narration according to chronological order' 1917), *donatisti* (*donatismo* 1830), *ebraismo* 'Judaism' (*ebraista* 'Judaist' 1911), *evangelista* (*evangelismo* 'evangelism' 1921), *galenisti* (*galenismo* 1834), *gesuitista* (*gesuitismo* 1668), *ipocratista* (*ippocratismo* 'Hippocratism' 20th century), *meteorista* (*meteorismo* 1788), *montanista* (*montanismo* 1891), *naturalista* (*naturalismo* 1849), *priscillianisti* (*priscillianesimo* 1835), *scotista* (*scotismo* 1960), *semplicista* (*semplicismo* 1904), *Tomisti* (*tomismo* 1749), *umanista* (*umanesimo* 'Humanism' 1891), *umorista* (*umorismo* 1875).

6.3. Classification of Old Italian forms

A first attempt to classify *-ismo* and *-ista* forms which are attested in the 13th-16th centuries may be based on their denotation; this leads to the following classes (but we repeat that this kind of classification is only an intermediate level; ongoing work will help us to go beyond this denotative classification):
A. *-ismo* forms[7]
1. IDEOLOGY, FIELD, CONCEPTION: *anatematismo*, *Antimacchiavellismo*, *arianismo*, *ateismo*, *Calvinismo*, *catechismo*, *Catolichismo*, *cristianesimo*, *ebraismo*, *gentilesimo*, *Giudaismo*, *iustinianismo*, *luteranismo*, *Mahumetisimo, Monachismo, paganesmo, papismo, Peripatecismo, soldanesimo*;
2. RHETORICAL DEVICE, LINGUISTIC PECULIARITY: *barbarismo, epitetismo, paralogismo*, *sillogismo, solecismo*;
3. BEHAVIOUR, ATTITUDE: *bardascismo*, *piratismo*, *puttanesimo, ruffianesimo, Schematismi*.

B. *-ista* forms[8]

[7] Some forms do not fit in this classification, it is the case of: *essorcismi*, *incantesimo*, *ostracismo*, *profanismo*, *stryanismi* and *trigonismo*.
[8] Outside of this classification remain *duellisti*, *essorcisti*, *nobilista*, *priorista* and *vangelista* (female).

1. FOLLOWER, DISCIPLE, SCHOLAR (to A1.): *albertista*, *albichista*, *alchimista*, *algebrista*, *Arnaldiste*, *ateista*, *averroista*, *cabalista*, *calvinista*, *confessionisti*, *copernichista*, *Donatisti*, *galenisti, Galileista, Gesuitista, ipocratista*, *Leoniste*, *macchiavellista*, *maumettista, meteorista, mineralista, montanisti, naturalista, Navarristi, papalista, papista, paulianista persianisti, priscillianisti, pseudotomisti, romanisti, scotista, sorbonista*, *Speroniste, Talmudista, Tomisti, (h)umanista, umorista, viglefisti*;
2. PROFESSIONAL, EXPERT: *abachista*, *canonista*, *casista*, *chimista*, *contista*, *computista*, *conclavista*, *copista*, *etimologista*, *fisionomista*, *fisionotomista*, *giurista*, *legista*, *notomista*, *palacista*, *scritturista*, *semplicista*;
3. AUTHOR, PAINTER etc.: first of all we find the hypernym *artista*; *barzellettista, cronichista, evangelista, figurista, novellista, problemista, prologhista*, *salmista*, *sommista, teoremisti*;
4. MUSICIAN: *citarista*, *flautista*, *organista*, *ribichista*;
5. PERSON CHARACTERIZED BY BEHAVIOUR (to A3.): *Achitofellista*, *regolista*, *spigolista*.

From the beginning the denotation of religious movements (particularly heretical or non-Christian ones) and the related disciples shows up:

- religious movements: a) *cristianesimo – cattolicesimo*; b) non-Christian: *paganesimo, ebraismo, giudaismo, gentilesimo, mahumetismimo*; c) heretical: *arianismo, calvinismo, luteranismo*
- disciples: a) *albertisti*; b) *maumettisti, talmudisti*; c) *arnaldisti, leonisti, speronisti, albichisti, calvinisti, confessionisti, donatisti, montanisti, paulianisti, priscillianisti, viglefisti*.

A short look reveals the chronological distribution of the forms within the classes, first for *-ismo* (table A) then for *-ista* (table B):

Table A

	A-1	**A-2**	**A-3**
13th century[9]	*cristianesimo*, *paganesmo*	*barbarismo*, *sillogismo*, *solecismo*	
14th century			*ruffianesimo*
15th century[10]			
16th century[11]	*anatematismo*, *Antimacchiavellismo*, *arianismo*, *ateismo*, *Calvinismo catechismo*, *Catolichismo*, *ebraismo*, *gentilesimo*, *Giudaismo*, *iustinianismo*, *luteranismo*, *Mahumetisimo*, *Monachismo*, *papismo*, *Peripatecismo*, *profanismo*, *soldanesimo*	*epitetismo*, *paralogismo*	*bardascismo*, *piratismo*, *puttanesimo*, *Schematismi*

[9] Unclear: *incantesimo*.
[10] Unclear: *ostracismo*.
[11] The indication 16th century includes attestations until 1650 as indicated above. Unclear: *essorcismi* (act of *esorcizzare*), *stryanismi* (act of bewitching), *trigonismo* (mathematical concept).

Table B

	B-1	**B-2**	**B-3**	**B-4**	**B-5**
13th century	*Arnaldiste*, *ipocratista*, *Leoniste*, *Speroniste*	*legista*	*artista*, *evangelista*, *salmista*	*citarista*	
14th century	*alchimista*	*canonista*, *giurista*			
15th century	*Maumettista*, *Talmudista*, *(h)umanista*	*notomista*	*cronichista*	*flautista*	
16th century[12]	*albertista*, *albichista*, *algebrista*, *ateista*, *averroista*, *cabalista*, *calvinista*, *confessionisti*, *copernichista*, *Donatisti*, *galenisti*, *Galileista*, *Gesuitista*, *macchiavellista meteorista*, *mineralista*, *montanisti*, *naturalista*, *Navarristi*, *papalista*, *papista*, *paulianista*, *persianisti*, *priscillianisti*, *pseudotomisti*, *romanisti*, *scotista*, *sorbonista*, *Tomisti*, *umorista*, *viglefisti*	*abachista*, *casista*, *chimista*, *contista*, *computista*, *conclavista*, *copista*, *etimologista*, *fisionomista*, *fisionotomista*, *palacista*, *scritturista*, *semplicista*	*barzellettista*, *figurista*, *novellista*, *problemista*, *prologhista*, *sommista*, *teoremisti*	*organista*, *ribichista*	*Achitofellista*, *regolista*, *spigolista*

No formations of A-3 'behaviour, attitude' occur until the 14th century, and B-5, which is the correlated *-ista* formation, is documented only in the 16th century. A control in Zingarelli 2008 and Devoto & Oli (2009) confirmed that there are no attested forms in relation to these classes. The only form cited for *-ista* is: *moralista* (first documentation indicated before 1492 by Z., 15th century by D&O) not found in our texts, but it is unclear when *moralista* came to be interpreted as a person behaving like a moralist and when as a follower of a moralist norm (i.e. belonging to A-1).

7. Conclusions and outlook

In this paper we have suggested some reflections on two classes of nominal suffixes which are closely related to each other from the morpho-lexical point of view: AGr. *-ismós*/*-istḗs* and It. *-ismo*/*-ista*. The comparison of these two languages has allowed us to describe two different derivational processes, both productive in the two languages. Furthermore, we have taken into consideration the diachronic relationship between the two languages, which is mediated by means of Latin. For this reason, our research on Italian *-ismo*/*-ista* formations has concerned first of all Old Italian texts, in order to verify whether these derivational processes were productive also in Old Italian.

[12] The indication 16th century includes attestations until 1650 as indicated above. Unclear *duellisti*, *essorcisti*, *nobilista* (scuola nobilista, a nobilista), *priorista* (book where the priors are registered), *vangelista* (feminine).

In Ancient Greek, the *-ismós/-istḗs* formations are correlated with the *-ízō* verb forms. So, *-ismós/-istḗs* noun forms can be described – from the morphological point of view – as derivatives, by means of the *-mo-* suffix, from derived *-ízō* verb forms. We have shown that the nominal derivational process seems to become autonomous already in the synchrony of Ancient Greek, and consequently *-ismós/-istḗs* derivational suffixes can be recognized. Nevertheless, the *-ízō* verbs and the *-ismós/-istḗs* nouns remain closely related to each other, from the lexico-syntactic point of view. This relationship does not concern Italian forms, even if the etymologically corresponding *-izzare/-eggiare* verb forms are productive in Italian too. There seems to be no systematic relationship between nouns in *-ismo*, *-ista* and verbs in *-izzare/-eggiare* in either Old or Modern Italian[13]. In relation to our formations with *-ismo* and *-ista* in Old Italian texts, we find only some sporadic parallels to verbs. For the interpretation 'behave like X' for example we find only *paganeggiare* 'to think or act according to pagan principles', *pirateggiare* 'to pirate, to commit piracy', *puttaneggiare* 'to whore, to play the whore' and *ruffianeggiare* 'to pimp, to pander' attested in Zingarelli 2008, whereas for the interpretation 'cause to be X' we find *cattolicizzare* 'catholicize' and *cristianizzare* 'to Christianize' attested only in the 19th and 20th century, whereas *ebraizzare* 'to Hebraize' and *giudaizzare* 'to Hebraize' have Ancient Greek predecessors; the only interesting formation is *paganizzare* 'to paganize' coupling with *paganeggiare* to *paganesimo*. Even if the paradigmatic relationship to the verbal suffixes in Old Italian is not systematic as in Ancient Greek, the topic should be analysed extensively.

In Old Italian *-ismo* and *-ista* are productive suffixes, but they did not reach their complete wide range of formations and interpretations until the 16th century. As far as *-ismo* is concerned, a real productive use of the suffix is not documented in our texts until the 16th century, whereas for *-ista*, already the 13th century shows productive formations. This closer look at the development of such productivity based on concrete textual occurrences is, to our knowledge, a new perspective within the study of these suffixes, together with the comparative look at Ancient Greek, which resulted in new insights. Neither suffix exhibited its peculiarity as regards the richness of possible lexical bases: they are limited to nouns, proper names, adjectives and verbs (so they behave virtually like other derivational suffixes). It remains to be analysed when the first formations with the missing bases (i.e. compounds, adverbs, idioms, numerals, prepositions, whole phrases or abbreviations) enter the language. An extensive study of the subsequent periods (17th century until now) should fill the gaps. A surprising "gap", in our opinion, is the lack of *-ista* forms denoting 'one who behaves like X' (B-5) until the 16th century. These interpretations clearly occur in Ancient Greek (e.g. *lakōnistḗs* 'one who imitates the Lacedaemonians') and the corresponding *-ismo* nouns (e.g. *ruffianesimo*) are documented already before the 16th century. A problem for the analysis of Old Italian (at least for the early periods) could also be the fact that the written language was still predominantly Latin.

Another point to be mentioned – and to be further investigated – concerns the fact that in both languages the great amount of *-ismo/-ista* and the *-ismós/-istḗs* forms appear to be created and therefore occur in "technical" that is philosophical, religious, historic, scientific, or rhetoric texts and are, by contrast, only rarely found in literary texts. An example is the one concerning the word *galileista* (§ 6.2 above). Other examples are AGr. lexemes *barbarismós* 'use of a foreign tongue or one's own tongue amiss, barbarism', *iōtakismós* 'doubling of ι as in *Troiia'*, *labdakismós* 'a defect of pronunciation', *soloikismós* 'incorrectness in the use of language, solecism', as far as rhetorical terminology is concerned. Thus, it is not by chance that Hellenistic culture, as regards Ancient Greek,

[13] The *-izzare* verb forms denote a meaning comparable to 'cause to be X, render X' but also 'practice, make use of X', whereas the *-eggiare* verb forms seem to have specialized in the interpretation 'behave like X': see La Fauci (2006, 2007, 2008).

and the 16th century, as regards Italian, are the two ages in which *-ismo*/*-ista* and *-ismós*/*-istḗs* show high productivity.

References

Aronoff Mark 1976. *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press.

Bauer Laurie 1997. Derivational paradigms. In Booij, Geert and Jaap van Marle (eds.) *Yearbook of Morphology 1996*, 243-256. Dordrecht, Boston & London: Kluwer Academic Publishers.

Braun Peter, Burkhard Schaeder and Johannes Volmert 2003. *Internationalismen II. Studien zur interlingualen Lexikologie und Lexikographie* (Reihe Germanistische Linguistik. Band 246). Tübingen: Niemeyer.

Chantraine Pierre 1933. *La formation des noms en grec ancien*. Paris: Champion.

Ciacci Serena 1999/2000. *I derivati in *mo/a in greco ed in antico indiano: un esempio di nominalizzazione*. Università di Pisa: tesi di laurea.

Dardano Maurizio 2009. *Costruire parole. La morfologia derivativa dell'italiano*. Bologna: Il Mulino.

Devoto Giacomo and Giancarlo Oli 2009. *Il Devoto-Oli 2009. Vocabolario della lingua italiana*. Mondadori: Milano.

Sabatini Francesco and Vittorio Coletti 1997. *DISC compact dizionario italiano Sabatini Coletti.* Ed. in CD-ROM, versione 1.1. Firenze: Giunti Multimedia.

Greimas Algirdas Julien 1969. Eléments d'une grammaire narrative. *L'Homme* 9 (3): 71-92.

Grossmann Maria and Franz Rainer 2004. *La formazione delle parole in italiano*. Tübingen: Niemeyer.

La Fauci Nunzio 2006. Verbi deonomastici e sintassi: sul tipo *catoneggiare*. In D'Achille, Paolo and Enzo Caffarelli (eds.) *Lessicografia e onomastica. Atti delle Giornate internazionali di studio (Università di Roma Tre, 16-17 febbraio 2006)*, Quaderni Internazionali di RIOn 2, 3-15. Roma: Società Editrice Romana.

La Fauci Nunzio 2007. Nomi propri, luoghi comuni. In: *Atti del XXII Congresso internazionale di Scienze onomastiche (Pisa, 28 agosto – 4 settembre 2005)*, vol. I: 605-13. Pisa: ETS.

La Fauci Nunzio 2008. Antonomasie. In: Cresti, Emanuela (ed.) *Prospettive nello studio del lessico italiano. Atti del IX Convegno internazionale della SILFI (Firenze, 14-17 giugno 2006)*, vol. I: 279-83. Firenze: University Press.

Lewis Charlton T. and Charles Short 1879. *A Latin Dictionary*. Oxford: Clarendon Press. Available on-line: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0059

Roché Michel 2007. Logique lexicale et morphologie: la dérivation en *-isme*. In Montermini, Fabio, Gilles Boyé and Nabil Hathout (eds.) *Selected Proceedings of the 5th Décembrettes: Morphology in Toulouse*, 45-58. Somerville, MA: Cascadilla Proceedings Project.

Ronzitti Rosa 2006. *I derivati in *-mo- della lingua vedica (Saṃhitā e Brāhmaṇa)*. Perugia: Guerra Edizioni.

Schwarze Christoph 1995. *Grammatik der italienischen Sprache*. Tübingen: Niemeyer.

Schwyzer Eduard 1953. *Griechische Grammatik. Erster Band*. München: C. H. Beck.

Tronci Liana 2010. Funzioni, forme, categorie. Una nota su costrutti con verbi in -ίζω. In Putzu, Ignazio, Giulio Paulis, Gianfranco Nieddu and Pierluigi Cuzzolin (eds.) *La morfologia del greco tra tipologia e diacronia*, 495-511. Milano: Franco Angeli.

Werner Jürgen 1980. Zum -ismus. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 33: 488-96.

Corpus

Buck Carl D. and Walter Petersen 1984. *A Reverse Index of Greek Nouns and Adjectives*. Hildesheim – Zürich – New York: Georg Olms Verlag.

Liddell Henry G., Robert Scott and Henry S. Jones 1996 [1843]. *Greek-English Lexicon*. Oxford: Clarendon Press.

Zingarelli Nicola 2002. *Lo Zingarelli 2003*. *Vocabolario della lingua italiana.* Bologna: Zanichelli. [Zingarelli 2008 in CD-ROM, 2007].

Romance clitics: cluster formation and allomorphy

Diego Pescarini
University of Padua
diego.pescarini@unipd.it

1. Introduction

The syntactic distribution of clitic elements in Romance depends on parameters like finiteness (for instance, in many languages clitics are postverbal with non-finite tenses), or clause typing (for instance, in many northern Italo-Romance varieties subject clitics are placed postverbally if the sentence is interrogative or exclamative)[*]. Combinations of two or more clitics, on the other hand, are not fully transparent to syntactic principles. In particular, the order of elements within the cluster and the morphological shape of the resulting compound cannot be derived straightforwardly via syntactic computation.

For instance, in a language like French in (1a), object clitics follow the same order of lexical arguments, namely, accusative > dative, while in others, like Italian in (1b), they exhibit the mirror order dative > accusative.

(1) a. Jean le lui donne
Jean it to.him/her gives
'Jean gives it to him/her'

b. Gianni glie lo regala.
Gianni to-him/her.CL it.CL gives
'Gianni gives it to him/her'

In addition, certain combinations are morphologically opaque. In Italian, for instance, when the reflexive clitic *si* in (2a) combines with the impersonal *si*, e.g. (2b), the resulting cluster does not correspond to a transparent sequence of two *si*'s, but the leftmost element of the cluster is replaced by the clitic *ci*, as shown in (2c).

(2) a. Carlo si lava ogni giorno
Carlo himself.CL wash.3.SG everyday
'Carlo washes everyday'

b. La macchina si lava ogni giorno.
The car.OBJ one.CL wash.3.SG everyday
'The car is washed everyday'

c. Ci/*si si lava ogni giorno.
Himself/herself/themselves.CL one.CL wash.3.SG everyday
'You wash everyday'

Grimshaw 1997, 2000, Maiden 2000, Pescarini 2010 among many others have argued that opacity normally follows from a dissimilation principle preventing the adjacency of identical clitic exponents. This principle accounts straightforwardly for systematic gaps like the one exemplified in the following set of Italian sentences, where the locative

[*] I am grateful to Anna Cardinaletti for helpful discussion of this material. This work is part of the FIRB project 'A grammatical survey of Italian dialects: fieldwork, data management, and linguistic analysis' (RBFR08KR5A).

pronoun *ci* can combine with any personal pronoun except for the homophonous 1pl clitic, ex. (3c):

(3) a. mi ci porta Micol.
me.CL there.CL brings Micol
'Micol brings me there'

b. ti ci porta Micol.
you.sg.CL there.CL brings Micol
'Micol brings you there'

c. ci (*ci) porta Micol.
us.CL there.CL brings Micol
'Micol brings us there'.

d. vi ci porta Micol.
you.pl.CL there.CL brings Micol
'Micol brings you there'

In other cases, however, morphological opacity cannot result from a ban on the co-occurrence of identical clitics. This is particularly true in the case of sequences formed by a 3rd person dative followed by a 3rd person accusative clitic. In this context, the exponent which normally expresses the 3p dative clitic (in round brackets in the following examples) must be replaced by another item, which is subject to cross-linguistic variation.

(4) a. Glie /*le lo presto.
Italian
To-her.CL it.CL lend.1.SG
'I lend it to her'

b. Juan se/*le lo compró.
Spanish
Juan to-him/her.CL it.CL bought
'Juan bought it for him/her/them'

c. cə (*i) u da.
Poggio
Imperiale, Southern Italy
to-him.CL it.CL give.3.SG (Manzini & Savoia 2005: 135-138)
'He gives it to him'

It is worth noting that this pattern – which cannot be due to a trivial dissimilation principle – is regularly attested in a consistent set of Romance languages. In particular, this irregularity is found in those languages in which the clusters of 3p clitics display the mirror order (like Italian in (1b)), while these combinations are normally transparent in those languages, like French in (1a), in which the dative clitic occupies the rightmost position in the cluster.

In what follows, I will argue that this correlation between clitic-order phenomena and morphological irregularities can receive a principled explanation if we assume – following Kayne 1994:19-21 and Cardinaletti 2008 – that sequences of clitic items can be either *split* or *cluster* in a stricter sense. According to Kayne 1994: 19-21, two clitics α, β can combine into either a "split" configuration like (5a), where clitics occupy distinct

syntactic projections, or a "cluster" configuration like (5b), in which the leftmost clitic moves from a lower position and left adjoins to the higher clitic.

(5) a [α [β]]

b. [β α [tβ]]

The hypothesis put forth by Cardinaletti 2008 is that morphological irregularities emerge in the latter type of cluster. In what follows I support Cardinaletti's analysis and argue for a finer account of the morphological irregularities displayed in Romance clitic combinations. In section 2 I summarize Cardinaletti's proposal; in section 3 I address a pattern of allomorphy in Italian and, finally, in section 4 I focus on a pattern of suppletivism that targets combinations of 3p clitics in various Romance languages.

2. Cardinaletti 2008: on different types of clitic clusters

Cardinaletti 2008 argues that, in Italian (and, arguably, in Romance), there are several types of clitic clusters. Her classification is based on the following diagnostics:

i. distributional gaps: certain clusters are not allowed in enclisis.
ii. vowel change: in certain clusters of Italian, the vowel of the leftmost clitic is *-e-* instead of the expected *-i-*.

With respect to the former parameter, Italian displays two types of clusters: those which can occur both in enclisis and in proclisis, (6) and (7), vs those which cannot occur in enclisis, (8). Cardinaletti argues that this asymmetry results from the underlying syntactic configuration as she assumes that split sequences are not allowed in enclisis.

Furthermore, unrestricted clusters can be divided into two sub-classes: one formed by clusters with vowel change vs another in which the two clitics are combined transparently.

The resulting taxonomy is as follows:
Type1: unrestricted clusters with vowel change, in (6);
Type2: unrestricted clusters without vowel change, in (7);
Type3: clusters which are allowed only in proclisis, in (8).

(6)

a. Mi ha dato un libro. [he] to-me has given a book	a'. Mi ha dato tre libri. [he] to-me has given three books
b. Me lo ha dato. [he] to-me it has given	b'. Me ne ha dati tre. [he] to-me of-them has given three
c. Pensa di darmelo. [he] thinks to give to-me it	c'. Pensa di darmene tre. [he] thinks to give to-me of-them three
d. Dammelo! give to-me it	d'. Dammene tre! give to-me of-them three

(7) a. Mi / Ti / Vi ci metterà.
[he] me / you:SG / you:PL there will-put

b. Pensa di mettermici / mettertici / mettervici.
[he] thinks to put me / you:SG / you:PL there

(8) a. Non mi/ti/vi/gli/le si parlò con la dovuta attenzione.
not to-me/you:SG/you:PL/him/her IMP spoke with the due attention

b. *Non sembra esser=mi/ti/vi/gli/le=si parlato con la dovuta attenzione.
[it] not seems [to] have=to-me/you:SG/you:PL/him/her=IMP spoken with the due attention

Possibly, the conclusion that Type2 combinations are clusters *stricto sensu* is too strong because, unlike Type1 combinations, they can be split in restructuring contexts, as shown in (9c,d), see also Pescarini (2012).

(9) a. **Ti ci** può portare lui
You there can bring he

b. può portar=**ti**=**ci** lui
can bring=you=there he

c. %**ti** può portar=**ci** lui[1]
You can bring=there he

d. %**ci** può portar=**ti** lui
There can bring=you he
'he can bring you there'

However, leaving Type2 combinations aside, Cardinaletti's claim that Type1 clusters correspond to a single morphosyntactic constituent can shed light on several morphological aspects which will be addressed in the following sections.

3. The *-i/e-* alternation revised

As previously mentioned, Italian shows a context-driven alternation targeting the leftmost clitic of certain clusters: before a 3p accusative clitic (e.g. *lo* 'him', *la* 'her'), or the partitive *ne* ('of it/them'), clitics end with *-e* instead of the expected *-i*. For instance, the clitic *mi* '(to) me' becomes *me*, see (10a), and the 3p m.sg clitic *gli* becomes *glie* /ʎe/, (10b).

(10) a. [me] lo porti [*mi]
to.me it bring.you
'You bring it to me'

b. [ʎe] ne porti due [*ʎi]
to.him of.it bring.you two
'You bring him two of it'

This pattern has received a good deal of attention at least since D'Ovidio (1886:71), who argues that *-e-* is a reflex of the etymological initial vowel of the second clitic *(e)lo* < ILLUM, *(e)ne* < INDE. According to this reconstruction, the derivation of the clusters above is as follows:

(11) a. ME ĬLLUM > M'ĬLLU > me lo 'it/him to me'

[1] The % marker means that the grammaticality of the constructions is subject to variation. In particular, informants usually accept (9c,d) when the sentence is uttered in a plausible context. On the contrary, when I tried to elicit grammatical judgments by means of a written questionnaire, the same construction was often judged ungrammatical.

b. ĬLLI ĬNDE > ILL'ĬNDE > gliene 'it/him to him/her'

This solution provides a clear and elegant account of both the etymology of *-e-* and its synchronic distribution. However, D'Ovidio's analysis – see also Meyer Lübke 1890, 1894, 1901 – has three major drawbacks.

First, if *-e-* was the reflex of ĭ (< ILLE, INDE), the resulting cluster would show a geminate sonorant, as in Florentine etymological geminates are normally maintained. The regular evolution would therefore be as follows:

(12) a. ME ĬLLUM > M'ĬLLU > *mello 'it/him to me'

b. ĬLLI ĬLLUM > ILL'ĬLLUM > *gliello 'it/him to him/her'

Second, as Parodi (1887:189-190) pointed out, in the 13th century, reflexes of ILLE, INDE, occupy the leftmost position of the cluster, as shown by the following examples, and, crucially, do not show traces of either gemination or initial *e-* (e.g. **ella mi*):

(13) a. che […] voi **la mi** concediate (Boccaccio, Filocolo 212)
that […] you.pl it.f to.me grant.subj
'that you grant it to me'

b. io **lo vi** dirò. (Novellino, p.128)
I it ti.you tell.fut
'I will tell it to you'

The fact that the archaic order of these clusters is accusative > dative contradicts D'Ovidio's hypothesis that the linking vowel *-e-* is a reflex of preserved ĭ in cluster-internal position. Rather, *-e-* is an innovation emerging as soon as dative clitics begin to occupy the leftmost position of the cluster (Melander 1929). This means that there is a strong correlation between the syntactic change leading to the order in (13b) and the phonological one determining the change from *-i* to *-e*.

Third, the apocopated allomorph *l'* (< *lo*) resyllabifies by means of the prosthetic vowel *i-*, namely *l'* → *il* (Vanelli 1992/1998, Renzi 1993, Renzi & Vanelli 1993). If the etymological *e-* (< ĭ-) had been still underlying, the insertion of a prosthetic segment like *i-* would have been unnecessary.

The alternative hypothesis is that clitic clusters form an autonomous prosodic constituent, and that this pattern of allomorphy is a side-effect of this exact prosodic configuration. In particular, Cardinaletti claims that Type1 clusters are phonological words[2] (see also Monachesi 1996), although this hypothesis is not supported by any conclusive evidence and, ultimately, it does not account for the *e/i* alternation.

[2] On the contrary, Cardinaletti 2008 argues that type2 clusters e.g. *ci si* form a syntactic constituent, but are not phonological words and, as a consequence, they do not display vowel change. According to Cardinaletti 2008, this observation is supported by the fact that *ci si* is never subject to [s]-sonorization, a PrW-internal process typical of northern speakers:

(i) ci [s]i va *[z]
'there IMP goes'
'One/we go(es) there'

In my opinion, however, this conclusion is quite weak. In fact, it is worth noting that voicing is not allowed when [s] is in morpheme-initial position (Nespor & Vogel 1986:124-129), for instance, after prefixes like *a, anti*, *pre* (in particular when the prefix is transparent):

(ii) a. a-[s]ociale *[z] 'asocial' (northern Italian)
b.anti-[s]oldato *[z] 'anti-soldier'

The only evidence that Type1 clusters form a prosodic constituent is that they are pronounced with a long vowel: [me:lo], [te:lo], but this observation is consistent with an alternative hypothesis, more restrictive than Cardinaletti's, that clusters might correspond to a smaller prosodic unit – the Foot – as argued by Peperkamp 1995, 1996, 1997. The hypothesis that clitic clusters are Feet, rather than phonological words, is consistent with the pattern of several southern Italian dialects like Neapolitan (Bafile 1992, 1994), in which the penultimate pronoun of enclitic clusters is stressed, as shown in (14b), while proclitic clusters are never stressed:

(14) a.pòrta-**tə** na bbirrə (Neapolitan) — single enclitic
bring.to-youself a beer
'bring a beer for youself'

b. pòrta-**té-nnə** assaj — enclitic cluster
bring-to.yourself-of.it a.lot.of
'bring a lot of it (beer) for youself'

(15) a. **tə** pòrtə na bbirrə (Neapolitan) — single proclitic
to.you bring.I a beer
'I bring you a beer'

b. **tə nə** pòrtə assaj — proclitic cluster
to.you of.it bring.I a-lot-of beer
'I bring you a lot of it (beer)'

This pattern is consistent with the prosodic hierarchies in (16), in which the leftmost clitic of the cluster corresponds to the penultimate syllable of the outer PrW and can therefore receive stress[3]. Proclitic clusters, on the contrary, do not occupy a PrW-final position and, as a consequence, cannot be assigned stress.

(16)
```
        *
        *
a. [[(por.ta)]PrW (te-nnə)]PrW          enclitic cluster

          *
          *
b. [(te-nnə) [(por.ta)]PrW]PrW          proclitic cluster
```

However, even if we adopt the plausible hypothesis that clusters are feet, this is not *per se* an explanation of the *-e/i-* alternation. In order to account for this phenomenon, we have to focus on the morpho-phonology of Old Italian, in which, according to Rohlfs 1966:178,

c. pre-[s]elezione *[z] 'preselection'

On the basis of (ii), it seems to me that [s] voicing cannot be considered a test to ascertain whether *ci si* is a PrW or not. Rather, on the basis of the data in (9c,d), I argue that the difference between Type1 (e.g. *glielo*) and Type2 clusters (e.g. *ci si*) is syntactic in nature, rather than phonological: following this analysis, the cluster *ci si* exhibit the vowel *-i-* because it is split in the Syntax (see also Pescarini 2012), while the prosodic status of Type2 clusters, on the contrary, is still an open question, because [s] voicing is not a reliable test.

[3] Neapolitan, unlike Italian, assigns stress to the outer PrW, namely, the PrW which contains the clitic clusters. For a thorough analysis of the data, the interested reader is referred to Peperkamp 1995, 1996, 1997.

final unstressed *-e* optionally became *-i*, giving rise to a series of alternations like *avante* > *avanti* 'before, in front of', *diece* > *dieci* 'ten', *longe* > *lungi* 'far', etc. Such a raising ended up differentiating the morphology of clitic pronouns (e.g. *mi, ti* 'me, you') from the one of their stressed counterparts, which still maintain the vowel *-e* of ME, TE, INCE, SE.

If so, Type1 clusters trigger foot formation and, as a consequence the raising rule is blocked because the leftmost clitics becomes the foot's head. In this position, the vowel *-e-* is therefore expected to surface instead of *-i-*, which is found in weak syllables.

This hypothesis encounters a counterexample as the *-i/e-* alternation targets also the 3p dative clitic *gli* 'to him/her' < ILLI, although in this case *-e-* cannot be a reflex of the original ending. Since the original ending of the dative clitic is *-i*, we would expect **glilo*, **gline*, instead of *glielo*, *gliene* (pron. /ʎelo/, /ʎene/, cf. (10b). Crucially, Old Italian is consistent with this prediction, as in the earliest documents the 3p dative clitic exhibits also the linking vowel *-i-*, as shown in (17), see also Cardinaletti 2010:444ff. Only in a later stage, the linking vowel of these clusters has become *-e-*, possibly in analogy with the morphology of the other clusters.

(17) a. che **gli le** demo p(er) una inpossta (LibrAmmBIR[4])
that to.him them gave.1pl for a tax
'that we gave them to him for a tax'

b. ché **gli ne** potrebbe troppo di mal seguire (Boccaccio, Dec. III, 3, p. 197)
because to.him of.it could too.much of bad(luck) follow
'because it could cause him too much misfortune'

In conclusion, the data above are consistent with Cardinaletti's view that Type1 clusters correspond to a single autonomous constituent both in the Syntax and in the Phonology. Following Kayne 1994, we can argue that this type of combinations is due to the incorporation of the dative clitic onto the accusative one and, following Peperkamp 1995, 1996, 1997, we can argue that Type1 clusters give rise to foot formation which in turn trigger the *-e/i-* alternation as a side effect of secondary stress. The other clitic combinations, on the contrary, are split in the syntax, do not undergo foot formation and, ultimately, were subject to vowel raising.

3.1. An aside on the morphology of P+D combinations

The above analysis, which relies on a desirable Syntax/Prosody isomorphism, can be weakened if other types of clitic clusters are taken into consideration. Cardinaletti 2008, in fact, notices in fact that Type1 clusters "display the same vowel that is found in the combinations of preposition and determiner such as *in* + *il* > *nel* 'in the' or *di* + *il* > *del* 'of the'".

This remark, however, is highly undesirable as it ends up contradicting Cardinaletti's own analysis. In fact, P+D sequences cannot be clusters *strictu sensu* (*à la* Kayne 1994) as the linear order P > D cannot be due to movement of P° past D°. Hence, if we want to maintain Cardinaletti's analysis of clitic clusters, we have to demonstrate that the underlying structure of Type1 clusters differs neatly from the one of P+D sequences.

Italian provides evidence in favour of such a distinction. Recall that the *-e-* of Type1 clusters cannot be considered the reflex of ĭ (< ILLE, INDE), otherwise the resulting cluster would show a geminate sonorant:

[4] Libro d'amministrazione dell'eredità di Baldovino Iacopi Riccomanni (La prosa italiana delle origini: I, Testi toscani di carattere pratico, a cura di Arrigo Castellani, Bologna, Pàtron, 1982, pp. 429-64 [testo pp. 433-64].)

(18) a. ME ĬLLUM > M'ĬLLU > *mello 'it/him to me'

b. ĬLLI ĬLLUM > ILL'ĬLLUM > *gliello 'it/him to him/her'

It is worth noting that P+D sequences, unlike pronominal clusters, are geminated and, following Formentin (1996), at least in some cases, this gemination must be a regular reflex of the *univerbation* of two separate though adjacent heads:

(19) a. DE ĬLLUM > D'ĬLLU > dello 'of the'

b. IN ĬLLUM > (I)N'ĬLLUM > nello 'in the'

On the basis of the asymmetry between (18) and (19), we can conclude that the *-e-* of Type1 clusters can be considered a clue of incorporation, while in P+D sequences *-e-* is a regular reflex of ĭ.

4. Suppletivism

In many Romance varieties, the 3p dative clitic is replaced by a suppletive exponent when it is clustered with another clitic element. As shown in (4), this normally happens in Type1 clusters. We can distinguish at least three main patterns of substitution on the basis of the etymology of the replacing item:
i. spurious *se* patterns, attested in Ibero-Romance and Campidanese Sardinian: in true clusters the etymological 3p dative *le/li* is replaced by the 3p reflexive element (with a non-reflexive interpretation).

(20) a. **ḍi** pottu unu libru. (Sarroch, Campidanese Sard.)
to-him.CL bring.1.SG a book
'I bring him a book'

b. **si**/***ḍi** ḍu pottu.
to-him.CL it.CL bring.1.SG
'I bring it to him'

ii. spurious locative patterns, attested in many Italo-Romance dialects, Logudorese Sardinian and Catalan (also in colloquial French, cf. (22)): in true clusters the etymological 3dat *le/li* is replaced by the locative clitic *ci/bi/hi/y.*

(21) bi/***li** l'appo datu (Log. Sardinian, Jones 1993:220)
to.him/her/them it have.1.sg given
'I gave it to him/her/them'

(22) %Donne-z-**y**-**en** (colloquial French)
Give-[z]-there-of.it
'Give him some of it!'

iii. spurious *ne* patterns (several southern Italian dialects): in true clusters the etymological 3dat *le/li* is replaced by the partitive element deriving from Lat. INDE.

(23) a.**i** da kkuistə (Rocca Imperiale, CS, Southern Italy
to.him/her/them gives this Manzini & Savoia 2005: 291)
'He/she gives this to him/her/them'

b. **n**/*__i u__ da
to him/her/them it gives
'He/she gives it to him/her/them'

Following Cardinaletti (2008), this phenomenon regards the 3p dative clitic because it is *bimorphemic* (Kayne 2000), i.e., it is formed by a root √*l*- followed by an agreement marker. Under this view, the restriction above can be reformulated as follows: bimorphemic clitics cannot occour in the left position of Type1 clusters (see also Cardinaletti 2010)

This generalization is supported by several Sardinian dialects, which exhibit this kind of phenomena only in clusters with the mirror order, i. e., in clusters in which the dative clitic occupies the ledtmost position. Consider, for instance, the following pattern from a number of Sardinian dialects (Manzini & Savoia 2005 vol. II:317.321). Crucially, the etymological form of the 3p dative clitic is *li* (24a), which occours when it follows another clitic element as in (24b), while when it occupies the leftmost position of the cluster, in (24c), it must be replaced by the 'spurious' exponent *bi*.

(24) a.**li** dana kustu (Ittiri SS, Padria SS, Luras OT,
to him/her gives this Siniscola, NU Galtellì NU, Bosa OR)
'He/she gives this to him/her'

b. **bi**/*__li lu__ dana
to him/her it gives
'He/she gives it to him/her'

c. **nde li/*bi** dana
of.them to him/her gives
'He/she gives some of them to him/her'

In conclusion, all these context-determined phenomena seem to result from the same, general operation: the substitution of a bimorphemic clitic – √*l*+Agr 'to him/her' – with a monomorphemic one. Monomorphemic exponents (like 1/2, 3p reflexive clitics) are always free to combine with other clitic forms without giving rise to suppletive patterns even if they occupy the leftmost position of the cluster.

On the basis of this tentative generalization, we can improve Cardinaletti's analysis in proposing that the operation responsible for the mirror order of Type1 clusters is a process of *root incorporation*. In a nutshell, when Type1 clusters are built, only the root of the dative clitic incorporates and this gives rise to the above cases of suppletivism.

First of all, I will assume that bimorphemic clitics are formed by a root expressing Person features (say, √{P}) followed by an agreement marker expressing Number and Gender:

(25) $[_{D^\circ}$ √{P} $[_{Agr}$ {G,N}]]

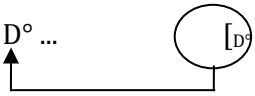
On the basis of these features, vocabulary items are inserted after syntactic operations have taken place (Halle & Marantz 1993):

(26) $[_{D^\circ}$ √{P} $[_{Agr}$ {G,N}]]
| |
/l/ /i/ → /li/ 'to him/her' (Sardinian)

When a bimorphemic clitic like *li* is clustered with an element referencing the higher internal argument, the root of the dative clitic undergoes incorporation. The stranded

agreement of the dative clitic normally remain unpronounced (but see below for some exceptions):

(27) [D° ... [$_{D°}$ √{P} [$_{Agr}$ {G,N}



As a consequence of this process, the (sub-)constituent √{P} cannot trigger the insertion of the item *l-* as in this position it cannot combine with a proper Agr marker.

(28) [√{P} + D° ... [$_{D°}$ t$_{√\{P\}}$ [$_{Agr}$ {G,N}
| |
**l* *lo*
to.him/her it/him

As the agreement material of the dative clitic is stranded in a discontinuous position, a monomorphemic element must fill the head hosting √{P}. 1/2 clitics cannot be inserted instead of *l*, as they cannot match the {P} specification of the root. As a last resort, a dummy clitic – subject to cross-linguistic variation (see Pescarini 2010) – is inserted, as schematized below:

(29) [√{P} + D° ... [$_{D°}$ t$_{√\{P\}}$ [$_{Agr}$ {G,N}
| |
bi *lu* Logudorese Sardinian
si *lu* Campidanese Sardinian
n *u* Rocca Imperiale, etc.

The hypothesis above is supported by the phenomenon of *parasitic plural* (Halle & Harris 2005, Kayne 2010, Manzini & Savoia 2009), which is attested in languages[5] in which Number is expressed by the plural suffix *-s*. In such languages, 3p plural clitics exhibit a trimorphemic exponent, as schematised below:

(30) [$_{D°}$ √{P} [$_{Agr}$ {G} [$_{Agr}$ {N}
| | |
/l/ /o/ /s/ = Sp. *los* 'them (m)'

The same analysis holds for the 3p dative clitic *les/lis* 'to them'. Interestingly, when the plural dative occurs in true clusters (for instance, before the 3^{A} clitic *lo* 'it/him'), it is replaced as usual by a dummy exponent (e.g. *bi*), but, crucially, its plural feature can be expressed by the morpheme *-s*, which in this case attaches to the right of the whole cluster as shown in (32). The resulting cluster does not mean 'them to him/her/them', but 'it to them':

(31) nara=**bi**=**lo**-**s** (Logudorese Sard., from Jones 1993)
tell=there=it-pl
'tell it to them'

In the light of the above analysis, the position of the plural suffix *-s* can be accounted for in terms of stranding of the agreement features of the dative pronoun, whose root has incorporated into the accusative clitic *lo*:

[5] The phenomenon is mainly attested in South American and Sardinian varieties. Parasitic plural is pervasive in the Catalan dialect spoken in Barcelona (Bonet 1991), but traces of parasitic plural are to befound also in Old French (Giampaolo Salvi's p.c. reported in Benincà & Poletto 2005: fn. 14)

(32) [√{P} + D° ... [$_{D°}$ t$_{√\{P\}}$ [$_{Agr}$ {G} [Agr {N}
| | |
bi *lo* *-s*

Jones 1993, focusing on Logodurese Sardinian, reports also cases of parasitic gender, i.e. cases in which the rightmost TV vowels expresses the gender of the dative clitic, rather than that of the accusative one:

(33) nara=**bi**=**l-a**-**s** (Logudorese Sard., from Jones 1993)
tell=there=3p-f-pl
'tell it to them.f'

Under the hypothesis above, the analysis of (34) is as follows:

(34) [√{P} + D° ... [$_{D°}$ t$_{√\{P\}}$ [$_{Agr}$ {G} [Agr {N}
| | | |
bi *l* *-a-* *-s*

5. Conclusions

In this contribution I presented a refined version of cardinaletti's 2008 analysis of Italian clitic clusters. I agreed with Cardinaletti in claiming that morphological irregularities follow from a tight isomorphism between the syntactic structure and the morpho-phonologic realization of certain clusters. In particular, I supported her hypothesis that Italian exhibit a peculiar subclass of clitic combinations which are due to the incorporation of the dative clitic onto the accusative one (*à la* Kayne 1994).

However, I improved or departed from Cardinaletti's proposal with respect to the following points:

i. I challenged the conclusion that Type2 clusters – those containing the locative *ci* – are clusters *stricto sensu* on the basis of evidence from restructuring constructions. Crucially, Type2 combinations, unlike Type1, can split in restructuring contexts, at least in a substandard/oral register.
ii. I rejected the idea that clusters are phonologic words as the observed lengthening of the vowel is compatible with the much more likely analysis that clusters are Feet (Peperkamp 1995, 1996, 1997).
iii. I improved the analysis of the *-e/i-* alternation by suggesting that it is due to the blocking of a historical raising process.
iv. I showed that the *-e-* we found in Type1 clusters and the *-e-* found in P+D sequences have a different nature.
v. I argued that the process giving rise to the suppletivism of Type1 clusters is due to a process of root incorporation. This provides a straightforward account of the parasitic plural pattern.

References

Bafile Laura 1992. *Fonologia prosodica e teoria metrica: accento, cliticizzazione e innalzamento vocalico in napoletano*. Università di Firenze: doctoral dissertation.

Bafile Laura 1994. La riassegnazione postlessicale dell'accento del napoletano. *Quaderni del Dipartimento di Linguistica – Università di Firenze* 5: 1-23.

Benincà Paola and Cecilia Poletto 2005. On some descriptive generalizations in Romance. In R. Kayne and G. Cinque (eds.) *The Oxford handbook of comparative syntax*, 221-258. New York – Oxford: Oxford University Press.

Bonet Eulalia 1991. *Morphology after syntax: Pronominal clitics in Romance*. MIT: doctoral dissertation.

Cardinaletti Anna 2008. On different types of clitic clusters. In Cécile De Cat and Katherine Demuth (eds.) *The Bantu–Romance Connection*, 41–82. Amsterdam: Benjamins.

Cardinaletti Anna 2011. Morphologically Complex Clitic Pronouns and Spurious *se* Once Again. In Vicenç Torrens, Linda Escobar, Anna Gavarró, Juncal Gutiérrez, *Movement and Clitics: Adult and Child Grammar*, Newcastle: Cambridge Scholars Publishing, 238-259

Cardinaletti, Anna 2010. Il pronome personale obliquo in G. Salvi and L. Renzi, *Grammatica dell'italiano antico*, vol. 1. Bologna: il Mulino, 414-450

D'Ovidio, Francesco 1886. Ricerche sui pronomi personali e possessivi neolatini. *Archivio Glottologico Italiano* 9: 25-101.

Formentin Vittorio 1996. Alcune considerazioni e un'ipotesi sull'articolo determinativo in area romanza. In L. Lugnani, M. Santagata, A. Stussi (eds.) *Studi offerti a Luigi Blasucci dai colleghi e dagli allievi pisani*, 257-272. Lucca: Maria Pacini Fazzi editore.

Grimshaw Jane 1997. The Best Clitic: constraint interaction in morphosyntax. In L. Haegeman (ed.) *Elements of Grammar*. *Handbook of Generative Syntax*, 169-196. Dordrecht: Kluwer.

Grimshaw Jane 2000. Optimal Clitic Positions and the Lexicon in Romance Clitic Systems. In G. Legendre, J. Grimshaw and S. Vikner (eds.) *Optimality Theoretic Syntax*, 205-240. Cambridge (Mass.), MIT Press.

Halle Morris and James Harris 2005. Unexpected Plural Inflections in Spanish: Reduplication and Methatesis, *Linguistic Inquiry* 36.2: 195-222.

Halle Morris and Alec Marantz 1993. Distributed Morphology and the Pieces of Inflection. In Kenneth Hale and S. Jay Keyser (eds.) *The View from Building 20*, 111-176.Cambridge (Mass): MIT Press.

Jones Michael Allan 1993. *Sardinian Syntax*. London: Routledge.

Kayne Richard 1994. *The Antisymmetry of Syntax*. Cambridge (Mass.): MIT Press.

Kayne Richard 2000. *Parameters and Universals*. Oxford: Oxford University Press.

Kayne Richard 2010. Toward a Syntactic Reinterpretation of Harris and Halle (2005). In R. Bok-Bennema, B. Kampers-Manhe and B. Hollebrandse (eds.) *Romance Languages and Linguistic Theory 2008, Selected papers from 'Going Romance' Groningen 2008*, 145-170. Amsterdam: John Benjamins, Amsterdam.

Maiden Martin 2000. Phonological Dissimilation and Clitic Morphology in Italo-Romance. In L. Repetti (ed.) *Phonological Theory and the Dialects of Italy*, 137-168. Amsterdam: Benjamins.

Manzini Maria Rita and Leonardo Savoia 2005. *I dialetti italiani e romance. Morfosintassi generativa*. Alessandria: Edizioni Dell'Orso.

Manzini Maria Rita and Leonardo Savoia 2009. Morphology dissolves into syntax: Infixation and Doubling in Romance languages. *Annali Online di Ferrara – Lettere*: 1-2009: 1/28.

Melander Johan 1929. L'origine de l'Italien me ne, me lo, te la, etc.. *Studia Neophilologica* 2: 169-203.

Monachesi Paola 1996. On the representation of Italian clitics. In U. Kleinhenz (ed.) *Interfaces in Phonology* (Studia Grammatica 41), 83-101. Berlin: Akademie Verlag.

Nespor Marina and Irene Vogel 1986. *Prosodic phonology*. Dordrecht: Foris Publications.

Parodi Ernesto Giacomo 1887. Illustrazioni linguistiche ai suddetti frammenti. *Giornale storco della letteratura italiana* 10: 178-196.

Peperkamp Sharon 1995. Enclitic stress in Romance. In A. Dainora, R. Hemphill, B. Luka, B. Need and S. Pargman (eds.) *Papers from the 31st Regional Meeting of the Chicago Linguistic Society, Volume 2: The Parasession on Clitics*, 234-249. Chicago: Chicago Linguistic Society.

Peperkamp Sharon 1996. On the Prosodic Representation of Clitics. In U. Kleinhenz (ed.) *Interfaces in Phonology* (Studia Grammatica 41), 102-127. Berlin: Akademie Verlag.
Peperkamp Sharon 1997. *Prosodic Words* (HIL dissertations 34). The Hague: Holland Academic Graphics.
Pescarini Diego 2010. *Elsewhere* in Romance: Evidence from Clitic Clusters. *Linguistic Inquiry* 41.3: 427-444.
Pescarini Diego 2012. Prosodic restructuring and morphological opacity. The evolution of Italo-Romance clitic clusters. University of Padua: ms. [lingBuzz/001477]
Renzi Lorenzo 1993. Da dove viene l'articolo *il*. In J. Kramer and G.A. Plangg (eds.) *Verbum Romanicum, Festschrift für Maria Iliescu*, 215-30. Hamburg: Buske.
Renzi Lorenzo and Laura Vanelli 1993. Storia e struttura dell'articolo italiano *il*. In *Actes du XX^e Congrès International de Linguistique et Philologie Romanes*, III, 293-305. Tübingen, Francke Verlag.
Rohlfs Gerhard 1966. *Grammatica storica della lingua italiana e dei suoi dialetti*. *Vol. I: Fonetica*. Turin: Einaudi.
Vanelli Laura 1992. Da "lo" a "il": storia dell'articolo definito maschile singolare in italiano e nei dialetti settentrionali. *Rivista Italiana di Dialettologia* 16:29-66. Republished as Ch. 8 of Vanelli 1998, *I dialetti italiani settentrionali nel panorama romanzo*, Roma: Bulzoni.

Gender selection and syntactic constructions: the case of Dutch double gender nouns

Chiara Semplicini
University of Perugia
semplicini@yahoo.it

1. Introduction

The grammatical category of gender is particularly interesting given its halfway status between morphology and syntax. In fact, gender is not only an inherent lexical property of nouns, that can be either morphologically marked (overt gender) or not (covert gender), but also a contextual property displayed through syntactic agreement (Hockett 1958; Corbett 1991; 2006). Accordingly, gender has two main functions: it allows nominal classification and ensures textual cohesion by means of cross-reference relations between the gender controller (noun) and gender targets (determiners and pronouns).

Nevertheless, gender is different from other morphological categories like case and number, because it allows for no choice (Leiss 1999), that is a noun can be in its singular or plural form, or in the accusative or nominative case, but it cannot be masculine in a context and feminine in another. In other words, the gender of a noun is given once for all.

This assumption is partially invalidated by 'double gender nouns' (DGNs), i.e. nouns that can take the agreements of more than one consistent pattern without any variation in meaning (Corbett 1991). A case in point is that of Dutch *aperitief* 'apéritif' in (1):

(1) a. *De aperitief heeft een lichte porttoets en bevat een alcoholpercentage van 15%.*
DET-C apéritif has a light taste of port wine and contains an alcohol percentage of 15%.[1]
'The apéritif has a light taste of port wine and contains an alcohol percentage of 15%'

b. *Het aperitief is de gelegenheid om een barbeque te organiseren.*[2]
DET-N apéritif is the occasion for a barbecue to organize
'The apéritif is the occasion to organize a barbecue'

The noun *aperitief* can take both common (1a) and neuter (1b) agreement without any apparent change in meaning. Nowadays Dutch nouns are commonly divided into *de*-words and *het*-words, while nouns like *aperitief* are listed in grammars and vocabularies as *de/het*-nouns: a definition that presumes the instability of gender to be 'inherent' for DGNs that, as a consequence, can trigger any kind of agreement (Booij 2002). The main problem is that Dutch DGNs cannot be treated as homonymous words, like *de ball* 'ball' and *het ball* 'dance'– where different genders associate with different meanings – and previous accounts on the subject generally assume dialectal variation to be responsible for these fluctuations (Pauwels 1938; De Vogelaer and De Sutter 2010). Accordingly, gender uncertainty for *de/het*-nouns is currently explained as an instance of lexical gender instability, without any further investigation.

[1] Available at: http://toerisme.hoogstraten.be/Aardbeienwijn-likeur-aperitief.html
[2] Available at: http://koersoepfrankrijk.nl/het-aperitief

The aim of this paper is to provide another explanation for Dutch DGNs treating the phenomenon from a different perspective: the analysis of these nouns as a matter of 'motivated' gender agreement. The paper is divided as follows: first of all the contemporary Dutch gender system will be described, then previous studies on Dutch pronominal gender and its ongoing recategorization will be briefly sketched before turning to a series of examples that allow explanation for these 'apparent' unmotivated fluctuations.

2. Previous studies on Dutch pronominal gender: the 'reinvention' of pronouns

Nowadays the Dutch gender system is described as a mismatched one (Booij and Audring 2009) given the paradigmatic unbalance between the number of gender values for controllers and targets. In fact, Dutch nouns divide into two classes: common *de*-nouns (resulting from the coalescence of original masculine and feminine nouns) and neuter *het*-nouns. The same distinction counts for all gender targets with the exception of personal pronouns that still discriminate between masculine (*hij* 'he'), feminine (*zij/ze* 'she') and neuter (*het* 'it'). The absence of a common gender pronoun – as it is the case in Scandinavian languages (Braunmüller 1999) – increases the degree of uncertainty to the extent that, in principle, all common gender nouns can trigger either the masculine or the feminine personal pronoun (Table 1).

AGREEMENT TARGET	**DETERMINER**	**ADJECTIVE**	**DEMONSTRATIVE PRONOUN**	**RELATIVE PRONOUN**	**PERSONAL PRONOUN**
genders	**common** de stoel 'the chair'	**common** de grot-*e* stoel 'the big chair' een grot-*e* stoel 'a big chair'	**common** deze/die 'this/that'	**common** die 'that'	**masculine** hij 'he'
	common de stoel 'the chair'	**common** de grot-*e* stoel 'the big chair' een grot-*e* stoel 'a big chair'	**common** deze/die 'this/that'	**common** die 'that'	**feminine** zij/ze 'she'
	neuter het huis 'the house'	**neuter** het grot-*e* huis 'the big house' een groot huis 'a big huis'	**neuter** dit/dat 'this/that'	**neuter** dat 'that'	**neuter** het 'it'

Table 1: The mismatch between Dutch gender controllers and gender targets

Actually, recent studies on Dutch pronominal gender reveal that the situation is much more complex than one could assume just by having a look at the table above: not only *de*-nouns can trigger both masculine and feminine personal pronouns, but they also agree with neuter pronouns, and neuter nouns can associate with common gender

targets as well, at least in the spoken language. Moreover, these agreement patterns are not random but demonstrate the existence of a tendency towards common gender for animate and inanimate highly individuated referents and the neuter for inanimate, low individuated entities (Fletcher 1987; Romijn 1996; De Vries 2001; De Vogelaer 2006; Audring 2006; 2009; De Vogelaer and De Vos 2011; De Vos 2009). This state of affairs presumes a progressive 'resemanticization' (Wurzel 1986) of Dutch gender towards a conceptual system based on different degrees of individuation (Audring 2009). In other words, the selection of pronouns seems to depend no longer on the lexical gender of the noun, but on the individuation properties of the referent (2):

(2) a. [count N] *artikel* 'article' (*het*-word)
dus ik zou 'm zo af kunnen sluiten
so I could PERS.PRO-M so off can close
'so I could just finish it like this'

b. [mass N] *olijfolie* 'olive oil' (*de*-word)
hoe 't geconserveerd wordt
how PERS.PRO-N conserved becomes
'how it is preserved'

As shown in (2), revised from Audring (2009: 88-95), the neuter noun *artikel* agrees with a masculine pronoun (*'m*), while the common gender noun *olijfolie* triggers the neuter *het*. Even though these agreement patterns are syntactically wrong, they make sense from a cognitive perspective: the noun *artikel* makes reference to a bounded (highly individuated) entity, whereas *olijfolie* is a mass noun referring to a substance, i.e. low individuated referent. Accordingly, Dutch pronouns are claimed to be undergoing a recategorization which follows the Individuation Hierarchy (Sasse 1993; Siemund 2008): the selection of pronouns depends on the conceptualization of the entity the noun refers to (masculine/(feminine) for more individuated referents and neuter for less individuated ones) rather than on the lexical gender of the controller (Figure 1).

Person (male) > Animal > Bounded Object > Unbounded object/Specific Mass > Unspecific Mass Person (female)

[+ individuated] [-individuated]

masculine/feminine > masculine > neuter > neuter

Figure 1: Dutch gender and the Individuation Hierarchy (revised from Audring 2009)

Besides, this change is in accordance with the predictions of Corbett's Agreement Hierarchy (1979; 1991; 2006): the probability to find semantic rather than syntactic agreement is higher for NP-external targets (personal, demonstrative and relative pronouns) than for NP-internal targets (articles and adjectives) and increases monotonically moving rightwards along the hierarchy (Figure 2). In other words, any change towards a (more) semantic gender system begins on the right pole of the scale (personal pronouns) exactly as it has been demonstrated to be happening in Dutch.[3]

[3] Similar recategorization phenomena have been proved to be at work in other Indo-European varieties: Danish (Braunmüller 1999), Norwegian (Enger 2004) and Swedish (Joseffson 2006); southern English varieties (Siemund 2008); northwest Spanish varieties (Lüdtke 2001; Fernández Ordoñez 2009) and central Italian dialects (Haase 1999).

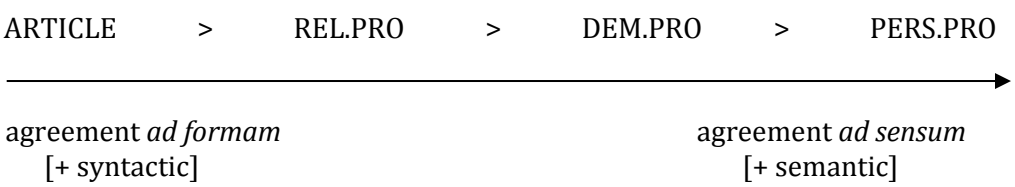


Figure 2: Corbett's Agreement Hierarchy and gender agreement

Nevertheless, previous research on Dutch gender claims that nominal targets (articles and adjectives) are not undergoing the same change (Audring 2009; Booij and Audring 2009): on the left of the hierarchy lexical gender continues to rule.[4] Accordingly, Dutch DGNs are particularly interesting to investigate, given that their lexical instability, if not regarded as a matter of dialect, could be driven by the same dynamics leading to the restructuring of pronominal gender along semantic pathways.

3. Dutch double gender nouns as a matter of gender agreement

Previous accounts on Dutch DGNs assume their gender instability to be lexically determined. Nevertheless, considering this instability as a matter of agreement may allow explanation for gender shifts exactly as it is the case for pronominal gender.

On the one hand gender is fixed and lexically determined for each noun (lexical gender) but on the other hand it is selected contextually each time a noun is used to refer to a specific entity the speaker has in mind and he wants to make understandable to the hearer, that is gender – as a property of NPs and pronominal targets – is a referential feature (Dahl 1999) and may serve the function of nominal 'perspectivization' (Weber 1999). In other words, if lexical gender is a fixed property of nouns that allows assignment to a certain class, gender targets (pronouns and determiners) reflect referential gender, that is gender is not pre-determined but triggered in context.

Provided that Dutch nouns are associated with a certain gender depending on the definite article they take,[5] talking about lexical gender for *de/het*-words does not makes any sense. As a matter of fact, for DGNs the distinction between gender assignment and gender agreement (Corbett 1991) is tautological. Once grammatical gender has become covert – as it is the case of contemporary Dutch - speakers cannot rely on the morphological shape of the noun to decide which gender class it belongs but may rely on semantic and pragmatic agreement patterns to structure their utterances, using gender as a cognitive cue to convey specific conceptualizations, disregarding its original grammatical function.

Actually, the analysis of Dutch DGNs in context confirms that different genders correlate with different morphosyntactic configurations: in particular the mass syntax tends to privilege neuter gender, while the count syntax prefers common gender. This let us suppose that referential gender interacts with both syntactic and semantic/pragmatic factors, not only at the pronominal level, but also in the nominal domain, a state of affairs which may indicate a step forward in the transition of Dutch gender as a whole.

Provided that the meaning of a noun can be coerced in context (Krifka 1995; Talmy 2000; Wisniewski 2009) and syntactic prominence is one of the primary factors in determining the salience of entities (Rose 2011), I assume that gender selection in definite NPs could be driven by the need to mark different degrees of

[4] Semantic agreement in the nominal domain is predicted to be very rare given the proximity of controller and targets (Corbett 1991).

[5] Dutch nouns do not display any longer morphological markers for gender: apart from some derivational suffixes that associate with a specific gender value (for instance nouns ending with the suffix *-heid* are feminine and all diminutives are neuter), gender is covert (Booij 2002).

(semantic/pragmatic) individuation, especially in cases where grammatical gender is impaired, i.e. DGNs.

4. Dutch double gender nouns and Individuation

DGNs have been generally considered as a niche phenomenon in the history of Dutch language, to the extent that even the number of items involved is uncertain and classifications are neither uniform nor coherent. For instance, simply considering different sources, i.e. dictionaries, grammars or lexical databases, the number of nouns listed as *de/het*-words changes a lot (Semplicini 2012a); moreover, some nouns are classified as either stable or unstable depending on the source (the noun *gordijn* 'curtain' is listed as either a *de/het*-word or a *het*-word according to different accounts). This terminological confusion arises the problem of determining the size of the phenomenon and presumes that Dutch DGNs may be less irrelevant than generally claimed.

A synchronic investigation of these nouns (Semplicini 2012a; Semplicini 2012b) has shown that they are characterized by three main properties: they refer to inanimate entities (objects, substances and abstracts), they are unstable for gender and share a high semantic affinity and polysemy. DGNs form not only synonymic pairs (*omslag/kaft* 'cover'; *afval/vuilnis* 'rubbish') but also complex semantic networks (*draad* 'wire, thread, line' - *koord* 'rope, line, string'- *kluwen* 'ball of string'; *fiber* 'fiber' - *flanel* 'flannel' - *katoen* 'cotton'; *hars* 'resin' - *gom* 'gum'- *plamuur/mastiek* 'mastic' etc.). Accordingly, *de/het*-nouns seem to constitute a particular area of the Dutch lexicon, a network of concepts which are synonymous at different levels and that generally allow different interpretations: substance/contained substance (*hars/plamuur* 'resin, mastik'); object/surface (*omslag/kaft* 'cover'); substance/object (*katoen/draad* 'cotton, yarn') etc. In other words, these nouns mainly refer to entities that can be conceptualized as either individuated or not.

From a cognitive perspective conceptualization represents a mid-way level between ontology and grammar: the ontological distinction individual vs. non individual (Quine 1960) is grammatically reflected by the count vs. mass syntax. Different perspectives on an entity are conveyed at the morphosyntactic level by different constructions (Krifka 1995): for instance the noun *beer*, that is a prototypical mass noun (3a), can be coerced into a countable one simply by choosing a marked syntactic configuration (3b).

(3) a. I drink beer (prototypical meaning)
b. I drink a beer (coerced meaning)

In other words, the meaning of a noun is not fixed but flexible and open to different interpretations (Talmy 2000). As a matter of fact, syntax is deeply affected by conceptualizations (or world knowledge): combinatorial possibilities and syntactic constructions are affected by the ways in which we conceive of their meanings (Parafragou 2005). Conceptualizations are variable and flexible because they strictly depend on the contingent communicative needs of the cognitive agent (Wisniewski et al. 2003; Wisniewski 2009): when we are referring to a certain entity we conceptualize it in different manners according to different 'scopes of predication' (Wisniewski 2009) that result in different syntactic constructions, i.e. for instance mass vs. count syntax. Not only words but also constructions have meaning (Langacker 1991) to the extent that if a lexical item is semantically incompatible with its syntactic context the meaning of the noun conforms to the meaning of the structure in which it is embedded (Michaelis 2005). For instance, recalling the example in (1) the noun *aperitief* conceptualizes two different referents: the apéritif as an occasion (abstract, i.e. non individuated) in (1b) and the apéritif as a drink (concrete, i.e. individuated) in (1a). Even though different genders for

aperitief may be a matter of chance, the analysis of DGNs in context has revealed that these shifts are less random than generally supposed.

5. Referential gender and syntactic constructions: some interesting cases

The investigation of Dutch DGNs in context shows that some syntactic constructions are more likely to correlate with a certain gender value and that within these configurations synonymic words tend to select the same definite article, i.e. the same gender: the *het*-article is the most preferred for non individuated interpretations, while the common gender article is the most suited for high individuated readings. Consider the case of *vuilnis* 'rubbish' in (4) and *appelmoes* 'apple puree' in (5):

(4) a. Winkelen tussen *het* *vuilnis* in Utrecht
Shops among DET-N rubbish in Utrecht
'Shops among the rubbish in Utrecht'[6]

b. Als vanzelfsprekend nam ik *de* *vuilnis* even mee
Obviously bring I DET-C rubbish always (with) me

naar beneden (ik woon in een flat), om 't beneden in de vuilcontainer te gooien.
downstairs (I live in a flat) to it outside in the rubbish container to throw

'Obviously I bring the rubbish downstairs (I live in a flat), to throw it outside into the rubbish container'[7]

(5) a. Verdeel *het* *appelmoes* over de plakken bladerdeeg en
Divide DET-N apple mousse over the pieces puff pastry and

klap ze dicht als een portefeuille.
close them as a wallet

'Divide the apple mousse on the pieces of puff pastry and close them as a wallet'[8]

b. *De* *appelmoes* van Koeleman wordt gemist. Op diverse internetafora
DET-C apple mousse of Koeleman is missed. On different internet forums

blijkt dat consumenten het betreuren dat nieuwe eigenaar Nederlands Glorie
seems that customers it regret that new owner Nederlands Glorie

de appelmoes uit het schap heeft gehaald.
DET-C apple mousse from the shelf has removed.

[6] Available at: nos.nl/artikel/157594-winkelen-tussen-het-vuilnis-in-utrecht.html
[7] Corpus of Spoken Dutch: Session fn008020
[8] Available at: eten-en-drinken.infonu.nl/recepten/1772-heerlijke-appelflappen.html

'The apple mousse of Koeleman won't exist any longer. On different internet forums customers are regretting that the new owner Nederlands Glorie has removed the apple mousse from the market shelf'[9]

As shown in these examples the neuter is triggered for unbounded conceptualizations (an indistinct amount of rubbish on the streets of Utrecht and the apple mousse to be spread on the puff pastry), while common gender is generally used for individuated referents: the rubbish sack and the apple mousse (pot) produced by Koeleman, respectively.
Similar tendencies are shown by DGNs sharing the same head like *haarlak* 'hairspray' and *nagellak* 'nail lacquer' and the specific syntactic configurations they enter in (6) and (7):

(6) a. In de winkel zag ik *deze* *haarlak* liggen
in the shop saw I DEM-C hairspray lay
'I saw this hairspray in the shop'[10]

b. Laat *het* *haarlak* even opdrogen
Let DET-N hairspray then dry up
'Let the hairspray dry up'[11]

(7) a. *De* *nagellak* van OPI wordt door tal van celebs
DET-C nail lacquer by OPI is by many celebrities

op handen gedragen.
on hands worn
'The nail lacquer by OPI is used by many celebrities'[12]

b. Het verwijderen van *het* *nagellak* werkt goed.[13]
The removing of DET-N nail lacquer works good
'The removing the nail lacquer is easy'

Even in these cases the neuter correlates with a low degree of individuation both for *haarlak* and *nagellak* (a substance that dries up or that can be easily removed), while common gender is triggered when the referent is conceptualized as a bounded object (the nail lacquer and the hairspray as products you can buy at the supermarket). The same can be observed in (8) and (9) for *talkpoeder* 'talc powder' and *waspoeder* 'detergent' respectively, where the *de*-article is preferred to refer to products (8a) and (9a) while the neuter appears in syntactic constructions conveying the idea of an unbounded mass (8b) and (9b):

(8) a. *De* *talkpoeder* staat bij ons altijd op de toilettafel
DET-C talc powder stand by us always on the toilet table
'At our place the talc powder is always on the toilet table'[14]

[9] Available at: www.evmi.nl/nieuws/marketing-sales/5358/consument-mist-koeleman-appelmoes.html
[10] Available at: www.ciao-shopping.nl/sr/q-schwarzkopf_junior_haarlak
[11] Available at: www.ciao-shopping.nl/Wella_Forte_Hairspray_Ultra_Sterk_1025602
[12] Available at: www.opi.nl/nagellak/nailstudio.htm
[13] Available at: www.ciao-shopping.nl/Herome_Caring_Nail_Polish_Remover_Pads_1032016
[14] Available at: www.yunomi.nl/artikel/talkpoeder-weetjes

b. *Het talkpoeder* is huidneutraal en vrij van parfum
DET-N talc powder is skin neutral and free of perfume
'The talc powder is neutral skin and without any added perfume'[15]

(9) a. *De waspoeder* van Klok Eco minimaliseert de kans
DET-C washing powder of Klok Eco minimizes the possibility

op huidproblemen
of skin problems

'The washing powder Klok Eco minimizes skin problems'[16]

b. Tot aan de uitvinding van *het waspoeder* was wassen zwaar handwerk.

Until the invention of DET-N washing powder was washing heavy handwork

'Until the invention of the washing powder the washing up was heavy handwork'[17]

All these examples display a tendency in the selection of the definite article which seems to confirm the possibility for gender to be contextually chosen according to different individuation degrees of the referent.

As a matter of fact, the investigation of intra-speaker variation – gender shifts produced in the same utterance by the same speaker – shows that different genders can be triggered also in the same utterance, and that such a variation is cognitively explainable recalling the same cognitive opposition, i.e. individuated vs. non individuated reading (Semplicini 2012a; 2012b). Consider *aperitief* in (10):

(10) a. *De perfecte aperitief*: fruitbier met chocolade
DET-C perfect-C apéritif: fruit beer with chocolate
'The perfect apéritif: fruit beer with chocolate'

b. Chocolade kan perfect *bij het aperitief* geserveerd worden
Chocolate can perfectly by DET-N apéritif served become
'Chocolate is perfectly suited to be served by the apéritif'[18]

In (10) the gender shift *de aperitief* > *het aperitief* recalls and confirms the distinction observed in (1): neuter for the 'occasion' and common gender for the bounded object. The different conceptualization is also conveyed through different syntactic roles: common gender for the subject position and neuter gender for the prepositional NP *bij* + DET + N.

As far as prepositional NPs are concerned, a very interesting case is provided by the construction *op* + DET + N: in (11) – (13) the neuter correlates with a low

[15]Available at: www.choosebeauty.nl/index.php?option=com_jreviews&task=listcategory&section=5&cat=997&dir=8&Itemid=20
[16] Available at: www.allergiewinkel.nl/product/klok-eco-waspoeder-wit/20734/
[17] Available at: www.schooltv.nl/vroegerenzo/2202052/wist-je-datjes/
[18] Available at: ttp://www.hln.be/hln/nl/39/Lekker-Eten/article/detail/1081148/2010/03/17/De-perfecte-aperitief-fruitbier-met-chocolade.dhtml

individuated interpretation of the object, but when the noun refers to a more individuated entity the speaker selects common gender. In other words, the preposition *op* 'on' coerces the interpretation of the noun, i.e. bounded object > unbounded surface, while the *de*-article is the most suited to appear in either subject or object position:

(11) a. Hij denkt dan ook dat de dief met een stoel *op het altaar*
He thinks then also that the thief with a chair on DET-N altar

heeft gestaan
has stond
'He also thinks that the thief got on the altar with a chair'

b. Uit politieonderzoek blijkt echter dat *de altaar* stoffig was.
From police investigation seems in fact that DET-C altar dusty was
'According to the police investigation came out that the altar was dusty'

(12) a. Er blijft altijd deeg achter *op het aanrecht*
There remains always dough behind on DET-N work surface
'Some dough always remains on the work surface'

b. Nadeel is dat je *de aanrecht* eerst goed schoon maken moet.
Disadvantage is that you DET-C work surface first good clean make must
'The disadvantage is that first of all you must clean the work surface'

(13) a. Eigen foto's en tekst *op het omslag*.
Own photos and texts on DET-N cover
'Your own photos and texts on the cover'

b. *Zachte omslag* geeft paperback uitstraling.
Soft-C cover gives paperback brightness
'Soft cover gives brightness to the paperback'

Something similar is shown by *vergiet* 'colander' in (14). Common gender appears in the NP in object position that conceptualizes the referent as a bounded object (14a), but when the colander represents the less individuated entity with respect to another object (the *deksel* 'cover' is more individuated than the colander) the speaker switches to neuter (14b):

(14) a. Plaats *de vergiet* erboven op
Put DET-C colander there over on
'Put the colander over it'

b. Plaats een deksel *bovenop het vergiet*
Put a cover over on DET-N colander
'Put a cover over the colander'

Other systematic interactions are provided by NPs that correlate with the verb *smelten* 'melt': the meaning of the verb coerce the interpretation of the noun that must necessarily refer to an unbounded substance, as in (15) for *drop* 'liquorice'.

(15) a. Doe *de drop* in het pannetje.

put DET-C liquorice in the little pan
'Put the liquorice in the little pan'

b. Een klein beetje water toevoegen kan helpen *het* *drop* beter
a little bit water add can help DET-N liquorice better

te laten smelten
to let melt
'Adding a bit of water can help to melt the liquorice'[19]

In fact, the relevance of this construction has been proved to be at work also for stable nouns. The noun *kaas* 'cheese' for instance is commonly considered a *de*-word and therefore NPs like **het kaas* are simply considered as gender mistakes. Anyway, a deeper look at examples like those in (15) and (16) presumes that these gender mistakes are quite systematic.

(16) Beleg het brood met kaas, vlees en groenten. Leg het even onder de grill
Fill the bread with cheese, meat and vegetables. Put it then under the grill

tot *het* *kaas* gesmolten is.
till DET-N cheese melted is.

'Fill the bread with cheese, meat and vegetables. Put it under the grill till the cheese melts down'[20]

Another correlation between gender selection and mass reading is displayed by NPs introduced by the quantifier *al* 'all' that occurs with definite mass nouns and definite plurals – it is not specified for number – and stands in complementary distribution with *heel* 'whole' that is restricted to definite singulars (Zwarts 1992).

(16) a. Men denkt dat men bang is om *de* *kauwgom* door te slikken,
One think that one scared is to DET-C chewing-gum through to swallow

en dat daardoor mensen *de* *kauwgom* ergens opplakken
and that therefore people DET-C chewing-gum anywhere stick

of op straat gooien
or on street throw

'One thinks that people are scared to swallow the chewing gum and therefore they either stick the chewing gum anywhere or throw (it) on the road.'

b. Intussen is *al het* *kauwgom* op straat een heel milieuprobleem
So is all DET-N chewing gum on street a whole environment problem

geworden want het blijft liggen.
become because it keeps stay.

[19] Available at: www.dropenzo.nl/drop-en-gerechten/drop-en-ijs/drop-en-ijs-maken-the-easy-way.html
[20] Available at: eten-en-drinken.infonu.nl/recepten/21532-recepten-met-oud-brood.html

'Accordingly, all the chewing gum on the road has become a problem for the environment, because it keeps staying there.'

Similar cases can be observed also for stable nouns referring to substances like *de boter* 'butter' and *de melk* 'milk':

(19) Als *al het boter* is gesmolten voeg je de tomatenpuree toe
When all DET-N butter is melted add you the tomato puree to
'When all the butter has melted down add the tomato puree'[21]

(20) Voeg niet in één keer *al het melk* toe, dat is niet altijd nodig.
Add not in a time all DET-N milk to, that is not always necessary
'Do not add all the milk at a time, that is not always necessary.'[22]

Further confirmations for the correlation between mass conceptualization and wrong grammatical gender come from the observation that instances of the construction *heel* + **het boter/*het melk* were not found.

Moreover, an investigation carried out on the web at different times revealed an increasing number of instances: from 13 to 15 instances of **het boter* in a few days. The most relevant fact is that all cases of deviation from grammatical gender are characterized by constructions that convey a low degree of individuation: **het boter smelten* 'to melt down the butter', **het boter mixen* 'to mix the butter', *al *het boter* 'all the butter', etc. Some instances of the query **het boter* are given in (21):

(21) a. ..smelt eerst *het boter* op hoog vuur in de pan. Wanneer *de boter*
melt first DET-N butter on high fire in the pan. When DET-C butter

lichtbruin is..
light brown is..

'First of all melt the butter in the pan on a high fire. When the butter becomes light brown..'

b. ..mix *het boter* met ei en suiker klop dit goed en voeg..
mix DET-N butter with egg and sugar beat this well and add..
'Mix the butter with egg and sugar, beat it and add..'

c1. ..met gestolde geklaarde boter. Een deel van *het boter*
with solidified-C finished-C butter. A piece of DET-N butter

gaan we gebruiken voor het deeg..
go we use for the dough..

'..with solidified ready butter. We're going to use a portion of the butter for the dough..'

c2. ..or het smelten verandert de moleculestructuur van *het boter..*
or the melting changes the molecular structure of DET-N butter
'or the melting (process) changes the molecular structure of the butter'

d. ..laat *het boter* smelten en bak de forel

[21] Available at: forum.scholieren.com/archive/index.php/t-448722.html
[22] Available at: www.smulweb.nl/recepten/1396150/Pepernootjes

let DET-N butter melt down and bake the trout
'Let the butter melt down and bake the trout'

e. *Het boter* laten smelten op een laag vuur
DET-N butter let melt down on a low fire
'Let the butter melt down on a low fire'

All these results seem to confirm that Dutch speakers are gradually losing their feeling for grammatical gender not only at the pronominal level but also for nominals. The observation of common agreement patterns based on different degrees of individuation rather than on lexical gender – not only for DGNs but also for stable *de*-words and *het*-words – let us suppose that gender fluctuations in the nominal domain are less marginal than currently claimed and that, especially in systems where gender is covert, referential gender may override lexical gender and strictly interact with morphosyntactic, semantic and pragmatic factors.

6. Conclusions

The analysis of Dutch DGNs in context has revealed some patterns that recall semantic pronominal agreement: the selection of an article instead of the other seems to depend on the cognitive perception of the referent the noun stands for in each specific utterance (the masculine/feminine article used for more individuated referents and the neuter article for less individuated ones).

In many cases the selection of gender seems to be related to either the particular syntactic construction the noun appears in or the syntactic prominence of the definite NP: this means that different gender values may be selected on the basis of specific communicative needs focusing on the semantic/pragmatic properties of the referent that are the most relevant in a certain context (Wisniewski 2003; 2009) and reflected by specific syntactic constructions and thematic roles.

Actually, it has been observed that certain syntactic constructions are more probable to trigger semantic rather than grammatical agreement not only for DGNs but also for stable gender nouns: for instance the recurrence of patterns like *op het + N*, *al het* + N, *het N + smelten* and the preference accorded to common gender for definite NPs with a higher degree of salience, i.e. in subject and object position, presume that once the sensitivity for grammatical gender has been lost speakers make their choices following their own cognitive perception of the referent which is also reflected at the morphosyntactic level by specific syntactic configurations.

These tendencies let us suppose that contemporary Dutch gender as a whole is subject to a 'recategorization' process which is leading to a (more) semantic system: the gradual loss of morphological endings (originally marking inflectional classes and gender) which led to the loss of sensitivity for grammatical gender has been triggering the restructuring of the category along semantic pathways.

Further confirmations come from a questionnaire study for native speakers in Flanders: in sentence completion task Flemish speakers produced gender shifts according to different perceptions of the referent. For instance, some informants associate the DGN *kluwen* 'tangle' with the neuter article *het* when it referred to the tangle as an 'indistinct amount of yarn', but to common gender (*de kluwen*) when it meant the 'yarn' as a bounded object (Semplicini 2012a).

This transition towards a conceptual system is demonstrated not only by increasing semantic pronominal agreement, but also by the analysis of definite NPs. Some systematic tendencies in the choice of definite articles for DGNs and, to a lesser extent, also for stable gender nouns suggest a step forward in the transition of Dutch gender based on different degrees of individuation and pragmatic salience.

In conclusion, DGNs should no longer be confined to the issue of dialectal variation and definitely deserve more attention because they represent a relevant step in the ongoing recategorization process and their analysis helps motivating not only the article selection for *de/het*-nouns, but also suggests another possible explanation for 'apparent' mistakes regarding stable gender nouns.

APPENDIX

The appendix contains those cases of intra-speaker gender variation for DGNs that have been quoted in distinct sentences in the body of the article.

- *de/het aperitief* (aperitif)

  De perfecte aperitief: fruitbier met chocolade
  De Oudenaardse brouwerij Liefmans serveert dit voorjaar zijn nieuwste fruitbier met donkere chocolade. [...] Chocolade kan perfect *bij het aperitief* geserveerd worden, oordeelde men bij Liefmans. "De bittere smaak van de chocolade harmonieert prima met het zoete fruitbier waardoor een nieuwe aperitiefbelevenis ontstaat", oordeelt Marc Coessens, meester-brouwer.

  The perfect aperitif-C: fruit beer with chocolate
  This year the brewery Liefmans in Oudenaarde serves its new fruit beer with dark chocolate. […] Chocolate is perfectly suited to be served *by the aperitif-N*, agree people by Liefmans. 'The bitter flavor of the chocolate goes very well with the sweet fruit beer whereby a new aperitif experience arises', asserted the master brewer Marc Coessens.

  http://www.hln.be/hln/nl/39/Lekker-Eten/article/detail/1081148/2010/03/17/De-perfecte-aperitief-fruitbier-met-chocolade.dhtml

- *de/het altaar* (altar)

  De bovenkant van het raamwerk bevond zich op 1.85 boven de altaartafel. Hij denkt dan ook dat de dief met een stoel *op het altaar* heeft gestaan om het paneel los te krijgen van boven.[...] waarom er geen voetafdrukken op het altaar waren. Als je binnen staat en het altaarstuk naar boven uit de lijst wil halen dan moet je *op het altaar* gaan staan. Uit politieonderzoek blijkt echter dat *de altaar* stoffig was en dat de stof niet was aangetast. Ook waren er geen voetafdrukken op het opgevouwen altaarkleed. Niets wees erop dat men *op het altaar* is geklommen! (Bron: Karl Hammer. Hij baseert zich op een politierapport van Hantsen.).

  The upper side of the window frame is 1.85 above the altar table. He also thinks that the thief got *on the altar-N* with a chair to get the panel from above. […] why there were no footprints *on the altar-N* If you are inside and you want to take the altar piece up from the frame then you have to *be on the altar-N*. According to the police investigation came out that *the altar-C* was dusty and that the dust had not been removed. Moreover there were no footprints on the bent altar cloth. Nothing indicates that someone has climbed *on the altar-N*! (Source: Karl Hammer. He goes on the basis of the report of the police of Hantsen).

  http://www.oogvanhorus.nl/03_Diefstal/02_reconstructie.htm

- *de/het aanrecht* (work surface)

  Brooddeeg kneden *op het aanrecht*
  Dan maar met de handen kneden *op het aanrecht.* [...] Nadeel is dat je *de aanrecht* eerst goed schoon moet maken. En dan nog een keer als je klaar bent, want er blijft altijd deeg achter *op het aanrecht*.

Bread dough kneading *on the work surface-N*
Therefore kneading with the hands *on the work surface-N*.[…] The disadvantage is that before you have to clean *the work surface-C*. And then again another time as you are ready because dough always remains *on the work surface-N*.

http://floorp.nl/broodbakmachine/hoe-het-begon-brood-bakken.htm

- *de/het omslag* (cover)

Handzaam en voordelig fotoboek.
Eigen foto's en tekst *op het omslag.*
Zachte omslag geeft paperback uitstraling.
Maak nu je eigen fotoboek
Fotoboeken *met zachte omslag*
Paperbacks waren altijd al populair. Niet alleen door de aantrekkelijke prijs, maar ook omdat ze gemakkelijk zijn mee te nemen en door te bladeren. Doordat je *het omslag* helemaal zelf kan samenstellen, krijg je een heel persoonlijk exemplaar. Net zoals bij de fotoboeken met *een harde omslag* heb je ook hier alle mogelijkheden om je fotoboek te verfraaien met achtergronden, teksten, scraps en pagina indelingen.

Photo books convenient and easy to handle
Your own photos and texts *on the cover-N*
Soft cover-C gives brightness to the paperback
Make now your own photo book
Photo books *with soft cover-C*
Paperbacks have always been popular. Not only because of the attractive price, but also because they are easy to handle and to flick trough. Given that you can yourself assemble *the cover-N* you get an highly personal copy. Just as in the case of photo books with *a soft cover-C* you have all the means to embellish your photo book with backgrounds, texts, scraps and pages distribution.

http://www.myphotofun.nl/producten/fotoboek-zachte-omslag.aspx

- *de/het vergiet* (colander)

Vul de pan met een laag water, laat het water koken en plaats *de vergiet* erboven op. Doe *in het vergiet* de rijst - die je vantevoren even gekookt heb - plaats een deksel *bovenop het vergiet* en laat het zo stomen.Let op: het water mag niet *in het vergiet* komen, dus doe niet teveel water in de pan!

Fill the pan with little water, let the water boil and put *the colander-C* over that. Put the rice *in the colander-N* – which you have already cooked – put a cover *over the colander-N* and let it steam. Be careful: the water must not pass *through the colander-N*, therefore do not put too much water in the pan!

http://eten-en-drinken.infonu.nl/recepten/4031-indische-snacks-lemper-hartig-gevulde-rijstrolletjes.html

- *de/het drop* (liquorice)

Wat moet je doen? Doe *de drop* in het pannetje en zet deze op een laag pitje. Laat *de drop* smelten en voeg naar smaak suiker toe, om het dropijs zoeter te krijgen. Blijf goed roeren. Als *de drop* niet makkelijk smelt zou je een beetje water toe kunnen voegen. Een klein beetje water toevoegen kan helpen om *het drop* beter te laten smelten zonder dat het aankoekt.

What do you have to do? Put *the liquorice-C* in the little pan and leave it on a low heat. Let the *liquorice-C* melt and add sugar as much as you like to get a sweeter liquorice ice cream. Keep mixing. *If the liquorice-C* does not melt easy you could add a bit of water. Adding a bit of water can help to melt *the liquorice-N* avoiding that it overcooks.

http://www.dropenzo.nl/drop-en-gerechten/drop-en-ijs/drop-en-ijs-maken-the-easy-way.html

- *de/het kauwgom* (chewing gum)

Kauwgom wordt gemaakt in een kauwgomfabriek. Men gebruikt daarbij natuurlijke gom 'de gom uit bomen' maar ook kunstmatige gom. Die maakt de fabriek zelf. Eerst gaan de gommen in een smeltmachine, die roert de gommen door elkaar. Gom smelt als het warm wordt, dus worden ze daar warm gemaakt. Daarna moet er suiker bij *de gom*. Daarvoor wordt poedersuiker gebruikt. Als de gommen weer koud zijn gaan ze in de kneedmachine en daar doen ze er suiker, een smaakje, en een kleurstof bij. Dan gaat *de kauwgom* in een machine die er een lint van maakt. Dan gaat het lint door een walsmachine. Daar wordt het plat gemaakt en dan komt het in de snijmachine en daar wordt het in stukjes gemaakt. *Het kauwgom* wordt afgekoeld en dan is het klaar.

[...]

Het probleem met kauwgom.

Gemiddeld eten we met iedereen van de wereld 100 000 ton kauwgom per jaar. Men eet gemiddeld per persoon 330 gram kauwgom per jaar, en 1 kauwgompje van sportlife, xylifresh, freedent of stimerol weegt 1.1 gram. Men eet dus gemiddeld 300 kauwgoms per persoon, per jaar. Met *al het kauwgom* op trottoirs en de onderkant van schoolbanken e.d. kan men de Queen Elizabeth II na bouwen. Men denkt dat men bang is om *de kauwgom* door te slikken, en dat daardoor mensen *de kauwgom* ergens opplakken of op straat gooien. [...] Intussen is *al het kauwgom* op straat een heel milieuprobleem geworden, want het blijft liggen. Daarom wil men een kauwgom maken die afbreekt in je mond. Dat hebben ze ook al ooit gedaan. Hij loste na 10 tot 15 minuten op. Het nadeel van *de kauwgom* was dat niemand hem lekker vond.

How is chewing gum made?

Chewing gum is made in a chewing gum factory. Where people use natural gum 'the gum from trees' but also synthetic gum. This is made in the factory itself. First of all gums are put into a mixing machine, that stirs the gums all together. When the gum becomes warm it melts and therefore it must be warmed. After, sugar must be added to *the gum-C* sugar powder is used. When the gums are cool again, they go into the cut machine where sugar, flavor and colouring are added. Then *the chewing gum-C* goes in a machine that transform it into a ribbon. Then the ribbon goes into the smoothing machine where it is made flat and in the cut machine where it is cut into pieces. *The chewing gum-N* is cooled and then it is ready.

[…]

The chewing gum problem.

On average we eat in the world 100.000 tons of chewing gum per year. On average 300 grams of chewing gum are eaten per person per year, and a little chewing gum of

sportlife, xylifresh, freedent or stimerol weighs 1.1 gram. Therefore 330 chewing gums are eaten per person, per year. With *all the chewing gum-N* on sidewalks and under school desks it could be possible to build Queen Elisabeth II. One thinks that people are scared to swallow *the chewing gum-C* and therefore they either stick *the chewing gum-C* anywhere or throw (it) on the road. […] Accordingly, *all the chewing gum-N* on the road has become a problem for the environment, because it-N keeps staying there. Therefore one should develop *a chewing gum which-C melts* in your mouth. It has been already made. It-M melts in 10 to 15 minutes. The disadvantage of *this chewing gum-C* was that nobody found it-M tasty.
http://huiswerk.leerlingen.com/bekijken.php?id=23242

References

Audring, Jenny 2006. Pronominal gender in spoken Dutch. In *Journal of Germanic Linguistics* 18: 85- 116.

Audring, Jenny 2009. *Reinventing Pronoun Gender*. Utrecht: LOT.

Booij, Geert and Jenny Audring 2009. Genus als probleemcategorie. *Taal en Tongval*. Themanummer 22:13-37.

Booij, Geert 2002. *The morphology of Dutch*. Oxford: OUP.

Braunmüller, Kurt 1999. Gender in North Germanic: A diasystematic and functional approach. In Unterbeck, Barbara and Matti Rissanen (eds), Gender in Grammar and Cognition, 25–53. Berlin: Mouton De Gruyter.

Corbett, Greville 1979. The Agreement Hierarchy. In *Journal of Linguistics* 15: 203-224.

Corbett, Greville 1991. *Gender*. Cambridge: CUP.

Corbett, Greville 2006. *Agreement*. Cambridge: CUP.

Dahl, Östen 1999. Animacy and the notion of semantic gender. In Unterbeck, Barbara and Matti Rissanen (eds), *Gender in Grammar and Cognition*, 99-116. Berlin: Mouton De Gruyter.

De Vogelaer, Gunther 2006. Pronominal Genus bij 'Zuid-Nederlandse' taalverwervers: grammaticaal of semantisch systeem?. In Hüning, Matthias, Vogl, Ulrike, Wouden van der, Ton and Arie Verhagen (eds) Nederlands tussen Duits en Engels, 89–102. Leiden: Stichting Nederlandistiek.

De Vogelaer, Gunther, and Gert De Sutter 2010. The geography of gender change: Pronominal and adnominal gender in Flemish dialects of Dutch. In *Language Sciences* 33: 192–205.

De Vogelaer, Gunther, and Lien De Vos 2011. Dutch gender and the locus of morphological regularization. *Folia Linguistica* 45: 245–281.

De Vos, Lien 2009. De dynamiek van hersematisering. *Taal en Tongval*. Themanummer 22. 80–109.

De Vries, Jelle 2001. *Onze Nederlandse spreektaal*. Den Haag: Sdu.

Enger, Hans-Olav 2004. On the relation between gender and declension: a diachronic perspective from Norwegian. In *Studies in Language* 28: 51-82.

Fernández Ordoñez, Inès 2009. The development of mass/count distinctions in Indo-European varieties. In Bubenik, Vit, Hewson, John and Sarah Rose (eds), *Grammatical change in Indo-European languages. Papers from the Workshop on Indo-European Linguistics at the 28th International Conference on Historical Linguistics, Montreal, 2007*, 55-70. Amsterdam: Benjamins.

Fletcher, William 1987. Semantic factors in Dutch gender choice. In William Fletcher (ed.), *Papers from the 2nd Interdisciplinary Conference on Netherlandic Studies, Georgetown University, June 7–9, 1984*, 51–63. Lanham: University Press of America.

Haase, Martin 1999. Reorganization of a gender system: The central Italian neuters. In Unterbeck, Barbara and Matti Rissanen (eds), *Gender in Grammar and Cognition*, 221–236. Berlin: Mouton De Gruyter.

Hockett, Charles Francis 1958. *A course in modern linguistics*. New York: Macmillian.

Josefsson, Gunlög 2006. Semantic and grammatical genders in Swedish – independent but interacting dimensions. In *Lingua* 116.9: 1346-1368.

Krifka, Manfred 1995. Common nouns: A contrastive analysis of Chinese and English. In Carlson, Gregory and Francis Jeffry Pelletier (eds), The generic book, 398–411. Chicago: University of Chicago Press.

Langacker, Ronald 1991. *Foundations of Cognitive Grammar* . Stanford, California: Stanford University Press.

Leiss, Elisabeth 1999. Gender in Old High German. In Unterbeck, Barbara and Matti Rissanen (eds), *Gender in Grammar and Cognition*, 237-257. Berlin: Mouton De Gruyter.

Lüdtke, Helmut 2001. Zu spanischen lo und zum Genus neutrum in Romanischen. In Gerd Wotjak (ed.), *Studien zum romanisch-deutschen und innerromanischen Sprachvergleich*, 185–189. Frankfurt: Lang.

Quine, William Van Orman 1960. *Word and Object*. Cambridge CA: MIT Press.

Michaelis, Laura 2005. Entity and event coercion in a symbolic theory of syntax. In Oestman, Jan-Ola and Mirjam Fried (eds), *Construction Grammar(s): Cognitive Grounding and Theoretical Extensions. Constructional Approaches to Language*, 45-88. Amsterdam: Benjamins.

Parafragou, Anna 2005. Relations between language and thought: Individuation and the count/mass distinction. In Cohen, Henri and Claire Lefebvre (eds), *Handbook of Categorization in Cognitive Science*, 255–275. Oxford: Elsevier.

Pauwels, Jan 1938. *Bijdrage tot de kennis van het geslacht der substantieven in Zuid-Nederland*. Tongeren: Michiels.

Romijn, Kirsten 1996. *Hoe doen we het? verwijzen naar linguïstische en cognitieve representaties met het voornaamwoord 'het'*. Amsterdam: Publicaties van het P.J. Meertens-Instituut.

Rose, Ralph 2011. Joint Information Value of Syntactic and Semantic Prominence for Subsequent Pronominal Reference. In Chiarcos, C., Claus, B. and M. Grabski (eds), *Salience. Multidisciplinary Perspectives on its Function in Discourse*, 81-103. Berlin: Mouton de Gruyter.

Sasse, Hans-Jürgen 1993. Syntactic categories and subcategories. Syntax. In Jacobs, Joachim, Stechow von, Arnim, Sternefeld, Wolfganf and Theo Vennemann (eds), *Ein internationales Handbuch zeitgenössischer Forschung/An international handbook of contemporary research*, 646–686. Berlin: Mouton De Gruyter.

Semplicini, Chiara 2012a. *Het/de kluwen doet denken aan de/het draad: an analysis of Dutch double gender nouns*. Ph.D Dissertation, University of Perugia.

Semplicini, Chiara 2012b. Dutch double gender nouns: arbitrary or motivated agreement? In *Journal of Germanic Linguistics* 24.2: 133–186. doi: 10.1017/S1470542712000025

Siemund, Peter 2008. *Pronominal gender in English. A study of English varieties from a cross-linguistic perspective*. London: Routledge.

Talmy, Leonard 2000. *Toward a cognitive semantics*. Cambridge, MA: MIT Press.

Weber, Doris 1999. On the function of gender. In Unterbeck, Barbara and Matti Rissanen (eds), *Gender in Grammar and Cognition*, 495-509. Berlin: Mouton De Gruyter.

Wisniewski, Edward, Lamb, Christopher and Erica Middleton 2003. On the conceptual basis of the count and mass noun distinction. In *Language and Cognitive Processes* 18: 583–624.

Wisniewski, Edward, Middleton, Erica, Trindel, Kelly and Mutsumi Imai. 2004. Separating the chaff from the oats: Evidence for a conceptual distinction between count noun and mass noun aggregates. In *Journal of Memory and Language* 50: 371–394.

Wisniewski, Edward 2009. On using count nouns, mass nouns, and pluralia tantum: What counts? In Francis Jeffry Pelletier (ed.), *Kinds, things and stuff. Mass terms and generics*, 166–190. New York: Oxford University Press.

Wurzel, Wolfgang 1986. Die Wiederholte Klassifikation von Substantiven. In *Zeitschrift für Phonetik, Sprachwissenschaft und kommunikation* 39: 76-96.

Zwarts, Joost 1992. *X'-syntax – X'-semantics: on the interpretation of functional and lexical heads*. Ph.D Dissertation, University of Utrecht.

The formal and functional architecture of inflectional morphology[1]

Gregory Stump
University of Kentucky
gstump@uky.edu

> It is the pervading law of all things organic and inorganic, of all things physical and metaphysical, of all things human and all things superhuman, of all true manifestations of the head, of the heart, of the soul, that the life is recognizable in its expression, that form ever follows function. This is the law.
>
> Architect Louis Sullivan,
> in "The tall office building artistically considered," *Lippincott's Magazine* (March 1896)

1. Introduction

In thinking about the architectural theme of the Eighth Mediterranean Morphology Meeting, I was drawn to the epigraph from Louis Sullivan, one of the pioneers of modernist architecture. Seen from a distance, the architecture of inflectional morphology conforms to Sullivan's law: the function of inflectional morphology is to relate content to form at the level of words, and the overall architecture of the component fits this function. But once we examine the architecture of inflection more closely, it is less clear that form invariably follows function in inflectional morphology. There are mismatches between form and function. In particular, if form straightforwardly followed function, we might expect words that are functionally alike to be expressed in formally identical ways, but this expectation is sometimes disconfirmed: the relation between content and form can be much more complicated. Two words can be alike in their morphosyntactic content but differ in the manner of their formal definition.

Sullivan's law does, however, raise the possibility of distinguishing morphological phenomena according to how closely they adhere to an ideal of optimal functionality. I assume that the most functional inflection is what Corbett (2009) has called canonical inflection. Canonical inflection is a kind of typological extreme relative to which actual inflectional systems can be calibrated. It encompasses both the notion of a canonical inflectional paradigm and the more general notion of a canonical system of inflectional paradigms (for some syntactic category). The characteristics of canonical inflection are paraphrased in (1).

(1) Canonical inflection (paraphrased from Corbett 2009)

a. Properties of a canonical inflectional paradigm

- **Exhaustivity**: Every compatible combination of the relevant morphosyntactic properties defines a cell.
- **Completeness**: Every cell has a realization.
- **Unambiguousness**: All realizations are distinct.
- **Freedom from stem alternation**: Every realization is based on the same stem.

[1] This paper was first presented at the Eighth Mediterranean Morphology Meeting, Cagliari, Sardinia, Italy, September 14-17, 2011. Several of those present at the meeting made helpful suggestions.

- **Morphotactic uniformity**: the same morphotactic pattern (e.g. stem+suffix) is used in every realization.

b. Properties of a canonical system of paradigms (for some syntactic category)

- **Parallelism**: All individual paradigms realize the same morphosyntactic property sets and all are canonical.
- **Distinctness**: Distinct paradigms are based on distinct stems and therefore have distinct realizations.
- **Uniformity of exponence**: Across paradigms, the same morphosyntactic property set is expressed by the same exponence.

The paradigm of the Breton inflecting preposition HERVEZ 'according to' in (2a) might be seen as nearly canonical: although it only shows a gender distinction in the 3sg and therefore fails the test of exhaustivity, it passes the tests of completeness, unambiguousness, freedom from stem alternation and morphotactic uniformity. The system of prepositional paradigms in which HERVEZ is embedded is somewhat less canonical; as a class, prepositional paradigms aren't parallel (some involve stem alternation, e.g. those of 'of' and 'against' in (2b)) and there is no uniformity of exponence (for example, 'according to' and 'against' belong to distinct conjugation classes).

(2) The paradigms of four inflecting prepositions in Breton

a.

	HERVEZ
	'according to'
1sg	*hervez-on*
2sg	*hervez-out*
3sgm	*hervez-añ*
3sgf	*hervez-i*
1pl	*hervez-omp*
2pl	*hervez-oc'h*
3pl	*hervez-o*

b.

	E	A	OUZH
	'in'	'of'	'against'
1sg	*enn-on*	*ac'han-on*	*ouzh-in*
2sg	*enn-out*	*ac'han-out*	*ouzh-it*
3sgm	*enn-añ*	*anezh-añ*	*out-añ*
3sgf	*enn-i*	*anezh-i*	*out-i*
1pl	*enn-omp*	*ac'han-omp*	*ouzh-imp*
2pl	*enn-oc'h*	*ac'han-oc'h*	*ouzh-oc'h*
3pl	*enn-o*	*anezh-o*	*out-o*

As this example suggests, purely canonical inflection isn't common; most inflectional phenomena are in some manner or other noncanonical, failing to exhibit one or more of the properties in (1).

Here, I wish to discuss an architecture for inflectional morphology that elucidates the ways in which noncanonical inflection deviates from canonical patterns. I begin with a general discussion of the assumed architecture of inflection and how it relates to canonical inflection. I then examine how the various noncanonical inflectional phenomena in (3) are situated with respect to this architecture.

(3) Noncanonical inflectional phenomena
- a. Defectiveness
- b. Syncretism
- c. Deponency
- d. Functor-argument reversal
- e. Suppletion

2. The architecture of content paradigms, form paradigms and paradigm linkage

I begin with the assumption that the purpose of inflectional morphology is to give **phonological** expression to lexeme + property set pairings supplied by the **syntax**. It is

therefore inherently an interface component, and this fact partially determines its formal architecture. But two additional facts are relevant to the details of this architecture. First, words possess grammatical properties to which rules of syntax and semantics are insensitive; these include membership in inflection classes and other morphomic categories. The architecture of inflectional morphology should entail this fact. Second, the same word may have (or may appear to have) distinct morphosyntactic properties for different purposes. Generally, morphosyntactic property sets serve the three purposes in (4):

(4) A word's morphosyntactic property set
  a. constrains its lexical insertion
  b. determines its semantic interpretation
  c. induces the introduction of its inflectional exponents

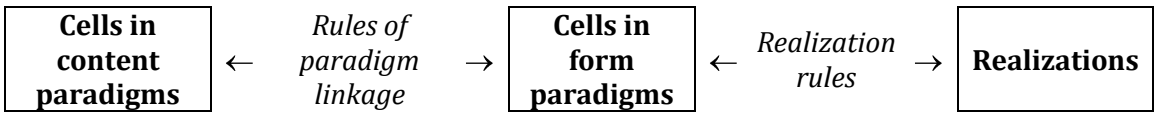
Yet, some words behave as if the property set serving purpose (4c) is different from the set serving purposes (4a) and (4b); for instance, the Latin deponent verb *hortātur* 's/he urges' has the property set in (5a) for purposes of lexical insertion and semantic interpretation, yet its inflectional exponence instead presumes the property set in (5b).

(5) *hortātur* 's/he urges'
  a. {3sg present indicative active}
  b. {3sg present indicative passive}

The architecture of inflectional morphology should account for this lack of parallelism.

In view of these considerations, I have argued in recent work (Stump 2002, 2006, 2007; Stewart and Stump 2007) that the architecture of inflection involves two levels of paradigmatic representation, one with syntacticosemantic relevance, the other with morphophonological relevance. On this view, the job of the inflectional rule system is to relate the two sorts of paradigms and to determine their realization. In particular, I assume that the architecture of inflectional morphology has the form in (6); that is, there are three kinds of representations (content cells, form cells, realizations) whose relation to one another is mediated by rules of paradigm linkage and realization rules.

(6) The architecture of inflectional morphology



On one hand, lexemes have content paradigms:

(7) a. A lexeme's **content paradigm** is its full inventory of content cells.
  b. A **content cell** for a lexeme L is a pairing ⟨L, σ⟩ of L with a morphosyntactic property set σ with which L may be associated in syntax; thus, the cells in the content paradigm of lexeme L specify the range of syntactic contexts in which L appears.
  c. The content paradigm of the Breton prepositional lexeme HERVEZ 'according to':
    ⟨HERVEZ, {1sg}⟩
    ⟨HERVEZ, {2sg}⟩
    ⟨HERVEZ, {3sgm}⟩
    ⟨HERVEZ, {3sgf}⟩
    ⟨HERVEZ, {1pl}⟩
    ⟨HERVEZ, {2pl}⟩

⟨HERVEZ, {3pl}⟩

d. Semantic interpretation of ⟨HERVEZ, {3pl}⟩:
λp∀x[x ∈ X_i → p ∈ **according-to**′(x)] 'according to them$_i$' (a set of propositions)

A lexeme's content cells determine lexical insertion and semantic interpretation; in particular, the realization of a content cell ⟨L, σ⟩ is insertable in a node specified for the morphosyntactic property set σ, and the content cell determines the semantic interpretation of the resulting structure. For instance, the realization *hervezo* of the content cell ⟨HERVEZ, {3pl}⟩ is insertable in a prepositional node specified as third-person plural, and the content cell entails that the resulting structure will have a semantic interpretation something like (7d).

While lexemes have content paradigms, stems, on the other hand, have form paradigms:

(8) a. A stem's **form paradigm** is its full inventory of form cells.
b. A **form cell** for a stem X is a pairing ⟨X, σ⟩ of X with a set σ of morphosyntactic properties for which X may be inflected; thus, the cells in the form paradigm of stem X specify the range of property sets for which X is inflectable.
c. **Inflection classes** are seen as properties of stems rather than of lexemes; it is therefore at the level of form paradigms (rather than that of content paradigms) that inflection-class distinctions are made.
d. The form paradigm of the Breton prepositional stem *hervez*$_{[Cl.1]}$ 'according to' (a member of prepositional inflection class 1):
⟨*hervez*$_{[Cl.1]}$, {1sg}⟩
⟨*hervez*$_{[Cl.1]}$, {2sg}⟩
⟨*hervez*$_{[Cl.1]}$, {3sgm}⟩
⟨*hervez*$_{[Cl.1]}$, {3sgf}⟩
⟨*hervez*$_{[Cl.1]}$, {1pl}⟩
⟨*hervez*$_{[Cl.1]}$, {2pl}⟩
⟨*hervez*$_{[Cl.1]}$, {3pl}⟩

A stem's form paradigm determines its realization; in particular, realization rules apply to each of a stem's form cells to determine its morphophonological expression. Thus, we might postulate the realization rules in (9) to account for the realization of the form cells in (8d); the resulting realizations are as in (2).

(9) Realization rules for Breton prepositions (Class 1)

	Form cell		Realization
a.	⟨$X_{[Cl.1]}$, {1sg}⟩	→	X*on*
b.	⟨$X_{[Cl.1]}$, {2sg}⟩	→	X*out*
c.	⟨X, {3sgm}⟩	→	X*añ*
d.	⟨X, {3sgf}⟩	→	X*i*
e.	⟨$X_{[Cl.1]}$, {1pl}⟩	→	X*omp*
f.	⟨X, {2pl}⟩	→	X*oc'h*
g.	⟨X, {3pl}⟩	→	X*o*

Content cells also have realizations; a content cell acquires its realization not directly (by means of realization rules), but indirectly—by association with a form cell and hence with its realization. The form cell with which a content cell is associated is its **form correspondent**. Thus, each of the cells in the content paradigm of the lexeme HERVEZ has, as its form correspondent, a cell in the form paradigm of the stem *hervez*$_{[Cl.1]}$, as in (10); each content cell shares the realization of its form correspondent.

(10) Correspondences between the content paradigm of HERVEZ and the form paradigm of *hervez*[Cl.1]

Content cell	Form correspondent	Shared realization
⟨HERVEZ, {1sg}⟩	⟨*hervez*[Cl.1], {1sg}⟩	*hervezon*
⟨HERVEZ, {2sg}⟩	⟨*hervez*[Cl.1], {2sg}⟩	*hervezout*
⟨HERVEZ, {3sgm}⟩	⟨*hervez*[Cl.1], {3sgm}⟩	*hervezañ*
⟨HERVEZ, {3sgf}⟩	⟨*hervez*[Cl.1], {3sgf}⟩	*hervezi*
⟨HERVEZ, {1pl}⟩	⟨*hervez*[Cl.1], {1pl}⟩	*hervezomp*
⟨HERVEZ, {2pl}⟩	⟨*hervez*[Cl.1], {2pl}⟩	*hervezoc'h*
⟨HERVEZ, {3pl}⟩	⟨*hervez*[Cl.1], {3pl}⟩	*hervezo*

The association of a content cell with its form correspondent is in general specified by a **rule of paradigm linkage**. Such associations are ordinarily effected by means of the universal default rule of paradigm linkage in (11).

(11) The universal default rule of paradigm linkage
Given a lexeme L having X as its σ-stem, the content cell ⟨L, σ⟩ has the form cell ⟨X, σ⟩ as its form correspondent.

Thus, given that the lexeme HERVEZ has *hervez*[Cl.1] as its sole stem (as in (12)), (11) entails that all of the correspondences in (10) will hold by default.

(12) Stem specification for Breton HERVEZ 'according to'
Given any relevant morphosyntactic property set σ, the lexeme HERVEZ has *hervez*[Cl.1] as its σ-stem.

In the definition of canonical paradigms, the default rule in (11) induces instances of paradigm linkage possessing the four characteristics in (13); extending Corbett's typology, I propose that these be regarded as the properties of canonical paradigm linkage.

(13) Canonical paradigm linkage
a. The relation between a lexeme's content cells and their form correspondents is a total function, i.e. every content cell has a form correspondent.
b. All of a lexeme's form correspondents share the same stem, i.e. all are drawn from the same form paradigm.
c. The relation between content cells and their form correspondents is one-to-one rather than many-to-one, i.e. there is no sharing of form correspondents.
d. A content cell's form correspondent is morphosyntactically faithful to it, i.e. it carries the same morphosyntactic property set.

Together, these properties characterize the canonical pattern of paradigm linkage schematized in (14).

(14) Canonical paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	⟨X, τ⟩

If all inflection were purely canonical, the assumption that inflection involves both content paradigms and form paradigms might seem unnecessarily complicated; one

might, for example, propose to eliminate any need for form paradigms (and hence any need for the rule of paradigm linkage in (11)) by formulating the realization rules in (9) directly in terms of content cells rather than in terms of form cells. But once one begins looking at a wider array of inflectional phenomena, the need to distinguish between content paradigms and form paradigms becomes apparent. In particular, there is a range of noncanonical inflectional phenomena involving overrides of the canonical pattern of paradigm linkage in (14).

Instances of paradigm linkage in noncanonical inflection lack one or more of the characteristics in (13), either because the default rule of paradigm linkage cannot apply, or because the application of the default rule of paradigm linkage is overridden by that of a language-specific rule of paradigm linkage, or because the default rule of paradigm linkage itself draws a lexeme's form correspondents from more than one form paradigm. I examine these various deviations from the canonical ideal in the inflectional morphology of the five phenomena in (3) (repeated here as (15)), each of which defines its own pattern of paradigm linkage.

(15) Inflectional phenomena with noncanonical paradigm linkage
   a. Defectiveness
   b. Syncretism
   c. Deponency
   d. Functor-argument reversal
   e. Suppletion

3. Deviations from canonical paradigm linkage

3.1. Defectiveness: lack of a form correspondent

One sort of noncanonical inflection is defectiveness. In the architecture proposed here, a defective lexeme has a content paradigm some of whose cells lack form correspondents (and hence realizations); this pattern of paradigm linkage is schematized in (16). An example is the Latin verb COEPISSE 'begin', which has perfect-system forms (perfect, pluperfect and future perfect) but no present-system forms (present, imperfect and future); the partial paradigm in (17) illustrates. This inflectional peculiarity can be attributed to the absence of a present-system stem in the stem specifications for COEPISSE, as in (18). If there is no present-system stem, then the default rule of paradigm linkage cannot apply to the present-system cells in COEPISSE's content paradigm, leaving these cells without form correspondents and therefore without realizations, as in (19).

(16) Defective paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	—

Cf. Canonical paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	⟨X, τ⟩

(17) The defective inflection of Latin COEPISSE 'begin'

	Present	Perfect
1sg	—	*coepī*
2sg	—	*coepistī*
3sg	—	*coepit*
1pl	—	*coepimus*
2pl	—	*coepistis*
3pl	—	*coepērunt*

(18) Stem specifications for the Latin verb COEPISSE 'begin'

Given any perfect-system (i.e. perfect, pluperfect or future perfect) property set σ, the lexeme COEPISSE has *coep* as its σ-stem.
Given any present-system (i.e. present, imperfect or future) property set σ, COEPISSE lacks a σ-stem.

(19) The content paradigm and form correspondents of Latin COEPISSE 'begin'

Content paradigm	Form correspondents	Shared realization
⟨COEPISSE, {1sg pres indic act}⟩	—	—
⟨COEPISSE, {2sg pres indic act}⟩	—	—
⟨COEPISSE, {3sg pres indic act}⟩	—	—
⟨COEPISSE, {1pl pres indic act}⟩	—	—
⟨COEPISSE, {2pl pres indic act}⟩	—	—
⟨COEPISSE, {3pl pres indic act}⟩	—	—
⟨COEPISSE, {1sg perf indic act}⟩	⟨*coep*, {1sg perf indic act}⟩	*coepī*
⟨COEPISSE, {2sg perf indic act}⟩	⟨*coep*, {2sg perf indic act}⟩	*coepistī*
⟨COEPISSE, {3sg perf indic act}⟩	⟨*coep*, {3sg perf indic act}⟩	*coepit*
⟨COEPISSE, {1pl perf indic act}⟩	⟨*coep*, {1pl perf indic act}⟩	*coepimus*
⟨COEPISSE, {2pl perf indic act}⟩	⟨*coep*, {2pl perf indic act}⟩	*coepistis*
⟨COEPISSE, {3pl perf indic act}⟩	⟨*coep*, {3pl perf indic act}⟩	*coepērunt*
etc.	etc.	etc.

Thus, one sort of noncanonical inflection involves lexemes whose stem specifications are incomplete; the relation between such a lexeme's content cells and their form correspondents is merely a partial function.

3.2. Syncretism: shared form correspondents

In other instances of noncanonical inflection, the default rule of paradigm linkage is applicable, but its application is overridden by a language-specific rule of paradigm linkage. Instances of syncretism involve overrides of this sort. In syncretism, distinct cells in the same content paradigm are realized by the same form cell; this pattern of paradigm linkage is schematized in (20) and (21). Consider, for example, the forms of the Latin noun BELLUM 'war', given in (22); these forms are all based on the stem *bell* specified in (23). There are two instances of syncretism among these forms. BELLUM is a neuter noun, and therefore participates in a general pattern in Latin, in which a neuter noun's nominative forms are always identical to their accusative counterparts; this is a directional syncretism, in the sense that the nominative seems to pattern after the accusative, at least in the singular. In addition, the forms of BELLUM reflect a default pattern of syncretism between the dative and the ablative; this is a nondirectional syncretism. In the proposed architecture of inflectional morphology, these instances of syncretism are accounted for by the special rules of paradigm linkage in (24) and (25), which cause certain content cells to have the same form correspondent and hence the same realization, as in (26).

(20) Syncretic paradigm linkage (directional)

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	

(21) Syncretic paradigm linkage (nondirectional)

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ/τ⟩
⟨L, τ⟩	

Cf. Canonical paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	⟨X, τ⟩

(22) The inflection of Latin BELLUM (neut.) 'war'

	Singular	Plural
Nom	*bellum*	*bella*
Gen	*bellī*	*bellōrum*
Dat	*bellō*	*bellīs*
Acc	*bellum*	*bella*
Abl	*bellō*	*bellīs*

(23) Stem specification for Latin BELLUM 'war'
For any case/number property set σ, the lexeme BELLUM has *bell* as its σ-stem.

(24) Special rule of paradigm linkage for neuter nouns
Where σ = {acc Y} and L is a neuter noun with σ-stem X,
the content cell ⟨L, {nom Y}⟩ has ⟨X, σ⟩ as its form correspondent.

(25) Special default rule of paradigm linkage for dative/ablative forms
Where σ = {dat Y} or {abl Y} and L is a noun with σ-stem X, the content cell ⟨L, σ⟩ has ⟨X, {dat/abl Y}⟩ as its form correspondent.

(26) The content paradigm and form correspondents of Latin BELLUM 'war' (neut.)

Content paradigm	Form correspondents	Shared realization
⟨BELLUM, {nom sg}⟩ ⟨BELLUM, {acc sg}⟩	⟨*bell*, {acc sg}⟩	*bellum*
⟨BELLUM, {gen sg}⟩	⟨*bell*, {gen sg}⟩	*bellī*
⟨BELLUM, {dat sg}⟩ ⟨BELLUM, {abl sg}⟩	⟨*bell*, {dat/abl sg}⟩	*bellō*
⟨BELLUM, {nom pl}⟩ ⟨BELLUM, {acc pl}⟩	⟨*bell*, {acc pl}⟩	*bella*
⟨BELLUM, {gen pl}⟩	⟨*bell*, {gen pl}⟩	*bellōrum*
⟨BELLUM, {dat pl}⟩ ⟨BELLUM, {abl pl}⟩	⟨*bell*, {dat/abl pl}⟩	*bellīs*

3.3. Other morphosyntactically unfaithful form correspondents

3.3.1. Deponency

Another sort of override of the default rule of paradigm linkage is that of deponency, in which form correspondents have seemingly set aside their association with the expected content cells. This pattern of paradigm linkage is schematized in (27). In Latin, for example, many verbs have both active and passive forms, e.g. LAUDĀRE 'praise', *laudat* 's/he praises', *laudātur* 's/he is praised'. Deponent verbs, however, are different: in finite contexts, they are only used in the active voice; yet, they have the form of passives.[2] Thus, only the active cells in the finite content paradigm of the deponent verb HORTĀRĪ 'urge' have form corresondents, yet these form correspondents are specified for passive rather than active morphology. The examples in (28) illustrate.

[2] The participial forms of a deponent verb add an extra dimension of complexity: as with nondeponent verbs, a deponent verb's present and future participles are active in both form and meaning, and its gerundive, passive in both form and meaning. The properties of these participial forms are not at issue here.

(27) Deponent paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, τ⟩
⟨L, τ⟩	—

Cf. Canonical paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	⟨X, τ⟩

(28) The present indicative inflection of Latin LAUDĀRE 'praise' and HORTĀRĪ 'urge'

LAUDARE 'praise'		HORTARI 'urge'	
Active	Passive	Active	Passive
laudō	*laudor*	*hortor*	—
laudās	*laudāris*	*hortāris*	—
laudat	*laudātur*	*hortātur*	—
laudāmus	*laudāmur*	*hortāmur*	—
laudātis	*laudāminī*	*hortāminī*	—
laudant	*laudantur*	*hortantur*	—

These peculiarities of deponent verbs are accounted for by the special rule of paradigm linkage in (29) together with stem specifications such as those in (30a): the rule of paradigm linkage in (29) causes a deponent verb's active content cells to have passive form correspondents; stem specifications such as (30a) cause a deponent verb's passive content cells to lack form correspondents. Thus, assuming the additional stem specification in (30b), (29) and (30) together induce the patterns of paradigm linkage exemplified in (31).[3]

(29) Special rule of paradigm linkage for deponents
Where σ is a finite active property set with τ as its passive counterpart and L is a deponent verb with τ-stem X, the content cell ⟨L, σ⟩ has ⟨X, τ⟩ as its form correspondent.

(30) Stem specifications for Latin HORTĀRĪ 'urge' and LAUDĀRE 'praise'
a. Given any finite passive property set σ, the lexeme HORTĀRĪ has *hortā* as its σ-stem.
Given any finite active property set σ, the lexeme HORTĀRĪ lacks a σ-stem.
b. Given any property set σ, the lexeme LAUDĀRE has *laudā* as its σ-stem.

(31) The content paradigms and form correspondents of Latin LAUDĀRE 'praise' and HORTĀRĪ 'urge'

Content paradigm	Form correspondents	Shared realization
⟨LAUDĀRE, {1sg pres indic act}⟩	⟨*laudā*, {1sg pres indic act}⟩	*laudō*
⟨LAUDĀRE, {2sg pres indic act}⟩	⟨*laudā*, {2sg pres indic act}⟩	*laudās*

[3] In Latin, the finite perfect passive forms of a nondeponent verbal lexeme L are periphrastic, consisting of an active present-system form of the auxiliary verb ESSE 'be' and a nominative form of L's perfect passive participle. Adopting the approach to inflectional periphrasis proposed by Bonami & Webelhuth (2010), I assume that in Latin, a nondeponent verbal lexeme L's finite perfect passive realizations are forms of ESSE that require a perfect passive participial form of L as a complement. In formal terms: where τ_1 is a passive perfect-system property set with τ_2 as its active present-system counterpart and X is the τ_1-stem of a verbal lexeme L, a rule of referral causes the realization of the form cell ⟨X, τ_1⟩ to be that of ⟨*es*, τ_2⟩, and this realization takes L's perfect passive participle as a complement. This analysis correctly entails that the perfect active cells in a deponent verb's content paradigm will have the same periphrastic expression as the perfect passive cells of a nondeponent verb's content paradigm.

⟨LAUDĀRE, {3sg pres indic act}⟩	⟨*laudā*, {3sg pres indic act}⟩	*laudat*
⟨LAUDĀRE, {1sg pres indic pass}⟩	⟨*laudā*, {1sg pres indic pass}⟩	*laudor*
⟨LAUDĀRE, {2sg pres indic pass}⟩	⟨*laudā*, {2sg pres indic pass}⟩	*laudāris*
⟨LAUDĀRE, {3sg pres indic pass}⟩	⟨*laudā*, {3sg pres indic pass}⟩	*laudātur*
etc.	etc.	etc.
⟨ HORTĀRĪ, {1sg pres indic act}⟩	⟨ *hortā*, {1sg pres indic pass}⟩	*hortor*
⟨ HORTĀRĪ, {2sg pres indic act}⟩	⟨ *hortā*, {2sg pres indic pass}⟩	*hortāris*
⟨ HORTĀRĪ, {3sg pres indic act}⟩	⟨ *hortā*, {3sg pres indic pass}⟩	*hortātur*
⟨ HORTĀRĪ, {1sg pres indic pass}⟩	—	—
⟨ HORTĀRĪ, {2sg pres indic pass}⟩	—	—
⟨ HORTĀRĪ, {3sg pres indic pass}⟩	—	—
etc.	etc.	etc.

3.3.3. Functor-argument reversal

A final kind of override of the default rule of paradigm linkage arises in instances of what Spencer & Stump 2011 call functor-argument reversal. This phenomenon, schematized in (32), is dramatically exemplified by Hungarian pronominal case forms. In Hungarian, nouns exhibit possessor marking, as in (33).

(32) Paradigm linkage with functor-argument reversal

Content cell	Form correspondent
⟨L, σ⟩	⟨*f*(σ), *g*(L)⟩

(33) Possessor marking on two Hungarian nouns (in the nominative)

Possessor	KÖNYV 'book'		HAZ 'house'	
	Possessee		Possessee	
	Singular	Plural	Singular	Plural
1sg	*könyv-e-m*	*könyv-e-i-m*	*ház-a-m*	*ház-a-i-m*
2sg	*könyv-e-d*	*könyv-e-i-d*	*ház-a-d*	*ház-a-i-d*
3sg	*könyv-e*	*könyv-e-i*	*ház-a*	*ház-a-i*
1pl	*könyv-ünk*	*könyv-e-i-nk*	*ház-unk*	*ház-a-i-nk*
2pl	*könyv-e-tek*	*könyv-e-i-tek*	*ház-a-tok*	*ház-a-i-tok*
3pl	*könyv-ük*	*könyv-e-i-k*	*ház-uk*	*ház-a-i-k*

This same morphology is used to express pronominal objects of postpositions, as in (34).

(34) Object-agreement paradigm of the Hungarian postposition MÖGÖTT 'behind'

mögött-e-m	'behind me'
mögött-e-d	'behind you (sg.)'
mögött-e	'behind her/him'
mögött-ünk	'behind us'
mögött-e-tek	'behind you (pl.)'
mögött-ük	'behind them'

In the Hungarian case system, case is marked suffixally on nouns, with the nominative case being unmarked; the nouns in (35) illustrate.

(35) Case forms of two Hungarian nouns

Case	KÖNYV 'book'		HAZ 'house'		Gloss
	Singular	Plural	Singular	Plural	
Nominative	*könyv*	*könyv-e-k*	*ház*	*ház-a-k*	(subject)
Accusative	*könyv-e-t*	*könyv-e-k-e-t*	*ház-a-t*	*ház-a-k-a-t*	(direct object)

Dative	*könyv-nek*	*könyv-e-k-nek*	*ház-nak*	*ház-a-k-nak*	(indirect object)
Inessive	*könyv-ben*	*könyv-ek-ben*	*ház-ban*	*ház-ak-ban*	'in a book …'
Illative	*könyv-be*	*könyv-ek-be*	*ház-ba*	*ház-ak-ba*	'into a book …'
Elative	*könyv-ből*	*könyv-ek-ből*	*ház-ból*	*ház-ak-ból*	'from inside of a book …'
Superessive	*könyv-ön*	*könyv-ek-en*	*ház-on*	*ház-ak-on*	'on a book …'
Sublative	*könyv-re*	*könyv-ek-re*	*ház-ra*	*ház-ak-ra*	'onto a book …'
Delative	*könyv-ről*	*könyv-ek-ről*	*ház-ról*	*ház-ak-ról*	'from the surface of a book …'
Adessive	*könyv-nél*	*könyv-ek-nél*	*ház-nál*	*ház-ak-nál*	'at a book …'
Allative	*könyv-höz*	*könyv-ek-hez*	*ház-hoz*	*ház-ak-hoz*	'towards a book …'
Ablative	*könyv-től*	*könyv-ek-től*	*ház-tól*	*ház-ak-tól*	'from (near) a book …'
Instrumental	*könyv-vel*	*könyv-ek-kel*	*ház-zal*	*ház-ak-kal*	'with a book …'
Causal-final	*könyv-ért*	*könyv-ek-ért*	*ház-ért*	*ház-ak-ért*	'as a book …'
Translative	*könyv-vé*	*könyv-ek-ké*	*ház-zá*	*ház-ak-ká*	'(change) into a book …'
Terminative	*könyv-ig*	*könyv-ek-ig*	*ház-ig*	*ház-ak-ig*	'up to a book …'
Essive-formal	*könyv-ként*	*könyv-ek-ként*	*ház-ként*	*ház-ak-ként*	'in the capacity of a book …'
Essive	*könyv-ül*	—	*ház-ul*	—	'in the capacity of a book …'

In the direct cases, pronouns exhibit a similar sort of inflection. In the oblique cases, however, pronominal case is expressed by case postpositions inflected for person and number, as in (36).

(36) Case forms of Hungarian personal pronouns

		First person	Second person	Third person
Singular	nominative	*én*	*te*	*ő*
	accusative	*engem(et)*	*téged(et)*	*őt*
	dative	*nek-e-m*	*nek-e-d*	*nek-i*
	inessive	*benn-e-m*	*benn-e-d*	*benn-e*
	illative	*bel-é-m*	*bel-é-d*	*bel-e, bel-é(-je)*
	elative	*belől-e-m*	*belől-e-d*	*belől-e*
	superessive	*rajt-a-m*	*rajt-a-d*	*rajt-a*
	sublative	*rá-m*	*rá-d*	*rá(-ja)*
	delative	*ról-a-m*	*ról-a-d*	*ról-a*
	adessive	*nál-a-m*	*nál-a-d*	*nál-a*
	allative	*hozz-á-m*	*hozz-á-d*	*hozz-á(-ja)*
	ablative	*től-e-m*	*től-e-d*	*től-e*
	instrumental	*vel-e-m*	*vel-e-d*	*vel-e*
	causal-final	*ért-e-m*	*ért-e-d*	*ért-e*
Plural	nominative	*mi*	*ti*	*ők*
	accusative	*minket ~ bennünket*	*titeket ~ benneteket*	*őket*
	dative	*nek-ünk*	*nek-tek*	*nek-i-k*
	inessive	*benn-ünk*	*benn-e-tek*	*benn-ük*
	illative	*bel-é-nk*	*bel-é-tek*	*bel-é-jük*
	elative	*belől-ünk*	*belől-e-tek*	*belől-ük*
	superessive	*rajt-unk*	*rajt-a-tok*	*rajt-uk*
	sublative	*rá-nk*	*rá-tok*	*rá-juk*
	delative	*ról-unk*	*ról-a-tok*	*ról-uk*
	adessive	*nál-unk*	*nál-a-tok*	*nál-uk*
	allative	*hozz-á-nk*	*hozz-á-tok*	*hozz-á-juk*
	ablative	*től-ünk*	*től-e-tek*	*től-ük*
	instrumental	*vel-ünk*	*vel-e-tek*	*vel-ük*
	causal-final	*ért-ünk*	*ért-e-tek*	*ért-ük*

Thus, the inflection of pronominal case in Hungarian involves a kind of functor-argument reversal. This pattern of paradigm linkage is accounted for by the special rule in (37). To a content cell consisting of a pronominal lexeme paired with an oblique case property,

rule (37) assigns a form correspondent consisting of the appropriate case stem paired with the appropriate set of pronominal properties; in this way, rule (37) defines instances of paradigm linkage such as those in (38). In (38), the inessive case of the first-person singular lexeme ÉN is inflected as the first-person singular form of the inessive stem *benn*; similarly, the superessive case of the first-person singular lexeme ÉN is inflected as the first-person singular form of the superessive stem *rajt*; and so on.

(37) Rule of paradigm linkage for oblique pronominal case forms
If L is a pronominal lexeme expressing person a and number b and X is a postpositional stem expressing oblique case c, the content cell ⟨L, {c}⟩ has ⟨X, {a b}⟩ as its form correspondent.

(38) The content paradigm and form correspondents of the Hungarian 1sg pronoun ÉN

Content paradigm	Form correspondents	Shared realization
⟨EN, {nominative}⟩	⟨*én*, {nominative}⟩	*én*
⟨EN, {accusative}⟩	⟨*én*, {accusative}⟩	*engem(et)*
⟨EN, {dative}⟩	⟨*nek*, {1sg}⟩	*nekem*
⟨EN, {inessive}⟩	⟨*benn*, {1sg}⟩	*bennem*
⟨EN, {illative}⟩	⟨*bel*, {1sg}⟩	*belém*
⟨EN, {elative}⟩	⟨*belől*, {1sg}⟩	*belőlem*
⟨EN, {superessive}⟩	⟨*rajt*, {1sg}⟩	*rajtam*
⟨EN, {sublative}⟩	⟨*rá*, {1sg}⟩	*rám*
⟨EN, {delative}⟩	⟨*ról*, {1sg}⟩	*rólam*
⟨EN, {adessive}⟩	⟨*nál*, {1sg}⟩	*nálam*
⟨EN, {allative}⟩	⟨*hozz*, {1sg}⟩	*hozzám*
⟨EN, {ablative}⟩	⟨*től*, {1sg}⟩	*tőlem*
⟨EN, {instrumental}⟩	⟨*vel*, {1sg}⟩	*velem*
⟨EN, {causal-final}⟩	⟨*ért*, {1sg}⟩	*értem*

Summarizing, the default rule of paradigm linkage may be overridden in instances of syncretism, deponency, and functor-argument reversal; each of these defines its own particular pattern of override, as in (39).

(39) Overrides of the default rule of paradigm linkage

Where L is a lexeme with stem X and σ, σ′ are distinct morphosyntactic property sets associated with L,

a.	L's inflection exhibits SYNCRETISM:	⟨L, σ⟩ and ⟨L, σ′⟩ have the same form correspondent.
b.	L's inflection is DEPONENT:	⟨X, σ′⟩ is the form correspondent of ⟨L, σ⟩ rather than of ⟨L, σ′⟩.
c.	L's inflection exhibits FUNCTOR-ARGUMENT REVERSAL:	the form correspondent of ⟨L, σ⟩ is ⟨*f*(σ), *g*(L)⟩.

3.4. Suppletion: form correspondents drawn from distinct form paradigms

A final noncanonical pattern of paradigm linkage arises in the inflection of lexemes exhibiting stem suppletion. In instances of suppletion, the cells in a lexeme's content paradigm have form correspondents based on two (or more) arbitrarily different stems; this pattern of paradigm linkage is schematized in (40). In Latin, for example, present-system cells in the content paradigm of the lexeme FERRE 'carry' have form

correspondents based on the stem *fer*; but the perfect-system cells in this content paradigm have form correspondents based on the stem *tul*. The examples in (41) illustrate.

(40) Suppletive paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X_1, σ⟩
⟨L, τ⟩	⟨X_2, τ⟩

Cf. Canonical paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	⟨X, τ⟩

(41) The suppletive inflection of Latin FERRE 'carry'

	Present	Perfect
1sg	*ferō*	*tulī*
2sg	*fers*	*tulistī*
3sg	*fert*	*tulit*
1pl	*ferimus*	*tulimus*
2pl	*fertis*	*tulistis*
3pl	*ferunt*	*tulērunt*

In the proposed architecture for inflectional morphology, this suppletive pattern of paradigm linkage arises not because of any override of the default rule of paradigm linkage, but simply because the stem specifications for FERRE give two suppletive stems in complementary distribution, as in (42). Given these specifications, the default rule of paradigm linkage produces the patterns of correspondence exemplified in (43).

(42) Stem specifications for Latin FERRE 'carry'
Given any present-system (i.e. present, imperfect or future) property set σ, the lexeme FERRE has *fer*[Third conjugation] as its σ-stem.
Given any perfect-system (i.e. perfect, pluperfect or future perfect) property set σ, the lexeme FERRE has *tul* as its σ-stem.

(43) The content paradigm and form correspondents of Latin FERRE 'carry'

Content paradigm	Form correspondents	Shared realization
⟨FERRE, {1sg pres indic act}⟩	⟨*fer*, {1sg pres indic act}⟩	*ferō*
⟨FERRE, {2sg pres indic act}⟩	⟨*fer*, {2sg pres indic act}⟩	*fers*
⟨FERRE, {3sg pres indic act}⟩	⟨*fer*, {3sg pres indic act}⟩	*fert*
⟨FERRE, {1pl pres indic act}⟩	⟨*fer*, {1pl pres indic act}⟩	*ferimus*
⟨FERRE, {2pl pres indic act}⟩	⟨*fer*, {2pl pres indic act}⟩	*fertis*
⟨FERRE, {3pl pres indic act}⟩	⟨*fer*, {3pl pres indic act}⟩	*ferunt*
⟨FERRE, {1sg perf indic act}⟩	⟨*tul*, {1sg perf indic act}⟩	*tulī*
⟨FERRE, {2sg perf indic act}⟩	⟨*tul*, {2sg perf indic act}⟩	*tulistī*
⟨FERRE, {3sg perf indic act}⟩	⟨*tul*, {3sg perf indic act}⟩	*tulit*
⟨FERRE, {1pl perf indic act}⟩	⟨*tul*, {1pl perf indic act}⟩	*tulimus*
⟨FERRE, {2pl perf indic act}⟩	⟨*tul*, {2pl perf indic act}⟩	*tulistis*
⟨FERRE, {3pl perf indic act}⟩	⟨*tul*, {3pl perf indic act}⟩	*tulērunt*
etc.	etc.	etc.

The various noncanonical patterns of paradigm linkage are not mutually exclusive; complex noncanonical patterns may arise. For example, suppletion and deponency coincide in the inflection of Old Icelandic preterite-present verbs. They are deponent in that they form their present tense as a strong verb ordinarily forms its past tense; they are suppletive in that they form their past tense with a separate, weak stem. The examples of ÞURFA 'need' in (44) illustrate. These forms imply the pattern of paradigm linkage schematized in (45) and exemplified in (46).

(44) Indicative and subjunctive paradigms of three Old Icelandic verbs
(Shaded forms of ÞURFA inflect like shaded forms of BRENNA;
heavy-bordered boxes enclose forms of ÞURFA and DUGA that inflect alike.)

				Strong BRENNA 'burn'	Preterite-Present ÞURFA 'need'	Weak DUGA 'help'
Indicative	Pres.	Sg.	1	*brenn*	*þarf*	*dug-i*
			2	*brenn-r*	*þarf-t*	*dug-ir*
			3	*brenn-r*	*þarf*	*dug-ir*
		Pl.	1	*brenn-um*	*þurf-um*	*dug-um*
			2	*brenn-ið*	*þurf-uð*	*dug-ið*
			3	*brenn-a*	*þurf-u*	*dug-a*
	Past	Sg.	1	*brann*	*þurf-ta*	*dug-ða*
			2	*brann-t*	*þurf-tir*	*dug-ðir*
			3	*brann*	*þurf-ti*	*dug-ði*
		Pl.	1	*brunn-um*	*þurf-tum*	*dug-ðum*
			2	*brunn-uð*	*þurf-tuð*	*dug-ðuð*
			3	*brunn-u*	*þurf-tu*	*dug-ðu*
Subjunctive	Pres.	Sg.	1	*brenn-a*	*þurf-a*	*dug-a*
			2	*brenn-ir*	*þurf-ir*	*dug-ir*
			3	*brenn-i*	*þurf-i*	*dug-i*
		Pl.	1	*brenn-im*	*þurf-im*	*dug-im*
			2	*brenn-ið*	*þurf-ið*	*dug-ið*
			3	*brenn-i*	*þurf-i*	*dug-i*
	Past	Sg.	1	*brynn-a*	*þyrf-ta*	*dyg-ða*
			2	*brynn-ir*	*þyrf-tir*	*dyg-ðir*
			3	*brynn-i*	*þyrf-ti*	*dyg-ði*
		Pl.	1	*brynn-im*	*þyrf-tim*	*dyg-ðim*
			2	*brynn-ið*	*þyrf-tið*	*dyg-ðið*
			3	*brynn-i*	*þyrf-ti*	*dyg-ði*

Source: Zoëga 1910.

(45) Paradigm linkage with deponency plus suppletion

Content cell	Form correspondent
⟨L, σ⟩	⟨X_1, τ⟩
⟨L, τ⟩	⟨X_2, τ⟩

Cf. Canonical paradigm linkage

Content cell	Form correspondent
⟨L, σ⟩	⟨X, σ⟩
⟨L, τ⟩	⟨X, τ⟩

(46) The content paradigm and form correspondents of the Old Icelandic preterite-present verb ÞURFA 'need'

Content paradigm	Form correspondents	Shared realization
⟨ÞURFA, {1sg pres indic}⟩	⟨*þarf*$_{[Strong]}$, {1sg past indic}⟩	*þarf*
⟨ÞURFA, {2sg pres indic}⟩	⟨*þarf*$_{[Strong]}$, {2sg past indic}⟩	*þarft*
⟨ÞURFA, {3sg pres indic}⟩	⟨*þarf*$_{[Strong]}$, {3sg past indic}⟩	*þarf*
etc.	etc.	etc.
⟨ÞURFA, {1sg past indic}⟩	⟨*þurf*$_{[Weak]}$, {1sg past indic}⟩	*þurfta*
⟨ÞURFA, {2sg past indic}⟩	⟨*þurf*$_{[Weak]}$, {2sg past indic}⟩	*þurftir*
⟨ÞURFA, {3sg past indic}⟩	⟨*þurf*$_{[Weak]}$, {3sg past indic}⟩	*þurfti*
etc.	etc.	etc.

4. Discussion

The proposed architecture of inflection has several important consequences. It clarifies the difference between morphological properties that have syntactic relevance (e.g. properties of case, tense, etc.) and purely morphological properties (e.g. properties of inflection-class membership): while the former are represented in both content paradigms and form paradigms, the latter are restricted to form paradigms. This means that because content cells are the interface of inflectional morphology with syntax and semantics, a word's syntax and semantics are invariably insensitive to its inflection-class membership.

Because the proposed architecture allows a content cell to have a different morphosyntactic property set from its form correspondent, it correctly entails that a word can have different morphosyntactic property sets for different purposes; specifically, it allows the morphosyntactic property set that determines exponence to differ from the set that determines lexical insertion and semantic interpretation.

The proposed architecture provides an explicit account of how canonical and noncanonical inflection differ: the former conforms to the canonical pattern of paradigm linkage in (14), while the latter invariably deviates from this pattern. The details of this distinction can be articulated in terms of the relation between content cells and form cells, as in (47).

(47) Relations between content cells and form cells

	Canonical inflection	Noncanonical inflection
The relation of a lexeme's content cells to their form correspondents	is a total function [every content cell has a form correspondent]	may be a partial function (as in cases of defectiveness)
	has a single form paradigm as its range	may have many form paradigms in its range (as in cases of suppletion)
	is one-to-one [each content cell has a form correspondent entirely to itself]	may be many-to-one (as in cases of syncretism) [sharing]
A content cell's form correspondent	is morphosyntactically faithful	may not be faithful (as in cases of deponency, functor-argument reversal and directional syncretism)

Finally, the proposed architecture implies the existence of virtual cells—that is, form cells that aren't form correspondents. The realization rules define virtual realizations for these cells. These virtual realizations may emerge as "exploratory expressions" (cf. (48)) in language change. Consider, for example, the case of the Latin deponent verb HORTĀRĪ 'urge'. In Classical Latin, the active form cells in (49) are merely virtual; even so, they serve as the basis for the active realizations in (50). These too are merely virtual, since they don't serve as the realization of any content cell. In late Latin, active forms of this verb begin to appear, as in (51) and (52) (Hippisley 2010). Assuming the present architecture, such forms existed "all along" as virtual realizations of form cells that didn't serve as form correspondents for any content cell. Their emergence is not the result of adding any new rules nor of changing any existing rules; rather, it is the inevitable effect of simply suppressing an override of the default rule of paradigm linkage.

(48) By **exploratory expressions** we mean expressions which are introduced through the ordinary operation of the grammar and which "catch on" and become fixed expressions and eventually are grammaticalized. […] It appears that most initial exploratory expressions are made by applying the rules of grammar in a regular way, but it may be that some perhaps also involve ignoring (breaking) existing rules of grammar. (Harris & Campbell 1995: 73)

(49) Imperfect indicative form paradigm of *hortā*

Active (virtual)	Passive
⟨*hortā*, {1sg impf indic act}⟩	⟨*hortā*, {1sg impf indic pass}⟩
⟨*hortā*, {2sg impf indic act}⟩	⟨*hortā*, {2sg impf indic pass}⟩
⟨*hortā*, {3sg impf indic act}⟩	⟨*hortā*, {3sg impf indic pass}⟩
⟨*hortā*, {1pl impf indic act}⟩	⟨*hortā*, {1pl impf indic pass}⟩
⟨*hortā*, {2pl impf indic act}⟩	⟨*hortā*, {2pl impf indic pass}⟩
⟨*hortā*, {3pl impf indic act}⟩	⟨*hortā*, {3pl impf indic pass}⟩

(50) Imperfect indicative realizations of *hortā*

	Active (virtual)	Passive
1sg	*hortābam*	*hortābar*
2sg	*hortābās*	*hortābāris*
3sg	*hortābat*	*hortābātur*
1pl	*hortābāmus*	*hortābāmur*
2pl	*hortābātis*	*hortābāminī*
3pl	*hortābant*	*hortābantur*

(51) *Hortabat* *caeteros* *Apostolus*
urge.3PL.IMPF.ACT other.ACC.PL Apostle.NOM.SG
"The Apostle exhorted the others…" Luculentius, *Commentary on Romans XII 6*; 5-6 c. AD

(52) *Petrus* *hortabat* *eos* *de* *cruce*
Peter.NOM.SG exhort.3SG.IMPF.ACT PRON.3PL.ACC from cross.ABL.SG
'Peter exhorted them [speaking down] from the cross.' [Gregorian chant, 10th c. AD]

The ultimate conclusion of the ideas developed here is in a sense unsurprising: just as words can be seen as concrete units of morphological form or as abstract units of grammatical analysis, so paradigms exhibit a similar dichotomy. Content paradigms specify the range of syntactic contexts in which a lexeme may appear; both the lexical insertion and the semantic interpretation of a given realization are sensitive to the content cell that it realizes. Form paradigms specify the distinctions to which rules of inflectional exponence are sensitive; they determine a lexeme's inventory of inflected forms. In canonical inflection, the two sorts of paradigm are parallel, but in noncanonical inflection, they exhibit a variety of mismatches.

References

Bonami Olivier and Gert Webelhuth 2010. The phrase-structure diversity of periphrasis: A lexicalist account. Paper presented at the Conference on the Typology of Periphrasis, University of Surrey, June 7-8, 2010.

Corbett Greville G. 2005. The canonical approach in typology. In Frajzyngier, Zygmunt, Adam Hodges and David S. Rood (eds.) *Linguistic diversity and language theories* (Studies in Language Companion Series 72), 25-49. Amsterdam : John Benjamins.

Corbett Greville G. 2009. Canonical inflectional classes. In Montermini, Fabio, Gilles Boyé and Jesse Tseng (eds.) *Selected proceedings of the 6th Décembrettes*, 1-11. Somerville, MA: Cascadilla Proceedings Project.

Harris Alice C. and Lyle Campbell 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.

Hippisley Andrew 2010. Paradigmatic realignment and morphological change. In Rainer, Franz, Wolfgang U. Dressler, Dieter Kastovsky and Hans Christian Luschützky (eds.) *Variation and change in morphology*, 107-27. Amsterdam: John Benjamins.

Spencer Andrew and Gregory Stump 2011. Hungarian pronominal case and the dichotomy of content and form in inflectional morphology. Manuscript.

Stewart Tom and Gregory Stump 2007. Paradigm Function Morphology and the morphology/syntax interface. In Ramchand, Gillian and Charles Reiss (eds.) *The Oxford handbook of linguistic interfaces*, 383-421. Oxford: Oxford University Press.

Stump Gregory 2002. Morphological and syntactic paradigms: Arguments for a theory of paradigm linkage. In Booij, Geert and J. van Marle (eds.) *Yearbook of Morphology 2001*, 147-180. Dordrecht: Kluwer.

Stump Gregory 2006. Heteroclisis and paradigm linkage. *Language* 82, 279-322.

Stump Gregory 2007. A noncanonical pattern of deponency and its theoretical implications. In Baerman, Matthew, Greville G. Corbett, Dunstan Brown, & Andrew Hippisley (eds.) *Deponency and morphological mismatches: Proceedings of the British Academy* 145, 71-95. Oxford: Oxford University Press.

Zoëga Geir T. 1910. *A concise dictionary of Old Icelandic*. London: Oxford University Press.

Encoding Unexpectedness by Aspect Inflection*

Niina Ning Zhang
National Chung Cheng University
Lngnz@ccu.edu.tw

1. Introduction

Mirativity is a grammatical category that denotes unexpectedness (e.g., DeLancey 2001). In this paper, I argue that the perfective aspect marker *le* in Mandarin Chinese can be used as a mirative marker, denoting unexpectedness from the viewpoint of a reference world. In Reading A of (1a), *le* encodes mirativity, and the reference world overlaps with the mental world of the speaker in this example. (1b) shows that if *le* does not show up with the same string of words, no mirativity reading is available.

(1) a. Zhe gen shengzi duan-le san gongfen.[1]
this CL rope short-PRF three centimeter
A. 'This rope is three centimeters shorter than expected.'
B. 'This rope becomes three centimeters shorter than before.'
C. 'This rope is three centimeters shorter (than another one in the context).'
這根繩子短了三公分。

b. Zhe gen shengzi duan san gongfen.
this CL rope short three centimeter
only reading: = reading C above.
這根繩子短三公分。

For a mirative reading, the measure expression such as *san gongfen* 'three centimeters' is not obligatory. (2a) and (2b) have no measure expression, and both still have a mirative reading. Note that in all of verbal *le* sentences, a degree achievement reading, as Reading B in (1a), is also available, but we will discuss the mirative reading only.

(2) a. Zhe gen shengzi duan-le.
this CL rope short-PRF
'This rope is too short.'
這根繩子短了。

b. Tang xian-le.
soup salty-PRF
'The soup is too salty.'
湯鹹了。

* For comments on earlier versions of this paper, I thank the audiences of the Eighth Mediterranean Morphology Meeting (MMM 8), Cagliari, Italy, Sept. 14-17, 2011, and the Seventh Conference of the European Association of Chinese Linguistics (EACL-7), Venice, Italy, Sept. 13-15, 2011. I am especially grateful to the following people for their helpful suggestions and challenges: Michael Erlewine, Yuanlu Chen, Guglielmo Cinque, Davide Fanciullo, Daniel Hole, James Huang, James Myers, Victor Pan, Waltraud Paul, Martin Schäfer, and Hunter Wu. I am responsible for all remaining errors.

[1] Abbreviations used in the Chinese examples: PRF: perfect aspect; CL: classifier; DE: associative marker; BA: causative marker; PRT: sentence-final aspect or clause-type particle.

The reference world of the mirativity can also be different from that of the speaker. It is Lulu's mental world in (3a), and an earlier mental stage of the speaker in (3b). In either case, the predicate where *le* occurs functions as an evaluative comment on the topic nominal, from the perspective of the reference world. According to Chang (2009: 2243), this special use of *le* occurs in topic-comment constructions only. He calls the *le* 'comment LE'.

(3) a. Lulu juede zhe gen shengzi chang-le yidian.
Lulu feel this CL rope short-PRF a.little
'Lulu feels that this rope is a little bit longer than expected.'
露露覺得這根繩子短了一點。

b. Wo cengjing yiwei zhe gen shengzi chang-le yidian. (shiji-shang bu shi)
I once think this CL rope short-PRF a.little fact-on not be
'I once thought that this rope was a little bit longer than expected. (in fact it is not.)'
我曾經以為這根繩子短了一點。(實際上不是。)

I will call the use of the aspect marker *le* in Reading A of (1a), (2), and (3) Mirative LE (M-LE), and the canonical perfect aspect use of *le*, as in Reading B of (1a), Temporal LE (T-LE). We will discuss the *le* in Reading C of (1a) in 4.2.

Chao (1968) states that this "use of *le* after adjectives is to express excess over some expected norm" (p. 692; also p. 89). A similar statement is found in Lü et al. (1999 [1980]: 355 "不表示有什麼變化，只表示某一性質偏離標準") and Shi (1988: 105-107). This use of *le* has also been mentioned in Huang (1987), Ljungqvist (2003; 2007), Chang (2009), and Chen & Shirai (2010: 26 fn. 9). Nevertheless, M-LE has received very little attention in the literature.

The research questions of the paper are:

A. Why may the perfect aspect marker *le* bring about this special reading?

B. How is this reading represented syntactically?

I will argue for a parallelism between counterfactual (CF) morphology and counter-expectedness (CE) morphology. I will show that the feature [exclusion] covers both the fake past tense in CF and the fake perfect aspect in CE.

The paper is organized as follows. In Section 2, I introduce the basic properties of M-LE constructions. In Section 3, I review previous analyses of M-LE. Then in Section 4 M-LE is compared with other non-temporal uses of *le* in Mandarin Chinese. A further comparison between M-LE and T-LE is made in Section 5. In this section, question B above is answered. In Section 6, linking to the use of past tense in CF constructions, I address question A above. Section 7 concludes the paper.

2. Basic properties of M-LE constructions

2.1. Individual-Level Predicates

As noted in Chen & Shirai (2010: 26 fn. 9), M-LE occurs with "stative verbs" or "verb phrases or sentences describing nondynamic situations". Precisely speaking, M-LE occurs with only individual-level predicates, such as *zai* 'narrow' in (4a), *tian* 'sweet' in (4b), and *gui* 'expensive' in (4c).

(4) a. Chuang zhai-le. Shui-bu-xia liang ge ren. (individual-level state)
bed narrow-PRF sleep-not-down two CL person
'The bed is too narrow. It cannot sleep two persons.'
床窄了。

b. Nai-cha tian-le.
milk-tea sweet-PRF
'The milk-tea is too sweet.'
奶茶甜了。

c. Zhe dongxi gui-le. (Chao 1968: 692)
this thing expensive-PRF
'This thing is too expensive.'
這東西貴了。

If the predicate is a stage-level predicate, only T-LE may occur. In all of the examples in (5), the predicate is a stage-level one and thus the *le* is not an M-LE.

(5) a. Yanhui kan-le yi bu dianying. (transitive)
Yanhui see-PRF one CL movie
'Yanhui has seen a movie.'
燕慧看了一部電影。

b. Yanhui lai-le. (unaccusative)
Yanhui come-PRF
'Yanhui has come.'
燕慧來了。

c. Yanhui shui-le san tian. (unergative)
Yanhui sleep-PRF three day
'Yanhui has slept for three days.'
燕慧睡了三天。

d. Wendu jiang-le shi du. (degree achievement)
temperature fall-PRF ten degree
'The temperature has fallen ten degrees.'
溫度降了十度。

e. Yanhui zui-le zhengzheng yi shangwu. (stage-level state)
Yanhui drunk-PRF whole one morning
'Yanhui has been drunk for the whole morning.'
燕慧醉了整整一上午。

f. Lulu e-le (san tian). (stage-level state)
Lulu hungry-PRF three day
'Lulu has been hungry (for three days).'

Not only verb types, but also the context needs to be considered. M-LE may not occur in a boundary context. Specifically, it may not occur with a durative adverbial, verb-reduplication, and frequency adverbial. All of these three contexts signal the so-called "provided temporal endpoint" of a situation (Yang 2011). In (5c), (5e), and (5f) above, the T-LE is compatible with a durative adverbial, but the intended M-LE in (6) is not compatible with such an adverbial. In (7a), the verb *zou* 'walk' with the T-LE is reduplicated, but the predicate *duan* 'short', with the intended M-LE in (7b), may not be reduplicated. In (8a), the repetitive adverbial *san ci* 'three times' is compatible with the T-LE, whereas the M-LE in (8b) is not compatible with the adverbial.

(6) *Zhe gen shengzi duan-le san nian.
this CL rope short-PRF three year

(7) a. Ta zai gongyuan-li zou-le-zou.

s/he at park-in walk-PRF-walk
'S/he walked a bit in the park.'

b. *Zhe gen shengzi duan-le-duan.
this CL rope short-PRF-short

(8) a. Ta ke-le san ci.
s/he cough-PRF three time
'S/he coughed three times.'

b. *Zhe gen shengzi duan-le san ci.
this CL rope short-PRF three time

M-LE is thus different from T-LE, which must occur with a dynamic predicate (Shen 2004) and "is not compatible with [-telic] situations unless endpoints of other types are provided" (Yang 2011: Sec. 3). M-LE is used in [-telic] situations only.

Accordingly, if a *le* sentence is ambiguous between a temporal and mirative reading, such as Reading A and Reading B of (9a), the occurrence of a stage-level predicate marker, such as *xialai* 'down' in (9b), excludes the latter reading:

(9) a. Niunai leng-le.
milk cold-PRF
A: 'The milk has become cold.' (Temporal)
B: 'The milk is too cold.' (Non-temporal, mirative)
牛奶冷了。

b. Niunai leng-xialai-le.
milk cold-down-PRF
'The milk has become cold.' (= Reading A above)
牛奶冷下來了。

The subject of an M-LE construction can be any nominal that may occur as the subject of an individual level predicate in Mandarin Chinese, including a proper name, as in (10a), the inner subject of the so-called double-subject constructions (Zhang 2009), as in (10b) and (10c), kind-denoting nominal phrase, as in (10d), and a quantity-denoting expression (Li 1998), as in (10e).

(10) a. Lundun da-le. Zhao yi ge xiao yidian de chengshi! (proper name)
London big-PRF seek one CL small a.bit DE city
'London is too big. Look for a smaller city!'
倫敦大了。找一個小一點的城市！

b. Zhe haizi yanjing xiao-le. (inner subject: relational noun)
this kid eye small-PRF
'This kid, his eyes are too small.'
這孩子眼睛小了。

c. Zhe ge xigua pi hou-le. (inner subject: relational noun)
this CL watermelon skin thick-PRF
'This watermelon, its skin is too thick.'
這個西瓜皮厚了。

d. Zhe zhong bu bo-le. (kind-denoting)
this kind cloth thin-PRF
'This kind of cloth is too thin.'
這種布薄了。

e. Wu gongjin yan duo-le. Wo zhi yao si gongjin. (quantity-denoting)

five kilo salt much-PRF I only want four kilo
'Five kilos of salt is too much. I want only four kilos.'
五公斤鹽多了，我只要四公斤。

The predicate of an M-LE construction can be a matrix one, as seen the above examples, or be embedded in a modifier, as in (11a), or be the secondary predicate of a complex predicate (for V1V2 constructions where V1 is a verb of creation), as in the rest examples in (11):

(11) a. Ni ba shang-le san cun de shengzi tiao chulai!
you BA long-PRF three inch DE rope pick out
'Pick out the ropes that are three inches longer than the expected length!'
你把長了三吋的繩子挑出來！

b. Yanhui ba mao-yi zhi-chang-le (yi cun).
Yanhui BA wool-sweater knit-long-PRF one inch
'Yanhui has knitted a wool-sweater, which is (one inch) too long.'
燕慧把毛衣織長了（一吋）。

c. Yanhui ba zi xie-da-le (yidian).
Yanhui BA character write-big-PRF a.little
'Yanhui has written the characters (a little bit) too big.'
燕慧把字寫大了（一點）。

d. Yanhui ba keng wa-qian-le (shi gongfen).
Yanhui BA pit dig-shallow-PRF ten centimeter
'Yanhui has dug a pit, which is (ten centimeters) too shallow.'
燕慧把坑挖淺了（10公分）。

In the examples in (11), the scope of the CE is the secondary predicate only, not including the matrix predicate. For instance, for (11b), the speaker has no evaluation of the activity *zhi* 'knit' conducted by Yanhui. What is evaluated is the property of the product of the knitting.

M-LE may occur in either a realis or irrealis context, as shown by (12a) and (12b), respectively.

(12) a. Lulu zuotian ba zi xie-xiao-le.
Lulu yesterday BA character write-small-PRF
'Lulu wrote the characters too small yesterday.'
陸露昨天把字寫小了。

b. Ruguo ni ba zi xie-xiao-le, jiu cai-diao.
if you BA character write-small-PRF then wipe-off
'If you write characters too small, erase them.'
如果你把字寫小了，就擦掉。

2.2. Gradable predicates and the optionality of a measure expression

A measure expression, such as *san gongfen* 'three centimeters' in (1a), may always occur after M-LE, to encode the differential degree from the expectation. However, data like (2a) and (2b) show that such an expression is not obligatory in the construction (contra Zhu 1982: 69, Liu 2007: 779).

Since a measure phrase is always allowed in an M-LE construction, the predicate of the construction must be gradable. Non-gradable predicates may not occur with M-LE, as shown in (13). Example (13d), is acceptable, but there is no CE reading, and thus the *le* is T-LE, rather than M-LE.

(13) a. *Mianbao fang-le.
bread square-PRF
*麵包方了。

b. *Chuanghu tuoyuan-le.
window oval-PRF
*窗戶橢圓了。

c. *Zhe fen dang'an jue-mi-le.
this CL file absolute-secret-PRF
*這份檔案絕密了。

d. Lulu ba shui shao-gan-le. (No CE reading)
Lulu BA water boil-dry-PRF
'Lulu boiled the water so much that the container was dried up.'
陸露把水燒乾了。

2.3. The interactions M-LE and excessive adverbs

If an evaluative excessive degree adverb (e.g., *tai* 'too', *guofen* 'too much', *guoyu* 'too much') occurs, M-LE is optional, as in (14).

(14) a. Zhe gen shengzi tai duan (-le).
this CL rope too short PRF
'This rope is too short.'
這根繩子太短（了）。

b. Lulu dui Duoduo guofen keqi (-le).
Lulu to Duoduo too polite PRF
'Lulu is too polite to Duoduo.'
露露對多多過份客氣（了）。

Sentences with neither *le* nor an excessive adverb do not express CE (also in (1b)):

(15) a. Zhe ge zhuozi da-**le** yi-dianr.
this CL table big-PRF a-little
'This table is a little bit too big.'
這個桌子大了一點。

b. Zhe ge zhuozi da yi-dianr.
this CL table big a-little
'This table is a little bit bigger.'
這個桌子大一點。

Our descriptive generalization is that M-LE and the excessive adverbs have the same semantic function, i.e. they both encode mirativity.[2]

3. Previous analyses

The special role of M-LE in the aspect morphology has been briefly mentioned in the literature (e.g., Chao 1968: 89, 692; Huang 1987: 202; Shi 1988: 102; Lü et al. 1999 [1980]: 355; Ljungqvist 2007: 209; Chang 2009: 2243). However, the most extensive study of M-LE is Liu (2007). He makes the following claims for an M-LE construction.

A. A measure expression is obligatory (Liu 2007: 779).

B. There is a *pro* to encode an individual for an implicit comparison.

C. There is also a covert exceeding verb, called *ex*. This verb is selected from the lexicon in its inflected form, i.e., *ex-le*, before it is adjoined by an adjective (p. 789).

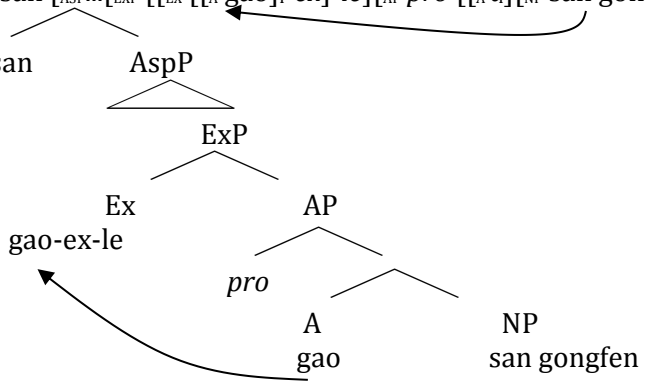
D. The assumed *pro* must be licensed by a head element, and it is *le* that licenses the *pro*.

E. The *pro* is hosted by the AP.

In this approach, (16a) is claimed to have the structure in (16b) (p. 789):

(16) a. Zhangsan gao-le san gongfen.
Zhangsan tall-ASP three centimeter
Intended: 'Zhangsan is three centimeters taller than expected.'
張三高了三公分。

b. [Zhangsan [$_{ASP}$...[$_{EXP}$ [[$_{EX}$ [[$_{A}$ gao]$_i$-*ex*]-*le*][$_{AP}$ *pro* [[$_{A}$ t$_i$][$_{NP}$ san gongfen]]]]]]]



Basically, in Liu's analysis, M-LE is the licensor of a *pro*, which represents an implicit compared individual. Let us call this approach *pro*-licensor (PL) approach. My comments on the claims of this PL approach are the following.

Taking Zhu (1982) for granted, Claim A asserts that a measure expression is obligatory for an M-LE construction. It thus wrongly rules out the acceptability of data like (2a) and (2b). The assumption is thus descriptively inadequate.

Generally, *pro* may be replaced by an overt pronoun, as shown in (17a), however, the assumed *pro* in the M-LE construction may not be replaced by a pronoun, as shown in (17b). Therefore, Claim B is also problematic.

(17) a. Lulu ting-le na ge xiaoxi yihou, {*pro*/ta} tebie gaoxing.

[2] The alternation between M-LE and an evaluative degree adverb also shows that the alleged rule that gradable adjectives must occur with *hen* 'very' in Mandarin Chinese is false. In the presence of M-LE, bare adjective can function as a predicate. This supports Chen (2010) and Grano's (2012) analysis that bare adjectives need to occur with a functional element in order to function as a predicate.

Lulu hear-PRF that CL news after *pro*/3SG especially happy
'After hearing the news, Lulu was especially happy.'
露露聽了那個消息以後，（他）特別高興。

b. Zhangsan gao-le (*ta) san gongfen.
Zhangsan tall-asp 3SG three centimeter
Intended: 'Zhangsan is three centimeters taller than expected.'
張三高了（*它）三公分。 （要表達意：張三比預料高三公分。）

If *pro* must be licensed by a head element, it is not clear why the assumed verb *ex* (Claim C) may not license the *pro*, and thus why *le* is necessary in the analysis. If a *pro* always needs an aspect marker in the containing clause, sentences like (17a) will be unacceptable, contrary to the fact. We conclude that contra Claim D, the function of M-LE is not to license a *pro*.

Finally, the syntactic structure proposed in the PL approach (p. 789) misrepresents the semantics of the construction. In (16b), *san gongfen* is base-generated as the complement of *gao* 'tall' and *pro* is the Spec of the AP (Claim E). The three elements form a complete functional complex (CFC) (Chomsky 1986), which is a predication-denoting configuration. Thus before the AP merges with *ex*, the AP, or CFC, encodes that *pro* is 3 centimeters tall. In other words, the individual that is compared with Zhangsan is 3 centimeters tall. This is not the meaning of (16a).

A similar analysis of the same construction in Cantonese is seen in Mok (1998: 113).

A more reasonable analysis in the PL perspective should be that the measure expression is the complement of ex or another relevant functional head (e.g., μ in Grano & Kennedy 2011: 25), or the complement of the combination of the adjective and μ (Grano & Kennedy 2011: 26, 41), rather than the complement of the adjective directly.

Mainly because of the problem of Claim D, we do not adopt this PL approach.

4. A comparison with other non-temporal uses of *le* in Mandarin Chinese

The goal of this section is to show how M-LE is different from other non-temporal uses of *le*.

4.1 A comparison with the non-temporal S-*le*

We first demonstrate that M-LE is syntactically different from the sentential *le* (S-*le*), which occurs at the end of a sentence. The M-LE in (1a) does not occur sentence-finally. The one in (2a) is also a verbal one, since it can be followed by a measure phrase such as *yi cun* 'one inch'. The one in (2b) can also be followed by *yidianr* 'a little bit'. Obviously, M-LE is not S-*le*.

A further question is whether S-*le* may also express CE reading at all. We find that in certain constructions, S-le does have a non-temporal reading, as seen in (18), but the construction expresses an assertion of a situation, rather than CE.[3]

[3] Soh (2009: 627) claims that temporal CE is expressed if the sentential particle *le* occurs with a stage-level predicate, as in (i).
(i) Wo bu chi mugua le.
I not eat papaya LE
'I don't eat papaya (, which I did before/contrary to what one may expect).'
In order to get the CE reading, one needs to know that the agent of (i) did eat papaya before. Thus, this is a kind of pragmatically-induced CE reading. The default reading of (i) is not CE. Instead, it

(18) Wo zui xihuan Lulu le.
I most like Lulu PRT
'I like Lulu the most.'
我最喜歡露露了。

If *le* occurs after a nominal predicate, which is not the syntactic position of a verbal *le*, no CE reading is attested. Instead, only a temporal reading is possible:

(19) Xiao shu 50 gongfen le.
small tree 50 centimeter PRT
'The small tree has grown up to 50 centimeters.'
小樹50公分了。

The *le* in (18), which is non-temporal, can be replaced with the sentence-final particle *la*, without a change of meaning. M-LE, which is a verbal suffix and thus never follows a nominal, may not be replaced by *la*. This is shown in the acceptability contrast between (20a) and (20b).[4]

(20) a. Wo zui xihuan Lulu la. (the same reading as (18) above)
I most like Lulu PRT
'I like Lulu the most.'
我最喜歡露露啦。

b. *Zhe gen shengzi duan-la 50 gongfen.
this CL rope short-PRT 50 centimeters
*這根繩子短啦50公分。

La can be treated as a complementizer of root declarative clauses, patterning with other sentence-final particles such as *a* or *ya* (exclamatory), and *ne* (interrogative). One shared property of these clause-typing particles is that they may not occur in an embedded sentence, as shown in (21a). However, M-LE may occur in an embedded clause, as in (21b). Therefore, M-LE is not a root-clause marker.

(21) a. Ruguo (ni renwei) [ta zui xihuan Xiaomei (*le)], ….
if you think he most like Xiaomai PRT
'If (you think) he likes Xiaomei most, …'
如果（你認為）[他最喜歡小梅 (*了)]，.....

---

just expresses a new state. For data like (i) to have a CE reading independent of the discourse, an adverb such as *juran* 'unexpectedly' should occur. If the discourse or pragmatic context allows, any sentence can show CE. Since normally people do not eat dirt, (ii) might express a CE.

(ii) Ta chi tu le.
he eat dirt PRT
'He has eaten dirt.'

But the CE reading denoted by M-LE does not need any special pragmatic condition.

[4] I assume that the *le* in the examples in (i) is also a variant of the particle *la*, which is an expressive marker. The expressive marker must occur with a degree adverb such as *ji* 'extremely' and *dai* 'extremely'. This is different from M-LE (contra Shi 1988: 102).

(i) a. Hao ji le! b. Hao-chi ji le! c. Shuai dai le!
goog extremely PRT goog-eat extremely PRT handsome extremely PRT
'Extremely good!' 'Extremely delicious!' 'Extremely handsome!'

b. Ruguo ni renwei [zhe gen shengzi duan-le], qing gaosu wo.
if you think this CL rope short-PRF please tell me
'If you think this rope is too short, please tell me.'
如果你認為這根繩子短了，請告訴我。

Thus syntactically, M-LE behaves like the verbal *le*, rather than the sentential *le*.

4.2. A comparison with another non-temporal verbal *le*

M-LE has not been distinguished from another non-temporal use of verb *le*, which occurs in the so-called transitive comparative constructions. For instance, (22) has three readings, listed in A, B, and C.

(22) Lulu gao-le san gongfen.
Lulu high-PRF three centimeter
露露高了三公分。

A. 'Lulu has become three centimeters taller than before.' (T-LE)
B. 'Lulu is three centimeters taller than another individual identifiable in the discourse context.' (non-temporal & non-M-LE)
C. 'Lulu is three centimeters taller than expected.' (M-LE)

Reading A is a degree achievement reading. The adjective *gao* 'high, tall' denotes a change of state in this context (Grano 2012: Section 3.3). The *le* is thus T-LE, a canonical perfect aspect marker.

(22) also has two non-temporal readings: B and C. Reading B is a transitive comparative reading (the term is from Erlewine 2007; also called Obligatory Measuring Comparatives in Mok 1998: 110; the Bare Comparative in Xiang 2005). The meaning that Lulu is three centimeters taller than another individual or other individuals identifiable in the discourse context is a pure narrative reading, without any evaluation from the speaker. The *le* is thus a non-temporal & non-mirative one.

Reading C is a mirative reading. The meaning that Lulu's height turns out to be three centimeters taller than expected can be found in the context that the speaker is selecting persons of a certain height to complete some special job. In this case, the *le* is M-LE.

We now discuss certain syntactic differences between the two non-temporal readings of the verbal *le* (i.e., reading B and C), leaving a comparison between readings A and C to Section 5.

In Liu's (2007) approach, the same syntax is given to Reading B and Reading C (i.e., M-LE). But the two readings are syntactically different, in five aspects.

<i> The occurrence of a measure expression. A measure expression is obligatory for a transitive comparative construction (Xiang 2005), but not for an M-LE construction (see my comment on Liu's point A in Section 3).

(23) a. Lulu gao-le Duoduo *(san gongfen/henduo/bu-shao). (transitive comparative)
Lulu tall-PRF Duoduo three centimeter/much/not-less
'Lulu is {three centimeters/much} taller than Duoduo.'
露露高了多多*(三公分/很多/不少)。

b. Lulu gao-le (san gongfen). (M-LE construction)
Lulu tall-PRF (three centimeter)
'Lulu is (three centimeters) taller than I expected.'
露露高了(三公分)。

This contrast indicates that an implicit verb (similar to Liu's implicit *exceed*) occurs in the transitive comparative construction, and its argument is the measure phrase.

In Grano & Kennedy (2011, G&K henceforth), in a transitive comparative construction, μ_{COMP} occurs. μ_{COMP} has double functions: it assigns Case to the standard nominal and it introduces a measure expression. Thus, the construction always has a measure expression.

<ii> The syntactic presence of an internal argument. In reading B, the individual identifiable in the discourse context can be encoded by an explicit nominal, as seen in the following (24a). Therefore, an argument may occur between *le* and the measure expression in the transitive comparative construction (see Section 3 for my comment on Liu's point B). But this is impossible for an M-LE construction, as seen in (24b) (We have seen the constraint in (17b)). The acceptability contrast between (24a) and (24b) shows that the M-LE construction is not a transitive comparative construction.

(24) a. Lulu gao-le Duoduo san gongfen. (transitive comparative construction)
Lulu tall-PRF Duoduo three centimeter
'Lulu is three centimeters taller than Duoduo.'
露露高了多多三公分。

b. *Lulu gao-le ta san gongfen. (intended M-LE construction)
Lulu tall-PRF 3SG three centimeter
Intended: 'Lulu is three centimeters taller than the height that I want.'
*露露高了它三公分。

Accordingly, a pronoun, which takes a nominal in the context as its antecedent, may occur between *le* and a measure expression in a transitive comparative construction, as in (25a), but not in an M-LE construction, as in (25b):

(25) a. Lulu bu xiang gen Duoduo$_i$ tiaowu, yinwei Lulu gao-le ta$_i$ san gongfen.
Lulu not want with Duoduo dance because Lulu tall-PRF 3SG three centimeter
'Lulu does not want to dance with Duoduo, because Lulu is three centimeters taller than him.
露露不想跟多多跳舞，因为露露高了他$_{多多}$三公分。

b. *Wo yao [170 gongfen de ren]$_i$. Lulu gao-le ta$_i$ san gongfen.
I want [170 centimeter DE person]$_i$, Lulu tall-PRF 3SG$_i$ three centimeter
intended: 'I want a person who is 170 centimeters tall. Lulu is taller than that.'
*我要[170公分的人]$_i$。露露高了他$_i$三公分。

For a transitive comparative construction, the second argument can be either explicit, as in (24a), or implicit, *pro* (G&K: 42). In the latter case, the surface string is identical to an M-LE construction. Therefore, (22) is ambiguous. For an M-LE construction, however, there is simply no syntactic position for a standard-denoting element (or an element that denotes the desired quality). The alleged *pro* in Liu's analysis does not exist.

Pragmatically, the desired quality is identifiable in the context of an M-LE construction. But the identification of an entity in pragmatics does not ensure a syntactic position for the entity in the syntactic structure. This is similar to middles in English, where no agent position is syntactically available. For the reading of (26a), of course an agent must be available: the book must be read by someone. But there is no agent position in the syntactic structure. No agent-oriented adverb is allowed in middles, as shown in (26b) and (26c). It has been assumed that only an agent can license an agent-oriented adverb, and the unacceptability of (26b) and (26c) indicates that middles have no agent in their syntactic structures (Stroik 1992).

(26) a. This book reads poorly.
b. *The book sold deliberately.
c. *Your books read intentionally.

<iii> The occurrence of *le*. In a transitive comparative construction, *le* does not have to show up. However, in an M-LE construction, by definition, *le* must show up. I have claimed that it is *le* that brings about the mirative reading in an M-LE construction.

(27) a. Lulu gao-(le) Duoduo san gongfen. (transitive comparative construction)
Lulu tall-PRF Duoduo three centimeter
'Lulu is three centimeters taller than Duoduo.
露露高(了)多多三公分。

b. Lulu gao-*(le) san gongfen. (M-LE construction)
Lulu tall-PRF three centimeter
Intended: 'Lulu is three centimeters taller than expected.'
露露高*(了)三公分。 要表達意：露露比預料高三公分。）

<iv> The replacement of *le* by *chu*. The *le* in a transitive comparative construction can always be replaced with *chu* 'exit, go beyond', as seen in (28a), but M-LE may not be replaced by *chu*, as seen in (28b).

(28) a. Lulu gao-{le/chu} Duoduo san gongfen. (transitive comparative construction)
Lulu tall-PRF/beyond Duoduo three centimeter
'Lulu is three centimeters taller than Duoduo.
露露高{了/出}多多三公分。

b. *Lulu gao-chu san gongfen. (M-LE construction)
Lulu tall-beyond three centimeter
Intended: 'Lulu is three centimeters taller than expected.'
*露露高出三公分。 （要表達意：露露比預料高三公分。）

Both transitive comparatives and *bi* constructions can express an unexpected reading if the second argument denotes the expected standard overtly. In this case, *chu* is always possible.

(29) a. Lulu gao-{*le/chu} wo yao de gaodu san gongfen. (transitive comparative)
Lulu tall-PRF 3SG/beyond I want DE height three centimeter
Intended: 'Lulu is three centimeters taller than the height that I want.'
露露高{?了/出}我要的高度三公分。

b. Lulu bi [wo yao de gaodu] gao-{le/chu} san gongfen. (*bi* construction)
Lulu than I want DE height tall-PRF/beyond three centimeter
'Lulu is three centimeters taller than the height that I want.'
露露比我要的高度高{了/出}三公分。

T-LE may not be replaced by *chu*, either.

(30) Lulu mai-{le/*chu} hua.
Lulu buy-{PRF/beyond flower
'Lulu bought flowers.'
露露買{了/*出}花。

This shows that M-LE patterns with T-LE, rather than the one in a transitive comparative construction.

<v> The requirement of a salient numeral scale. Certain predicates of transitive comparative constructions require the measure phrase to denote a salient numeral scale, as seen in the contrast between *gao* 'tall' in (31a) and *piaoliang* 'pretty' in (31b) (see Xiang 2005; also see Grano & Kennedy 2011: 3-4).

(31) a. Lulu gao-(le/chu) Lili {san gongfen/henduo}.
Lulu tall-PRF/beyond Lili three centimeter/much
'Lulu is (three sentimeters/much) taller than Lili.'
露露高（了/出）麗麗 {三公分/很多}。

b. *Lulu piaoliang-(le/chu) Lili {henduo/yidian}.
Lulu pretty-PRF/beyond Lili three centimeter/much
*露露漂亮（了/出）麗麗 {很多/一點}。

The constraint on the transitive comparative construction is not seen in either *bi*-comparative or M-LE constructions, as seen in (32a) and (32b), respectively.

(32) a. Lulu bi Lili piaoliang {henduo/yidian}.
Lulu than Lili pretty much/a.little
'Lulu is {much/a little} more pretty than Lili.'
露露比麗麗漂亮{很多/一點}。

b. Lulu (tai) piaoliang-le yidian. Daoyan yao zhao xiangmao pingping de.
Lulu too pretty-PRF a.little director want seek appearance plain DE
'Lulu is a little too pretty. The director looks for a person with a plain appearance.'
露露(太)漂亮了一點。導演要找相貌平平的。

In G&K (p. 35; 39), the contrast between (31a) and (31b) is accounted for by the selectional restrictions of μ. They claim that "*gao* but not *congming* uses a scale that supports measurement, and so is able to combine with μ" (p. 39).

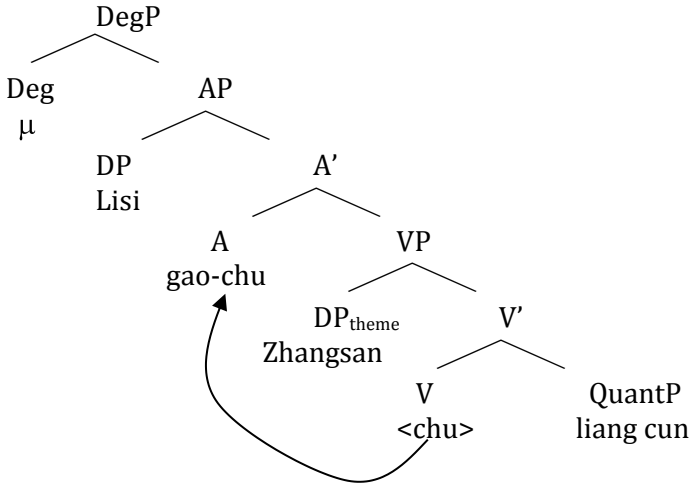
Based on the above five contrasts, we conclude that M-LE constructions are different from transitive comparative constructions. The *le* in the latter construction is in fact a variant of *chu* 'go beyond'.

4.3. The syntactic structure of transitive comparatives

We have seen that the non-temporal *le* in a transitive comparative construction is systematically different from M-LE. In this section, we present our syntactic structure for transitive comparative constructions, leaving the structure of M-LE constructions to 5.3. The syntactic structure of the transitive comparative (33a) is (33b):[5]

(33) a. Lisi gao-(chu/le) Zhangsan liang cun. (transitive comparative)
Lisi tall-exit/PRF Zhangsan two inch
'Lisi is two inches taller than Zhangsan.'

b.



Major characteristics of the structure in (33b) are the following.

Chu assigns Case to *Zhangsan* and takes the measure expression *liang cun* 'two inch' as its complement.

The measure expression *liang cun* is obligatory because it is the internal argument of *chu*. The measure expression is c-commanded by μ, and gets licensed by the latter semantically (cf. Svenonius & Kennedy 2006; G&K: Sec. 2.3). Note that a gradable adjective does not have to occur with a measure expression in other constructions, and thus it is *chu*, rather than μ, that is responsible for the obligatory presence of the measure expression in transitive comparative constructions.[6]

[5] The syntactic integration of a DegP, headed by μ, into a gradable predicate in comparative constructions varies cross-linguistically. In English, a measure phrase c-commands the standard-denoting DP, and therefore, the Deg takes a comparative AP as its complement (G&K p. 17), whereas in Mandarin Chinese, as observed by Xiang (2005), the standard-denoting DP c-commands the measure phrase, and therefore, the adjective may c-command DegP at a certain step of the derivation (cf. G&K p. 22, 24-25).

(i) a. John is three centimeters taller than *Mary*.
b. Duoduo bi Lulu gao san gongfen
Duoduo than Lulu tall three centimeter
'Duoduo is three centimeters taller than Lulu.'
多多比露露高三公分。
c. Duoduo gao Lulu san gongfen.
Duoduo tall Lulu three centimeter
'Duoduo is three centimeters taller than Lulu.'
多多高露露三公分。

[6] The function of *chu* is similar to that of *duo*- 'more' in (i) (see Bhatt 2011). Both *chu* and *duo* require the presence of an expression to denote the excessive degree.

In (33b), *Chu* moves to *gao*. This is similar to a complex predicate construction in Mandarin Chinese, where the head of a lower predicate moves to the right of the head of the higher predicate. For instance, in (34), *feng* 'mad' moves to the right of *qi* 'anger', forming a compound *qi-feng*.

(34) Lisi qi-feng-le Zhangsan
Lisi anger-mad-PRF Zhangsan
'Lisi angered Zhangsan such that the latter became mad.'

In a transitive comparative construction, since *le* may always be replaced by *chu*, they can be morphological variations of the same head element. This is similar to the replacement of S-*le* for *la* discussed above. Moreover, neither *chu* nor *le* has to be overt. Formal elements may occur in various functional head positions, and alternatively, a certain functional head may be realized by various forms. We thus adopt a syntactic position-oriented approach rather than a lexical item-oriented approach (see Zhang 2012).

Although many details of my analysis are different from G&K, the two analysis share the hypothesis that *chu* is responsible for licensing the Case of the second argument and the obligatory occurrence of the measure expression, in a transitive comparative construction. However, G&K does not discuss *le*, and Liu (2007) does not mention *chu* of the construction.

Conclusion: M-LE is different from other non-temporal uses of *le*, either S-*le* or the *le* in a transitive comparative construction.

5. A comparison with T-LE

M-LE constructions and T-LE constructions share formal similarities, although they are obviously different semantically.

5.1. Similarities

First, like T-LE, M-LE may not occur with an element that cannot function as a predicate alone, e.g., a non-predicative adjective. (35a) has neither a temporal nor a mirative reading. Such an element is not compatible with an excessive or mirative adverb, either, as shown in (35b).

(35) a. *{yiqian/benlai/lilai/so-wei}-le
past/original/always/so-called}-PRF
*{以前/本來/歷來/所謂}了

b. *tai {yiqian/benlai/lilai/so-wei}
too past/original/always/so-called}
*太{以前/本來/歷來/所謂}

Second, like T-LE, neither *chu* nor *la* substitution is possible, as shown above (4.1 and 4.2).

Third, if the linguistic context already encodes the relevant meaning, *le* does not have to occur; otherwise, *le* must show up. For T-LE, if the adverb *yijing* 'already' occurs, the aspect marker is optional.

(i) Lulu duo-chi-le *(yi ge pingguo).
Lulu more-eat-PRF one CL apple
'Lulu ate one more apple.'
露露多吃了一個蘋果。

(36) a. Lulu yijing shuijiao, bu yao dasheng shuohua.
Lulu already sleep not should loud speak
'Lulu has already slept. You should not speak loudly.'
露露已經睡覺，不要大聲說話。

b. Lulu shuijiao-le, bu yao dasheng shuohua.
Lulu sleep-PRF not should loud speak
'Lulu has slept. You should not speak loudly.'
露露睡覺了，不要大聲說話。

c. *Lulu shuijiao. (No perfect aspect reading)
Lulu sleep
*露露睡覺。

For M-LE, if the excessive adverb *tai* 'too' occurs, the mirative marker is also optional.

(37) a. Lulu tai gao.
Lulu too tall
'Lulu is too tall.'
露露太高。

b. Lulu gao-le.
Lulu tall-le
'Lulu is too tall.'
露露高了。

c. Lulu gao. (No CE reading)
Lulu tall
'Lulu is taller than others.'
露露高。

Fourth, a measure phrase is optional in both cases.

(38) a. Lulu shui-le (san tian).
Lulu sleep-PRF three day
'Lulu slept for three days.'
露露睡了（三天）。

b. Zhe gen shengzi duan-le (san gongfen).
this CL rope short-PRF three centimeter
'This rope is (three centimeters) shorter than expected.'
這根繩子短了（三公分）。

Fifth, if the semantic context is in conflict with the relevant reading, *le* may not occur. Specifically, if the semantic context is in conflict with the perfectiveness, T-LE may not occur:

(56) a. Lisi xiang pang (*le) san gongjin.
Lisi want fat PRF three kilogram
'Lisi wants to gain three kilograms.' (Lin 2004:87)

b. Lisi mei nian gao (*le) yi gongfen.
Lisi every year tall PRF one centimeter
'Lisi grows a centimeter every year.' (Lin 2004:87)

Similarly, if the semantic context is in conflict with the CE reading, M-LE may not occur:

(56) a. Wo xiwang shengzi duan (*-le) yidian.
I hope rope short-PRF a.little
'I hope that the rope is a little shorter.'
b. Wo xiwang shengzi (*tai) duan.
I hope rope too short
'I hope that the rope is short.'

5.2. Differences

The first difference between M-LE and T-LE is in the semantic types of predicates: T-LE (perfect aspect marker), by definition, is for stage-level predicates, whereas M-LE is for individual-level predicates only (see 2.1).

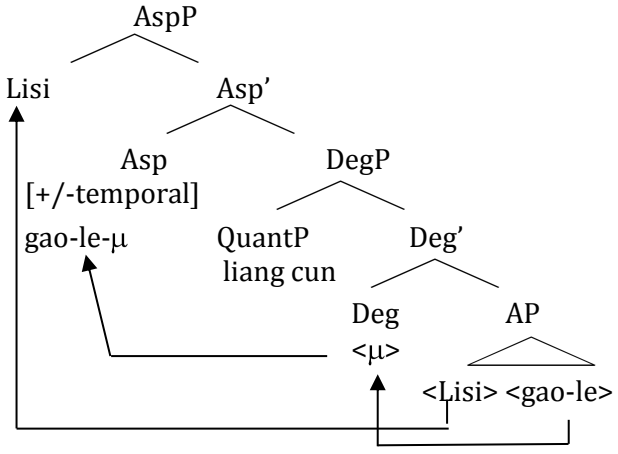
Second, with respect to the semantic types of the predication, there is no constraint on T-LE, however, M-LE occurs in a topic-comment construction (Chang 2009: 2243) which denotes a categorical judgment. In all of our examples, M-LE occurs with an adjective, which has only one argument.

5.3. The syntactic position of the verbal *le*

T-LE and M-LE are both related to the functional head Asp. For both the temporal reading and the CE reading of (39a), the structure is the same, as in (39b).

(39) a. Lisi gao-le liang cun.
Lisi tall-PRF two inch
Degree achievement reading: 'Lisi has grown two more inches.'
[+temporal]
CE reading: 'Lisi is two more inches taller than expected.'
[-temporal]

b.



One major property of the structure in (39b) is that a measure expression is always licensed by the functional element μ, which is base-generated at the head of a DegP (Svenonius & Kennedy 2006; G&K: Sec. 2.3). In (39b), the dependency of *liang cun* 'two inches' on μ is represented as a Spec-Head relation.

Another major property of the structure in (39b) is that the AP selected by μ is gradable, since only gradable adjectives allow degree or measure expressions. From a different perspective, DegP is always integrated with a gradable AP, since the latter provides a dimension for the former.

In summary, certain formal properties of T-LE and M-LE are the same, and they are captured by the same functional head Asp in syntactic structures.

6. Temporal morphology and the feature Exclusion

Why may the perfect aspect marker *le* bring about this special reading in an M-LE construction? How is this reading represented syntactically? These are our empirical issues.

6.1. Perfect aspect morphology and CE

I will show that there is a parallelism between counter-expectation (CE) and counter-factual (CF) constructions: fake perfect aspect in the former and fake past tense in the latter.

Soh (2009) lists a few formal differences between the verbal *le* and the sentential *le*, and shows that only the former must have a perfective or termination reading, in a temporal domain (Soh 2009:627; contra X. Liu 1988). In the last section, we have shown that M-LE patterns with T-LE, i.e., the verbal *le*, rather than the sentential *le*. It is clear that it is perfective morphology that may express the CE reading.

Past-tense may have a counterfactual (CF) reading, encoding a modal of exclusion or remoteness, in English (e.g., Iatridou 2000, Huddleston and Pullum 2002).

(40) a. I wish you <u>were</u> here. (conveys 'You are not here now')
b. I wish the rain <u>would stop</u>. (conveys 'The rain does not stop now.')

Using past-tense to encode CF reading is also seen in many other tense languages (Iatridou 2000, among others), including Japanese (Nishiguchi 2007). The parallel CF reading is even found in the past tense of nominals (Lecarme 2008).

CF reading is also seen in the doubly remote interpretation of past perfect constructions. In this case, the reading is CF to the past (e.g. Iatridou 2000: 232; also see Huddleston and Pullum 2002: 151).

(41) a. If he were smart, he would be rich. (conveys 'he is not smart" and "he is not rich')
b. If he had been smart, he would have been rich. (conveys 'he was not smart' — in general or on one particular occasion — and 'he was not rich')

According to Iatridou (2000: 235), in CF constructions, "we are dealing with past tense morphology that does not receive a past tense interpretation. I will refer to such occurrences of past morphology that do not receive a temporal past interpretation as *fake past* or *fake tense*."

Our data show that the perfect aspect marker *le* may express CE, which is also a modal of exclusion from a certain anchor. Similar to fake past tense, in an M-LE construction, we are dealing with perfect morphology that does not receive a perfect interpretation: fake perfect or fake aspect.

Generally speaking, tense and aspect is a pair of closely related temporal notion. We can see that human language uses the non-temporal readings of tense and aspect morphology to represent the exclusion of a situation from a certain perspective respectively.

(42) a. exclusion from the real world (CF)
b. exclusion from the anchored mental world (CE)

Previous studies of M-LE (e.g., Shi 1988, Ljungqvist 2007, Liu 2007, Chang 2009) did not see this parallelism between tense and aspect morphology. We claim that M-LE reflects the parallelism.

6.2. The grammatical ingredients of CE

6.2.1. CF

According to Iatridou (2000: 246), the representation of fake past in CF is (43) and (44) (C is probably a short-form for Comment, which is in construal with a topic, T).

(43) T(*x*) excludes C(*x*).
(44) a. *T(x)* stands for "Topic(x)" (i.e., "the x that we are talking about").
b. *C(x)* stands for "the x that for all we know is the x of the speaker"

Iatridou uses the feature [exclusion] in her analysis. The variable x can range over times or worlds. When x ranges over times, we get:

(45) a. T(t): the time interval (set of times) that we are talking about
b. C(t): the time interval (set of times) that for all we know is the time of the speaker (i.e., utterance time)

(43) thus derives (46), when its range is the time dimension. (46) means that the topic time is in the past with respect to the utterance time.

(46) The topic time excludes the utterance time.

Note that following Klein (1994), Iatridou (2000: 246) emphasizes that past tense is about the relation between the utterance time and the topic time, rather than that between the utterance time and the situation (or event) time.

Now when x ranges over worlds, we get:

(47) a. T(w): the worlds that we are talking about (topic worlds)
b. C(w): the worlds that for all we know are the worlds of the speaker (actual world)

(43) thus derives (48), when its range is the world dimension. (48) covers a CF reading.

(48) The topic world excludes the actual world.

Similarly, in Arregui's (2008) semantic study of past tense of CF, she also makes an effort of "blurring the boundaries between times and worlds."

6.2.2. CE

Let us now extend this theory of CF to CE. In addition to Iatridou's (43), we propose (49):

(49) T(x) excludes C_{ref}(x).
(50) a. *T(x)* stands for "Topic(x)" (i.e., "the x that we are talking about"). (= (44a))
b. $C_{ref}(x)$ stands for "the x that for all we know is the x of the reference"

When x ranges over times, we get:

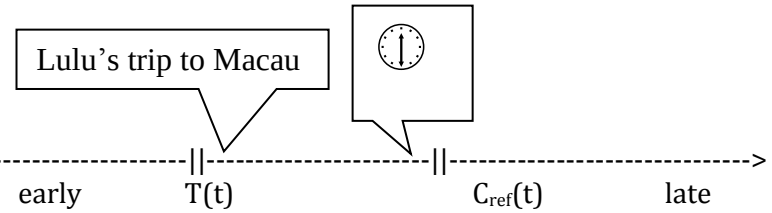
(51) a. T(t): the time interval (set of times) that we are talking about (= (45a))
b. C_{ref}(t): the time interval (set of times) that for all we know is the time of the reference

(49) thus derives (52), when its range is the time dimension:

(52) The topic time excludes the reference time. (cf. (46))

(52) covers perfect aspect. We use (53) to illustrate this. The topic time of this example is the time of Lulu's trip to Macau and the reference time is six o'clock (of any day, either yesterday or tomorrow). The former time is earlier than the latter time, and therefore, the perfect marker *le* (T-LE) is used.

(53) a. {Zuotian/Mingtian} liu dian de shihou, Lulu yijing qu-**le** Aomen.
yesterday/tomorrow six o'clock DE time Lulu already go-PRF Macau
'By six o'clock {yesterday/tomorrow}, Lulu {had/has} made his trip to Macau.'
{昨天/明天}六點的時候，露露已經去了澳門。



When x ranges over worlds, we get:

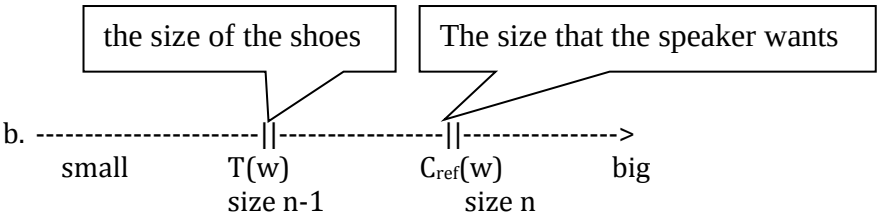
(54) a. T(w): the worlds that we are talking about (topic worlds) (= (47a))
b. C_{ref}(w): the worlds that for all we know are the worlds of the reference

(49) thus derives (55), when its range is the world dimension:

(55) The topic world excludes the reference world. (cf. (48))

(55) covers a CE reading. For instance, in (56), the topic world is the size of the shoes and the reference world is the size that the speaker wants. The former is one number smaller than the latter. The speaker finds the size of the shoes is unexpected, therefore, M-LE is used to encode the mirativity.

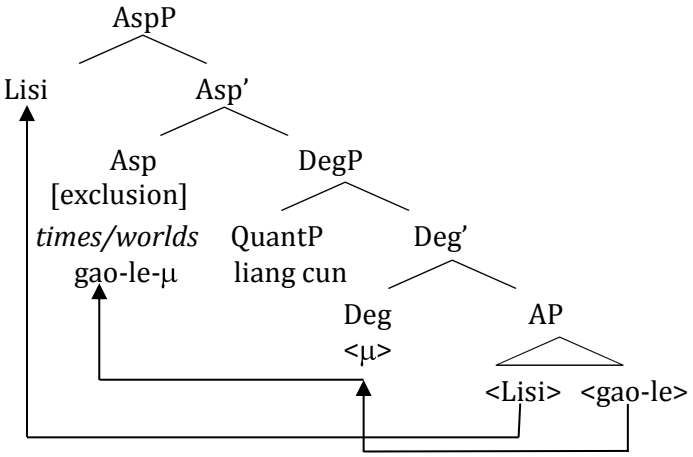
(56) a. Zhe shuang xie xiao-**le** yi hao.
This pair shoe small-PRF one number
'This pair of shoes is one number smaller than expected.'
這雙鞋小了一號。



We now refine our structure in (39) into the following:

(57) a. Lisi gao-le liang cun.
Lisi tall-PRF two inch
Degree achievement reading: 'Lisi has grown two more inches.' *times*
CE reading: 'Lisi is two more inches taller than expected.' *worlds*

b.



6.2.3. A comparison of CE and CF

With respect to the anchor, CF is anchored to the world of the utterance; whereas CE is anchored to the world of a reference. In both cases, it is a kind of displacement feature that encodes a mismatch between the topic world and an anchor.

The past tense morpheme in English has an [exclusion] feature in the CF use. The perfect aspect morpheme *le* in Mandarin Chinese also has a [exclusion] feature in the CE use.

With respect to their functions, neither the fake past tense nor M-LE is the unique way to express counter-X. There are other morphosyntactic ways other than past tense to express CF, cross-linguistically, as mentioned in Iatridou (2000). In Mandarin Chinese, expressions such as *fouzedehua* (否則的話), *burandehua* (不然的話), *yaobushi* (要不是) are used to encode CF, rather than past tense. Likewise, there are other ways to express CE than perfect aspect in Mandarin Chinese, e.g., the use of mirative adverbs.

With respect the syntactic conditions for the occurrence of the CF and CE markers, there are also some language-specific and idiosyncratic constraints. Fake past tense occurs in subordinate clauses only, either in conditional clauses or complement clauses of modal auxiliaries in English (Iatridou 2000; Huddleston and Pullum 2002: 149). The distributions of the fake past are more restricted than the regular temporal past tense morpheme of the language. But fake past tense allows both individual-level and stage-level predicates, and both gradable and ungradable predicates, etc.

Fake perfect aspect occurs with individual-level predicates only. The distributions of M-LE are different from T-LE. But like T-LE, it is found in both matrix and subordinate clauses.

6.2.4. The interactions between adverbs and CF/CE markers

The interactions between adverbs and CF/CE markers follow the language-specific rules (Iatridou 2000: 249). Also, the optionality of the M-LE in the presence of a relevant adverb, as in (14), follows the properties of aspect markers in the language. In English, non-fake past tense is obligatory. Accordingly, the fake past tense of remoteness is also obligatory. In Mandarin Chinese, perfect aspect is not obligatory, if an adverb expresses the relevant meaning. Accordingly, M-LE is not obligatory if an adverb expresses the relevant meaning (Section 2.3). If a mirative adverb occurs without M-LE, the exclusion feature in the null Asp establishes a dependency with the adverb.[7]

6.3. Degree adverbs and measure expressions in M-LE constructions

Each gradable predicate allows only one measure expression such as *san cun* 'three inches'. Each gradable predicate also allows one or two degree adverbs, such as *feichang* 'very', *tai* 'too', *xiangdang* 'quite', *jiqi* 'very', *tebie* 'especially', *gewai* 'extremely', *guofen* 'too'. In (58a) both *tai* and *guofen* occur, and in (58b), both *tebie* and *gewai* occur (note that all excessive adverbs belong to degree adverbs).

(58) a. Ni zhe ge ren tai guofen jiangjiu mianzi.
you this CL person too too care face
'You care about your face too much.'

b. Lulu tebie gewai guanxin qixiang yubao.
Lulu especially especially concern weather forecast
'Lulu especially concerns weather forecast.'

However, between a measure expression and a degree adverb, only one of them may occur, with the same predicate in Mandarin Chinese. In the two examples in (59), either the adverb *feichang* 'very', or the measure expression *190 gongfen* '190 centimeters' occurs, but the two may not occur together. In the two M-LE constructions in (60), either the adverb *tai* 'too', or the measure expression *san cun* 'three inches' occurs, but the two may not occur together.

(59) a. Lulu feichang gao (*190 gongfen).
Lulu very tall 190 centimeter
'Lulu is very (*190 centimeters) tall.'

b. Lulu (*feichang) gao 190 gongfen.
Lulu very tall 190 centimeter
'Lulu is (*very) 190 centimeters tall.'

(60) a. Zhe zhang zhuozi tai kuan-le (*san cun).
this CL table too wide-PRF three inch
'This table is (three inches) too wide.'
這張桌子太寬了(*三吋)。

[7] We do not discuss Iatridou's (2000) analysis of imperfect aspect in CF readings in English. First, we are dealing with perfect, rather than imperfect aspect morphology here. Second, her analysis of the imperfect aspect in CF constructions has been further discussed and challenged in Ferreira (2011). In this paper, I adopt Iadridou's analysis of the past tense morphology of CF only.

b. Zhe zhang zhuozi (*tai) kuan-le san cun.
this CL table too wide-PRF three inch
'This table is three inches too wide.'
這張桌子(*太)寬了三吋。

Cover (2009: 72) claims that a degree adverb and a measure expression do not co-occur (e.g., *three centimeters* (**very*) *high*) because they compete for the same predicate position. But if they are both predicates, it is not clear why they do not form a complex predicate construction. Also, it is not clear why *too* may occur with a measure expression, as seen in the translation of (60b). Assume both measure expressions and degree adverbs are licensed by DegP (e.g., Svenonius & Kennedy 2006; G&K). My hypothesis is that in Mandarin Chinese, one DegP does not license two types of elements at the same time: degree adverb and measure expression.

One apparent exception to the above generalization is that the word *yidian* 'a bit' may occur with a degree adverb. In (61a), the degree adverb *tai* 'too' occurs with *yidian*, and in (61b), the degree adverb *guofen* 'too' also occurs with *yidian*.

(61) a. Zhe zhang zhuozi tai kuan-le yidian.
this CL table too wide-PRF a.bit
'This table is a bit too wide.'
這張桌子太寬了一點。

b. Lulu guofen mingan-le yidian.
Lulu too sensitive-PRF a.bit
'Lulu is a bit too sensitive.'
露露過份敏感了一點。

According to G&K (p. 5), *a bit, a little* and their Chinese counterpart *yidian* is ambiguous between true measure phrases and degree modifiers. Since two degree modifiers may co-occur, as seen in (58), my hypothesis is that in examples like those in (61), *yidian* is a degree modifier rather than a measure expression.

In Mandarin Chinese, although degree adverbs usually precede a predicate, as seen in *tai* and *guofen* in (61), certain degree words may follow a predicative adjective, as seen in (62). Thus our above hypothesis is not ad hoc.

(62) a. Na zhang zhuozi da ji la.
that CL table big extremely PRT
'That table is extremely big.'

b. Ni zou man yidian!
you walk slow a.bit
'Walk a little bit slower!'
你走慢一點！

When *yidian* is used as a degree adverb, it seems to be a post-predicate counterpart of the pre-predicate degree adverb *youdian* 'a bit'. The two examples in (63) mean the same, and the two examples in (64) also mean the same.

(63) a. Zhe ge shouji zhong-le yidian.
this CL cell-phone heavy-PRF a.bit
'This cell-phone is a bit too heavy.'
這個手機重了一點。

b. Zhe ge shouji youdian zhong-le.
this CL cell-phone a.bit heavy-PRF
'This cell-phone is a bit too heavy.'
這個手機有點重了。

(64) a. Tang zhu-xian-le yidian.
soup cook-salty-PRF a.bit
'The soup has been cooked a bit too salty.'
湯煮鹹了一點。

b. Tang youdian zhu-xian-le.
soup a.bit cook-salty-PRF
'The soup has been cooked a bit too salty.'
湯有點煮鹹了。

We have two remaining puzzles. First, if the pre-adjective degree adverb is *shaowei* 'a bit' or *duoshao* 'somehow', the occurrence of *yidian* to the right the adjective is obligatory, as seen in (65).

(65) a. Zhe tiao qunzi shaowei duan-le *(yidian).
this CL skirt a.bit short-PRF a.bit
'This skirt is a bit too short.'
這條裙子**稍微**短了*(一點)。

b. Zhe tiao qunzi duoshao duan-le *(yidian).
this CL skirt somehow short-PRF a.bit
'This skirt is somehow too short.'
這條裙子**多少**短了*(一點)。

Second, unlike degree adverbs such as *tai* 'too', *guofen* 'too', *youdian* 'somehow' (see (61a), (61b), and (63b), respectively), adverbs such as *hen* 'very', *tebie* 'especially', and *feichang* 'very' may not occur in an M-LE construction.

(66) *Zhe gen shengzi {hen/tebie/feichang} duan-le.
this CL rope very/especially/very short-PRF

At this stage of the research, I do not have an account for these puzzles.

6.4. The C-domain of M-LE constructions

All evaluatives must be anchored to the reference world of a person. Thus, CE must be anchored to a reference world of a person. In the absence of an explicit anchor, the speaker is the anchor. (67a) (= (3a)) may, but (67b) (= (2a)) may not, be followed by a sentence meaning 'but I don't think so'.

(67) a. Lulu renwei zhe gen shengzi duan-le. (Dan wo bu zheme renwei)
Lulu think this CL rope short-PRF but I not so think
'Lulu thinks this rope is shorter than expected.' ('but I don't think so.')
露露認為這根繩子短了。但我不这么认为。

b. Zhe gen shengzi duan-le. (*dan wo bu zheme renwei)
this CL rope short-PRFbut I not so think
'This rope is too short.' (*'but I don't think so.')

這根繩子短了。(*但我不这么认为。)

If (67b) above is paraphrased as (68) below, the semantic contradiction is explicit.

(68) *I claim that this rope is too short, but I do not accept this claim.

The anchoring of an evaluation is also represented in syntactic structures. All evaluative sentences have Evaluative Phrase, subcategorized by the head of Speech Act Phrase in the Complementizer Domain (C-Domain) (Cinque 1999; see Tenny & Speas 2003 for the difference of this theory from Ross's 1970 Performative Hypothesis).

Moreover, the reference world is either hosted by the Spec of Speech Act Phrase, i.e. Speaker (Speas 2004: 265), or the subject of a verb of thought (Speas 2004: 267, Tenny & Speas 2003: 335) (e.g. *renwei* 'think'). The former situation is seen in (67b), and the latter is seen in the first conjunct of (67a). In (67a), the subject of the verb *renwei* 'think' is *Lulu*, and thus *Lulu* represents the reference world.

Finally, the reason for Speaker as the default reference world is that "a feature for matrix Speaker is globally available" (Tenny & Speas 2004: 8), and "if we assume that consistent indexing is the unmarked case, each indexing disjunct from speaker has an additional cost" (Speas 2004: 266).

7. Summary

The two research questions raised in the introduction section of this paper are:

A. Why may the perfect aspect marker *le* bring about the mirative reading?

B. How is this reading represented syntactically?

I have shown that the mirative reading of M-LE comes from the feature [exclusion], and M-LE has the same syntactic position as the canonical T-LE. Specifically, I have shown that M-LE occurs with individual level gradable predicates only, and that M-LE constructions are syntactically different from transitive comparative constructions. I have also argued for a parallelism between counterfactual (CF) morphology and counter-expectedness (CE) morphology. I have shown that the feature [exclusion] covers both the fake past tense in CF and the fake perfect aspect in CE.

From the perspective of the mirativity studies, one can see that similar to CF, the ways of encoding mirativity vary in different languages and within the same languages (e.g., via functional words or adverbs). From the perspective of the aspect marker studies, one can see that Chinese *le* brings us to more facts of temporal and modal morphology and clause structures.

References

Arregui, Ana. 2008. On the Role of Past Tense in Resolving Similarity in Counterfactuals. In Atle Grønn (ed.) *Proceedings of SuB12*, Oslo: ILOS, 17–31.

Bhatt, Rajesh. 2011. Ordinary and differential comparatives. Paper presented at the International Workshop on the Syntax-Semantics Interface, Institute of Linguistics Academia Sinica, Taipei, June 18, 2011.

Chang, Li-Hsiang. 2009. Stance uses of the Mandarin LE constructions in conversational discourse. *Journal of Pragmatics* 41: 2240-2256.

Chao, Yuen-Ren. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chen, Jidong & Yasuhiro Shirai. 2010. The development of aspectual marking in child Mandarin Chinese. *Applied Psycholinguistics* 31: 1-28.

Chen, Yuanlu. 2010. Degree Modifications and Boundedness of Adjectives in Mandarin. MA thesis, National Chung Cheng University.

Chomsky, Noam. 1986. *Knowledge of language: its nature, origin and use.* New York: Praeger.

Cinque, Guglielmo. 1999. *Adverbs and Functional Heads. A Cross-linguistic Perspective*. New York: Oxford University Press.

Corver, Norbert. 2009. Getting the (syntactic) measure of Measure Phrases. *The Linguistic Review* 26: 67-134.

DeLancey, Scott. 2001. The mirative and evidentiality. *Journal of Pragmatics* 33: 369-382.

Erlewine, Michael. 2007. A new syntax-semantics for the Mandarin *bi* comparative. MA Thesis, University of Chicago.

Ferreira, Marcelo. 2011. Displaced aspect in counterfactuals: towards a more unified theory of imperfectivity. Paper presented at I Encontro Internacional de Sintaxe e Semantica & suas Interfaces, Pucrs, Aug. 15-26, 2011.

Grano, Thomas. 2012. Mandarin *hen*, universal markedness, and gradable adjectives. To appear *Natural Language and Linguistic Theory*.

Grano, Thomas and Chris Kennedy. 2010. Mandarin Transitive Comparatives and the Grammar of Measurement. Ms. University of Chicago.

Huang, Lillian Meei-jin. 1987. Aspect: A general system and its manifestation in Mandarin Chinese. PhD dissertation, Rice University.

Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Iatridou, Sabine. 2000. The grammatical ingredients of counterfactuality. *Linguistics Inquiry* 31: 231-270.

Klein, Wolfgang. 1994. *Time in language*. London: Routledge.

Lecarme, Jacarme. 2008. Tense and modality in Nominals. In J. Guéron & J. Lecarme (eds.) *Time and Modality*, Heidelberg: Springer, 195-245.

Li, Yen-hui Audrey. 1998. Argument Determiner Phrases and Number Phrases, *Linguistic Inquiry* 29**:** 693-702.

Liu, Chen-Sheng. 2007. The chameleon in the Chinese individual exceed comparative. *Language and Linguistics* 8.3: 767-797.

Lin, Jimmy. 2004. Even structure and the encoding of arguments: The syntax of the Mandarin and English verb phrase. Ph.D. Dissertation, MIT, Cambridge.

Liu, Xunning. 1988. Xiandai hanyu ciwei 'le' de yufa yiyi. *Zongguo Yuwen* 5: 321–330.

Ljungqvist, Arin Marita. 2003. Aspect, tense and mood: context dependency and the marker le in Mandarin Chinese. PhD thesis, Lund University.

Ljungqvist, Arin Marita. 2007. *Le*, *guo* and *zhe* in Mandarin Chinese: a relevance-theoretic account. *Journal of East Asian Linguistics* 16:193-235.

Lü, Shuxiang et al. 1999. *Xiandai Hanyu Babai Ci* [800 Words in Chinese]. Beijing: Shangwu Press (1st edition, 1980).

Mok, Sui-Sang. 1998. Cantonese exceed comparatives. Ph.D. Dissertation, University of California, San Diego.

Nishiguchi, Sumiyo. 2007. Temporal Dynamic Semantics of Factual Counterfactuals. In A. Sakurai et al. (eds.), *New Frontiers in Artificial Intelligence: JSAI 2003 and JSAI 2004,* 438-448, Berlin, Heidelberg: Springer.

Ross, John R. 1970. On declarative sentences. In R. A. Jacobs & Peter S. Rosenbaum (eds.) *Readings in English Transformational Grammar*, Waltham, Mass.: Ginn.

Shen, Li. 2004. Aspect Agreement and Light Verbs in Chinese: A Comparison with Japanese, *Journal of East Asian Linguistics* 13: 141-179.
Shi, Ziqiang. 1988. The present and past of the particle "le" in Mandarin Chinese. PhD. dissertation, University of Pennsylvania.
Smith, Carlota. 2008. Time with and without tense. In Guéron & Lecarme (eds.) *Time and Modality*, Heidelberg: Springer, 227-249.
Soh, Hooi Ling. 2009. Speaker presupposition and Mandarin Chinese sentence-final *-le*: a unified analysis of the "change of state" and the "contrary to expectation" reading. *Natural Language and Linguistic Theory* 27: 623-657.
Speas, Margaret. 2004. Evidentiality, logophoricity, and the syntactic representation of pragmatic features. *Lingua* 114: 255-276.
Stroik, Thomas. 1992. Middles and movement. *Linguistic Inquiry* 23: 127-137.
Svenonius, Peter and Christopher Kennedy. 2006. Northern Norwegian degree questions and the syntax of measurement. In Mara Frascarelli (ed.) *Phases of interpretation*, Volume 91 of *Studies in Generative Grammar*, 133–161. Berlin: Mouton de Gruyter.
Tenny, Carol & Margaret Speas. 2003. Configurational properties of Point of View roles. In A. DiSciullo (ed.) *Asymmetry in Grammar,* John Benjamins, 315-344.
Tenny, Carol & Margaret Speas. 2004. The interaction of clausal syntax, discourse roles and information structure in questions. Handout for the Workshop of Syntax, Semantics and Pragmatics of Questions, Nancy, Aug. 13 2004.
Xiang, Ming. 2005. Some topics in comparative constructions. PhD thesis, Michigan State University, East Lansing, MI.
Yang, Suying. 2011. The parameter of temporal endpoint and the basic function of –le. *Journal of East Asian Linguistics*. To appear.
Zhang, Niina Ning. 2009. The Syntax of Relational-Nominal Second Constructions in Chinese. *Yuyanxue Luncong* 39. Beijing: Peking University Press, 257-301.
Zhang, Niina Ning. 2012. *De* and the Functional Expansion of Classifiers. *Language and Linguistics* 13 (3): 569-582.
Zhu, Dexi. 1982. *Yufa Jiangyi* [Lectures on grammar], Beijing, Shangwu Press.

Fitting into morphological structure: accounting for Sorani Kurdish endoclitics

Géraldine Walther
Sorbonne Paris Cité, Université Paris Diderot & CNRS
Laboratoire de Linguistique Formelle
geraldine.walther@linguist.jussieu.fr

0. Abstract: In this paper, we investigate the notion of morphological structure in the light of the behaviour of Sorani Kurdish mobile person markers (MPM) (Samvelian, 2007). We propose a formal analysis of Sorani Kurdish transitive verb inflection taking into account the complete Sorani verb paradigm, including the past tense forms that had been neglected so far. We argue that some structure is needed, even within a realisational approach, to account for the otherwise erratic behaviour of Sorani Kurdish MPMs. Our analysis is couched within a variant of *PFM* (Stump, 2001).

Keywords Sorani Kurdish, mesoclisis, endoclisis, morphology-syntax interface, morphological structure, lexeme-based morphology, non-canonical person marking, PFM.

1. Introduction

The question whether or not there should be structure specified within morphology has provoked numerous debates among contemporary morphologists. Lieber (2006) even goes so far as to call it "a veritable war concerning the relationship between syntax and morphology, with positions ranging from the militantly lexicalist to the staunchly syntacticalist." While the lexicalist, non-morphemic position (Anderson, 1992; Beard, 1995; Stump, 2001) argues that there can be no structure in morphology, the opposed syntactic position, building on the notion of morpheme (Lieber, 1992; Baker, 1988), argues for the exact opposite, namely that morphological structure is of the same kind as syntactic structure.

The existence of structure in morphology also questions the *Lexical Integrity Hypothesis* (Bresnan, 1995), usually broadly admitted among lexicalists. In the introduction to her work on Udi endoclitics, Harris (2002) points out the following:

> "Many linguists have assumed the correctness of the Lexical Integrity Principle, the hypothesis that words are composed according to morphological principles that differ in kind from the syntactic principles responsible for the composition of sentences. According to this hypothesis, the morphological composition of a word is not accessible to the rules of syntax, and items manipulated by syntactic rules (i.e., words) cannot be manipulated by the morphology (see, for example (Di Sciullo and Williams, 1987; Bresnan, 1995))."

However, data of the kind described, for example, by Harris (2002) for Udi show that the Lexical Integrity Hypothesis is seriously challenged by some clitic placement phenomena. This is also illustrated by Crysmann (2002) on European Portuguese, where clitics occur inside words, and by Crysmann (2006) who reanalyses the data presented by Kupść and Tseng (2005) for Polish, where elements described as having affixal status "float of" leaving their word domain to attach with other syntactic units.

The clitic placement data in these papers constitutes a particular challenge for a-morphous morphology. In Crysmann's (2002) analysis for example, word-internal,

morph-based structure is needed to account for the *endoclitic*[1] behaviour of European Portuguese person markers.

In this paper, we adopt a lexicalist point of view. In particular, we follow an inferential-realisational approach to word-form realisation (Zwicky, 1985; Anderson, 1992; Corbett and Fraser, 1993; Stump, 2001; Brown and Hippisley, 2012). However, drawing on evidence from known (Samvelian, 2007) and original data from Sorani Kurdish endoclitics, we claim that the solution to the question on morphological structure lies on some intermediate golden path between morphemic and a-morphous views of morphology. We show that some morphological structure needs to be at least partly accessible in order to account for Sorani clitics' particular placement properties. In particular, we argue that their seemingly erratic interaction with Sorani perfect tenses, which only superficially resemble periphrasis, can only be explained through internal morphological structure.

2. Sorani Kurdish data

The data used in this paper has been mainly extracted from reference grammars (McCarus, 1958; Thackston, 2006; MacKenzie, 1961). To a lesser extend we have also used information contained in (Blau, 1980, reprinted 2000), and various other descriptions (Garzoni, 1787; Justi, 1880; Soane, 1913; Beidar, 1926; Wahby, 1956; Bedir Khan and Lescot, 1970). Some original data has also been directly obtained from native speakers of the language.

2.1. General presentation

Sorani Kurdish is a Western Iranian language mostly spoken in the Iraq-Iran border regions. It possesses a written standard based on the dialects of the cities of Sulaymaniyah and Arbil (Iraq). We use the more common name *Sorani* to refer to its standardised dialect, corresponding to what Haig (2010) refers to as *Suleimani,* a dialect of the Central Group of the Kurdish languages. Sorani mainly distinguishes itself through its complex verbal morphology and in particular its intricate system of *"endoclitic" or "mobile" person markers,* henceforth MPM (Samvelian, 2007).

Sorani Kurdish resembles most Iranian languages in the fact that it possesses only a very limited amount of synthetic verbal lexemes (around 300). Most traditional verbal meanings are expressed through complex verbal predicates built from a light verbal head and a predicative element which can be either a noun or an adjective, or even an adposition or a preverb or postverb (MacKenzie, 1961; Blau, 2000).

2.2 Sorani Kurdish verbal inflection and person marking

General verbal inflection pattern Sorani verb forms roughly consist of a set of prefixes and suffixes clustered around a given stem. The relative order of the affixes is illustrated by the position classes shown in Table 3.

[1] Crysmann (2006) uses the term *mesoclisis* for insertion of clitics between identifiable morphs, which allows for reserving the term *endoclisis* for insertion inside the stem. Stem-internal cliticisation is described for Udi verbs in (Harris, 2002), alongside inter-morph insertion. However, (Harris, 2002) uses the term endoclisis as a denomination for both types of word-internal placement, as defined by the original coining of the term by Zwicky (1977). In this paper, we thus follow Harris (2002) and Zwicky (1977) in using the term *endoclisis* as a general term for cases where a clitic appears inside the word thereby breaking its phonological integrity. Note that Anderson (2005) mentions both terms as synonyms and then resorts to using *endoclisis* as a general term as well.

Prefix(es) - Stem - Suffix(es)

Stems Most traditional descriptions of Sorani morphology concur in stating the existence of two distinct verbal stems: one for the present tense forms and one for the past and non-finite forms.

Example: KIRDIN "to do" (STEM1 *ke,* STEM2 *kird)*

(1) *da-**kird** =ê*
IND-***do***-P3.PRES
"He does/makes" (PRESENT INDICATIVE)

(2) *da-**ke**-Ø*
IND- ***do*** -P3SG.PAST
"He did/made" (IMPERFECT INDICATIVE)

(3) *bi-**ke**-m-aja*
SUBJ-***do***-P1SG.PAST-SUBJ.PAST
"I would have done/made" (PRETERITE SUBJUNCTIVE)

In addition to these two traditional stems, Bonami and Samvelian (2008) suggest the existence of a third stem for the passive forms due to observable irregularities occurring during stem selection for these forms. The relevant data can be found in (McCarus, 1958) of which some examples are given in Table 1.

	INFINITIVE	STEM1	PASSIVE STEM	TRANSLATION
STEM1	*kuštin*	*kuš*	*kuš–râ/rê*	to kill
STEM2	*ûtn*	*l'ě*	*ût–râ/rê*	to say
STEM2	*bistin*	*bye*	*bst–râ/rê*	to hear
STEM1 MINU END-VOWEL	*birdin*	*be*	*b–râ/rê*	to carry
STEM1 MINUS END-VOWEL	*kirdin*	*ke*	*k–râ/rê*	to do, to make
STEM1 MINUS END-VOWEL	*dân*	*de*	*d–râ/rê*	to give
MODIFIED STEM1	*xwardn*	*xo*	*xû–râ/rê*	to eat
MODIFIED STEM2	*girtin*	*gir*	*gîr–râ/rê*	to take

Table 1: Irregular stem formation for Sorani passive forms

Moreover, the data in reference grammars (McCarus, 1958; MacKenzie, 1961; Blau, 2000; Thackston, 2006) shows a fourth possible stem for imperative forms. Hence up to two additional stems may be stipulated for Sorani verbs (Walther, 2011). Examples are given in Table 2.

INFINITIVE	STEM1	IMP. SING.	IMP. PL.	TRANSLATION
kirdin	*ke*	*bi–ke*	*bi–ke–n*	to do, to make
roîštn	*ro*	*bi–ro*	*bi–ro–n*	to go
bûn	*bi*	*bi–bi–e*	*bi–bi–(i)n*	to be, to become
čûn	*č*	*bi–č–e*	*bi–č–(i)n*	to go
girtin	*gir*	*bi–gir–e*	*bi–gir–(i)n*	to take, to grasp
nûsîn	*nûs*	*bi–nûs–e*	*bi–nûs–(i)n*	to write
da nîštn	*da nîš*	*da bi–nîš–e*	*da (bi–)nîš–(i)n*	to sit down
birdin	*be*	*bi–be–re*	*bi–be-n*	to carry
dan	*de*	*bi–de-re*	*bi–de–n*	to give
xistin	*xe*	*bi–xe–re*	*bi–xe–n*	to pull
hatin	*yě*	*wer–e*	*wer-in*	to come

Table 2: Irregular stem formation for Sorani Kurdish imperative forms

Inflectional affixes Inflectional affixes cluster around the verbal stems: prefixes mostly convey tense, aspect and mood *(*TAM*)* and polarity features, whereas suffixes may encode TAM and person information. Thus the affix distribution can be represented in terms of position classes (as used in PFM (Stump, 2001)). The distribution is illustrated in Table 3.

-2	-1	0	(P)	1	2	3
ma	da	STEM	râ	+û	im	aja
na	bi	(1, 2, 3, 4)	rê	+bû	î(t)	a
nâ				b	ê(t)	
					a	
					în	
					in	
					Ø	
				n		

Table 3. Sorani Kurdish verbal position classes within a PFM analysis

Sorani passives Passives are formed by inserting the sequence *–rê/–râ* between the stem and the other suffixes (shaded column in Table 3). Apart from these markers, the passive inflection exactly matches the paradigms of active intransitives.[2] Active transitive verbs, on the other hand, follow other, specific patterns.

Person marking Sorani is traditionally considered to display three sets of personal endings: PE1 for the present verb forms and the perfect subjunctive derived from STEM1, PE2 for imperfective, preterite and past perfect forms derived from STEM2, and PE3, being identical to the enclitic present forms of the verb *bûn* 'to be', for the remaining perfect forms. In fact those three paradigms only differ in the third person singular, as shown in Table 4.

[2] Descriptive grammars implicitly present *–rê/–râ* as just another set of inflectional suffixes for transitive verbs (McCarus, 1958; MacKenzie, 1961; Blau, 2000; Thackston, 2006). Walther (2011) gives however arguments for a derivational analysis of Sorani passives.

Personal Endings		**PE3** Enclitic personal endings	
PE1 Present personal endings		*–m*	*–în*
–m	*–în*	*–î(t)*	*–n*
–î(t)	*–n*	*–a*	*–n*
–ê(t)	*–n*	**Mobile person markers MPM**	
PE2 Past personal endings			
–m	*–în*	*–m*	*–mân*
–î(t)	*–n*	*–t*	*–tân*
–ø	*–n*	*–î*	*–yân*

Table 4. Person marking

The MPM constitute a fourth set of person markers. They mark the object in the present tense, while the standard person markers are used for subject marking. In the past tense a morphological reversal in the sense of Baerman (2007) occurs: MPMs become subject markers whereas standard person markers are used for object marking.

Moreover, clitic MPMs (Samvelian, 2007) do not appear in a fixed slot among the positions represented in Table 3 when attaching to a verb form; they may insert in various positions following a seemingly erratic pattern.

1. When it is an endoclitic, the MPM inserts after the first prefix, in verb-internal second position.
    a. The first prefix may be the indicative present and imperfect prefix *da–*,
    b. the subjunctive and imperative prefix *bi–*,
    c. or the negative prefix (be it *na–, nâ–* or *ma–*).
2. The third singular person marker (be it standard or MPM) always follows the object marker.
3. Be it a subject or object marker, the first singular person marker always precedes any plural marker.

Tables 5 to 9 show the placement of the endoclitic (marked through the use of angle-brackets <>) for the indicative forms with positive polarity. Other present or past forms containing prefixes *bi–*, *na–*, *nâ–*, *ma–* behave like the present or imperfect respectively.

SUBJECT=PE1 / OBJET=MPM	P1SG	P2SG	P3SG	P1PL	P2PL	P3PL
P1SG		da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1
P2SG	da=◇=STEM1–PE1		da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1
P3SG	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1
P1PL	da–STEM1–PE1 =◇	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1
P2PL	da–STEM1–PE1 =◇	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1
P3PL	da–STEM1–PE1 =◇	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1	da=◇=STEM1–PE1

Table 5. MPM placement in the present paradigm

SUBJECT=PE1 / OBJET=MPM	P1SG	P2SG	P3SG	P1PL	P2PL	P3PL
P1SG		da=◇=STEM2–PE2	da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da=◇=STEM2–PE2	da=◇=STEM2–PE2
P2SG	da=◇=STEM2–PE2		da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da=◇=STEM2–PE2	da=◇=STEM2–PE2
P3SG	da=◇=STEM2–PE2	da–◇–STEM2–PE2	da–STEM2–PE2 =◇	da–◇–STEM2–PE2	da–◇–STEM2–PE2	da=◇=STEM2–PE2
P1PL	da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da=◇=STEM2–PE2	da=◇=STEM2–PE2
P2PL	da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da=◇=STEM2–PE2	da=◇=STEM2–PE2
P3PL	da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da–STEM2–PE2 =◇	da=◇=STEM2–PE2	da=◇=STEM2–PE2	da=◇=STEM2–PE2

Table 6. MPM placement in the imperfect paradigm

SUBJECT=PE1 / OBJET=MPM	P1SG	P2SG	P3SG	P1PL	P2PL	P3PL
P1SG		STEM2 =◇=PE2	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2 =◇=PE2	STEM2 =◇=PE2
P2SG	STEM2 =◇=PE2		STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2 =◇=PE2	STEM2 =◇=PE2
P3SG	STEM2 =◇=PE2	STEM2 =◇=PE2	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2 =◇=PE2	STEM2 =◇=PE2
P1PL	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2 =◇=PE2	STEM2 =◇=PE2
P2PL	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2 =◇=PE2	STEM2 =◇=PE2
P3PL	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2–PE2 =◇	STEM2 =◇=PE2	STEM2 =◇=PE2	STEM2 =◇=PE2

Table 7. MPM placement in the preterite paradigm

SUBJECT=PE1 / OBJET=MPM	P1SG	P2SG	P3SG	P1PL	P2PL	P3PL
P1SG		STEM2–û=◇=PE2	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û=◇=PE2	STEM2–û=◇=PE2
P2SG	STEM2–û=◇=PE2		STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û=◇=PE2	STEM2–û=◇=PE2
P3SG	STEM2–û=◇=PE2	STEM2–û=◇=PE2	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û=◇=PE2	STEM2–û=◇=PE2
P1PL	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û=◇=PE2	STEM2–û=◇=PE2
P2PL	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û=◇=PE2	STEM2–û=◇=PE2
P3PL	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û–PE2 =◇	STEM2–û=◇=PE2	STEM2–û=◇=PE2	STEM2–û= ◇=PE2

Table 8. MPM placement in the present perfect paradigm

SUBJECT=PE1 / OBJET=MPM	P1SG	P2SG	P3SG	P1PL	P2PL	P3PL
P1SG		STEM2–bû=◇=PE2	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2
P2SG	STEM2–bû=◇=PE2		STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2
P3SG	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2
P1PL	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2
P2PL	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2
P3PL	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû–PE2 =◇	STEM2–bû=◇=PE2	STEM2–bû=◇=PE2	STEM2–bû= ◇=PE2

Table 9. MPM placement in the past perfect paradigm

Remnant split ergativity Finally, transitive Sorani verbs display remnant features of split ergativity (Haig, 2010): for transitive verbs in the present tenses, as well as for intransitive verbs, subject marking is handled as described above by using one of the three sets of standard person markers. In present forms of transitive verbs, MPMs only function as optional object marking. However, in the past tenses of transitive verbs, a morphological reversal (Baerman, 2007) occurs: the standard person markers are replaced with the MPMs as subject-verb agreement markers. The standard person markers are then used for object marking. Their presence is optional when another NP already functions as direct object in the sentence.

3. A previous formal analysis of Sorani Kurdish clitic placement

3.1. Placement properties of the MPM (Samvelian, 2007)

A first account of the clitic placement in Sorani Kurdish is given by Samvelian (2007). The author states that MPMs usually adopt (internal or external) second position.

Thus she specifies the following MPM placement rules:

1. If the VP is not verb-initial, the MPM is in enclitic second position in the VP.
2. If the VP is verb-initial, the MPM is an endoclitic within the verb.

These general rules show that Sorani MPMs can be either enclitics or endoclitics.

Concerning verb-internal placement, Samvelian (2007) also notes the following more specific constraints — the first one being the general rule, as illustrated in as presented in Examples (4) and (5), and the second one a more specific override thereof, cf. Examples (6) and (7).

a) If the VP is verb-initial, the MPM is an endoclitic within the V and occurs after its first morph.

(4) xward=**man**=in
eat.STEM=**P1PL**=P3PL
"We ate them."

(5) na=**man**=xward-in
NEG=**P1PL**=eat.STEM-P3PL
"We did not eat them."

b) The third singular person subject marker (be it standard or MPM) always follows the object marker.

(6) xward-in=**î**
eat.STEM-3PL=**3SG**
"He ate them."

(7) na-xward-in=**î**
NEG-eat.STEM-3PL=**3SG**
"He did not them."

The constraints presented by Samvelian (2007) have been formalised by Bonami and Samvelian (2008). The verb-external placement appears to be straightforward, once the VP internal second-position placement has been stipulated. However, the verb-internal placement still needs to be further investigated — especially since the authors do not study the complete paradigm. In this paper, we therefore focus on the sole verb-internal placement.

3.2 Formalisation of the MPM placement (Bonami and Samvelian, 2008)

Bonami and Samvelian (2008) formalise the clitic placement described by Samvelian (2007) for the present and the past non-perfect tenses.[3] For this purpose, they use concepts introduced by Crysmann (2002) for treating the verb internal insertion of European Portuguese clitic person markers. The idea underlying the analysis by Crysmann (2002) is that morphology operates on *morph lists* rather than unstructured phonology. Morph lists are an intermediate structured level between individual morphs and words. They contain elements that are either independent morphs or opaque morph clusters. The assumption underlying the idea of morph-lists is that elements of heterogeneous nature such as clitics do have access to elements of morph-lists (individual morphs or morph-clusters) but not to the structure within morph-list elements, and in particular do not access the inner structure of the clusters within the morph-list.

[3] Bonami and Samvelian (2008) also formalise the imperative forms. We do not give an explicit formalisation of the imperative forms here since, for those forms, the MPM placement is fully covered by the rules for the general case placement presented by Bonami and Samvelian (2008).

Bonami and Samvelian (2008) use the notion of morph list to account for the placement constraints on Sorani clitics in the following way.

As stated by Samvelian (2007) and illustrated in the data contained in McCarus (1958), MPM placement is not phonologically determined. One cannot, for example, deduce the placement of the MPM from the location of word-stress. Thus, contrarily to the behaviour of Pashto endoclitics, for example, as described by Anderson (2005), Sorani MPM display placement properties that appear to be strictly morphologically determined.

Bonami and Samvelian (2008) state that whenever the MPM is verb internal (i.e., the verb is VP initial), the default placement for the MPM is to occur in verb internal second position within the morph list. A simple example thereof is given in (4) and (5).

(4) xward=**man**=in
eat.STEM=**P1PL**=P3PL
"We ate them."

(5) na=**man**=xward-in
NEG=**P1PL**= eat.STEM-P3PL
"We did not eat them."

In this case the morph list can be formalised as simply being a list of monomorphic elements. The endoclitic (typeset in boldscript in the examples), being itself a morph list consisting of a single element, inserts between the first and second element of the verb's morph list.

$$\begin{bmatrix} \text{MORPHS} \left\langle xward \right\rangle \oplus \left\langle \boldsymbol{man} \right\rangle \oplus \left\langle in \right\rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} - perfect \\ \text{TENSE } preterite \\ \text{POLARITY } positive \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{MORPHS} \left\langle na \right\rangle \oplus \left\langle \boldsymbol{man} \right\rangle \oplus \left\langle xward \right\rangle \oplus \left\langle in \right\rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} - perfect \\ \text{TENSE } preterite \\ \text{POLARITY } negative \end{bmatrix} \end{bmatrix}$$

Yet, whenever there is a third singular person marker, it always follows the object marker. This is modelled by a morphological compact operation creating opaque morph-sequences whose inner structure is no longer accessible to the endoclitic MPM. When confronted with compact units (or *morphological stems),* MPMs must attach on the outside. They do so in second position, as specified in the general rule *a).* The relevant operation is given below.[4]

[4] The formal representation of the compact operation given in this paper slightly differs from the representation presented by Bonami and Samvelian (2008). In (Bonami and Samvelian, 2008), compact operates on morph-lists. In this paper we propose a simplified representation that allows for representing the operation and the attached constraints in one feature structure (cf. [SUBJECT [[PERSON *3*] ∧ [NUMBER *sing*]]]).

(6) xward-in=**î**
eat.STEM-3PL=**3SG**
"He ate them."

(7) na-xward-in=**î**
NEG-eat.STEM-3PL=**3SG**
"He did not them."

$$\texttt{compact}\left(\left[\begin{array}{l}\text{MORPHS}\left\langle \textit{xward, in}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT}-\textit{perfect}\\ \text{TENSE}\ \textit{preterite}\\ \text{POLARITY}\ \textit{positive}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON}\ \textit{3}\\ \text{NUMBER}\ \textit{sing}\end{array}\right]\end{array}\right]\right)\longrightarrow\left[\begin{array}{l}\text{MORPHS}\left\langle \textit{xwardin}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT}-\textit{perfect}\\ \text{TENSE}\ \textit{preterite}\\ \text{POLARITY}\ \textit{positive}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON}\ \textit{3}\\ \text{NUMBER}\ \textit{sing}\end{array}\right]\end{array}\right]$$

The complete verb form thus corresponds to the following representation.

$$\left[\begin{array}{l}\text{MORPHS}\left\langle \textit{xwardin}\right\rangle\oplus\left\langle \textit{î}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT}-\textit{perfect}\\ \text{TENSE}\ \textit{preterite}\\ \text{POLARITY}\ \textit{positive}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON}\ \textit{3}\\ \text{NUMBER}\ \textit{sing}\end{array}\right]\end{array}\right]$$

If the verb form also contains prefixes, as for example the negative verb forms hereafter, the compact operation not only clusters the stem to the person marker but also includes the prefix(es), as illustrated below.

$$\texttt{compact}\left(\left[\begin{array}{l}\text{MORPHS}\left\langle \textit{na, xward, in}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT}-\textit{perfect}\\ \text{TENSE}\ \textit{preterite}\\ \text{POLARITY}\ \textit{negative}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON}\ \textit{3}\\ \text{NUMBER}\ \textit{sing}\end{array}\right]\end{array}\right]\right)\longrightarrow\left[\begin{array}{l}\text{MORPHS}\left\langle \textit{naxwardin}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT}-\textit{perfect}\\ \text{TENSE}\ \textit{preterite}\\ \text{POLARITY}\ \textit{negative}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON}\ \textit{3}\\ \text{NUMBER}\ \textit{sing}\end{array}\right]\end{array}\right]$$

Hence the following representation for the resulting verb form.

$$\begin{bmatrix} \text{MORPHS} \left\langle \textit{naxwardin} \right\rangle \oplus \left\langle \hat{\textit{i}} \right\rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} - \textit{perfect} \\ \text{TENSE } \textit{preterite} \\ \text{POLARITY } \textit{negative} \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{PERSON } \textit{3} \\ \text{NUMBER } \textit{sing} \end{bmatrix} \end{bmatrix}$$

3.3. Extending the existing analysis

In addition to the two constraints formalised by Bonami and Samvelian (2008), McCarus (1958) cites the following constraint, which also overrides the general rule *a).*

c) Be it a subject or object marker, the first singular person marker always precedes any plural marker.

(8) xward=m=**in**
eat.STEM=P1PL=**P3PL**
"I ate them."

(9) xward-im=**tan**
eat.STEM-P1SG=**P2PL**
"You ate me."

(10) na=**m**=xward-in
NEG=**P1SG**= eat.STEM=P3PL
"I didn't eat them."

(11) na-xward-im=**tan**
NEG-eat.STEM=P1SG=**P3PL**
"You didn't eat me."

The second rule is not taken into account in (Bonami and Samvelian, 2008), but doing so only requires two minor modifications of their model and does not invalidate their analysis *per se.*

1) When the present tense verb has a first person singular object and a plural subject, or when a past tense verb has a first person singular subject and a plural object,[5] the placement of the clitic occurs according to the general case placement, i.e., in second position of a morph-list containing solely monomorphic elements. Hereafter we provide examples for a positive, respectively negative, preterite verb form.

[5] Obviously, since the object is plural and thus not a third person singular, rule *b)* does not apply either.

Positive verb form:

$$\begin{bmatrix} \text{MORPHS} \langle xward \rangle \oplus \langle \boldsymbol{m} \rangle \oplus \langle in \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} - perfect \\ \text{TENSE } preterite \\ \text{POLARITY } positive \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{PERSON } 1 \\ \text{NUMBER } sing \end{bmatrix} \\ \text{OBJECT} \begin{bmatrix} \text{NUMBER } pl \end{bmatrix} \end{bmatrix}$$

Negative verb form:

$$\begin{bmatrix} \text{MORPHS} \langle na \rangle \oplus \langle \boldsymbol{m} \rangle \oplus \langle xward \rangle \oplus \langle in \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} - perfect \\ \text{TENSE } preterite \\ \text{POLARITY } negative \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{PERSON } 1 \\ \text{NUMBER } sing \end{bmatrix} \\ \text{OBJECT} \begin{bmatrix} \text{NUMBER } pl \end{bmatrix} \end{bmatrix}$$

2) Whenever there is a first person singular subject and a plural object in the present or a first person singular object and a plural subject in the past, another compact operation, similar to the one for rule *b),* applies. Stipulating this additional compact operation suffices to account for the cluster consisting of the stem and the first person singular object affix.

$$\texttt{compact}\left(\begin{bmatrix} \text{MORPHS} \langle xward,\ im \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} - perfect \\ \text{TENSE } preterite \\ \text{POLARITY } positive \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{NUMBER } pl \end{bmatrix} \\ \text{OBJECT} \begin{bmatrix} \text{PERSON } 1 \\ \text{NUMBER } sing \end{bmatrix} \end{bmatrix}\right) \longrightarrow \begin{bmatrix} \text{MORPHS} \langle xwardim \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} - perfect \\ \text{TENSE } preterite \\ \text{POLARITY } positive \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{NUMBER } pl \end{bmatrix} \\ \text{OBJECT} \begin{bmatrix} \text{PERSON } 1 \\ \text{NUMBER } sing \end{bmatrix} \end{bmatrix}$$

For the verb forms with prefixes, such as a negative verb form, the compact operation generates clusters containing the prefix(es), the stem, and the first person singular affix, just as for modelling Example (7).

$$\texttt{compact}\left(\begin{bmatrix}\text{MORPHS}\left\langle na,\ xward\ ,im\right\rangle\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT}-perfect\\ \text{TENSE}\ preterite\\ \text{POLARITY}\ negative\end{bmatrix}\\ \text{SUBJECT}\begin{bmatrix}\text{NUMBER}\ pl\end{bmatrix}\\ \text{OBJECT}\begin{bmatrix}\text{PERSON}\ 1\\ \text{NUMBER}\ sing\end{bmatrix}\end{bmatrix}\right)\longrightarrow\begin{bmatrix}\text{MORPHS}\left\langle naxwardim\right\rangle\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT}-perfect\\ \text{TENSE}\ preterite\\ \text{POLARITY}\ negative\end{bmatrix}\\ \text{SUBJECT}\begin{bmatrix}\text{NUMBER}\ pl\end{bmatrix}\\ \text{OBJECT}\begin{bmatrix}\text{PERSON}\ 1\\ \text{NUMBER}\ sing\end{bmatrix}\end{bmatrix}$$

Thus the complete verb forms from Examples (9) and (11) can be represented as follows.

Positive verb form:

$$\begin{bmatrix}\text{MORPHS}\left\langle xwardim\right\rangle\oplus\left\langle \boldsymbol{tan}\right\rangle\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT}-perfect\\ \text{TENSE}\ preterite\\ \text{POLARITY}\ positive\end{bmatrix}\\ \text{SUBJECT}\begin{bmatrix}\text{PERSON}\ 2\\ \text{NUMBER}\ pl\end{bmatrix}\\ \text{OBJECT}\begin{bmatrix}\text{PERSON}\ 1\\ \text{NUMBER}\ sing\end{bmatrix}\end{bmatrix}$$

Negative verb form:

$$\begin{bmatrix}\text{MORPHS}\left\langle naxwardim\right\rangle\oplus\left\langle \boldsymbol{tan}\right\rangle\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT}-perfect\\ \text{TENSE}\ preterite\\ \text{POLARITY}\ negative\end{bmatrix}\\ \text{SUBJECT}\begin{bmatrix}\text{PERSON}\ 2\\ \text{NUMBER}\ pl\end{bmatrix}\\ \text{OBJECT}\begin{bmatrix}\text{PERSON}\ 1\\ \text{NUMBER}\ sing\end{bmatrix}\end{bmatrix}$$

4. Accounting for the perfect tenses

4.1 The yet unaccounted data

Bonami and Samvelian (2008) state that MPMs always adopt a second position, be it internal (i.e., in the morph-list) or external (i.e., in the VP). They explicitly choose to analyse only verb forms in the present and non-perfect past tenses. After extending their analysis with constraint *c)*, the resulting model holds for all transitive verb forms they analyse. Yet, if one takes into account the whole paradigm, including the perfect tenses, the situation gets much trickier. As shown in Examples (12) to (15), in perfect tenses,

MPMs seem to appear in third (or fourth) position, after the perfect maker *–û* or the past perfect marker *–bû* respectively.

(12) xward-û=**man**=in
eat.STEM-û=**P1PL**=P3PL
"We have eaten them."

(13) xward-û-in=**î**
eat.STEM-û-P3PL=**P3SG**
"He has eaten them."

(14) xward-bû=**man**=in
eat.STEM-bû=P1PL=**P3PL**
"We had eaten them."

(15) xward-bû-in=**î**
eat.STEM-bû-P3PL=**P3SG**
"He had eaten them."

4.2. A periphrastic analysis of the perfect tenses

This contradiction seems to disappear if the perfect tenses are analysed as periphrastic. Given the apparent, and at least historically confirmed, presence of forms of the auxiliary *bûn* 'to be' in the perfect tenses, this looks plausible. Indeed, the personal endings used for the present perfect (PE3) are equivalent to the enclitic forms of the verb *bûn* 'to be'. Thus the present perfect forms can be re-analysed as the past participle (i.e., STEM2 + *–û)* plus the inflected forms of the auxiliary *bûn*, which match the personal endings used for the present perfect (PE3). In the past perfect, we can make out the verb *bûn* in the past tense: the past perfect appears like a combination of a STEM and the past forms of *bûn* (*bû–* + PE2). Thus the present and past perfect forms, whose features only differ in tense, indeed show auxiliary forms that also only differ in tense, since for building the past perfect the auxiliary *bûn* is simply used in the past tense.

In such a periphrastic analysis, we would simply have to say that MPMs attach in auxiliary- second-position for the past perfect and as external second position clitics in the (present) perfect.

(16) [$_V$ xward] [$_V$ bû=**man**=in]
[$_V$ eat.STEM] [$_V$ AUX.STEM =**MPM.P1PL**=P3PL]
"We had eaten them." (PAST PERFECT)

(17) [$_V$ xward-û]=**man**=[$_{AUX}$ in]
[$_V$ eat.STEM-*û*]=**MPM.P1PL**=[$_{AUX}$ AUX.P3PL]
"*We have eaten them."* (PRESENT PERFECT)

4.3. Drawbacks of a periphrastic analysis of Sorani perfect tenses

However, this analysis runs into difficulties when considering the following facts:[6]

[6] While the indicative present perfect tenses seem to constitute a convincing case for canonical periphrasis (due to the presence of synthetic subjunctive forms), nothing in the verbal paradigm raises expectations as to the existence of past perfect forms. Their periphrasticity would therefore be at best non-canonical in the sense of Brown *et al.* (2012). Yet, one would be tempted to state that if periphrastic present perfect forms exist and if there are forms for the past perfect, then the latter

1. In present perfect constructions, MPMs indeed attach to the first element of the VP containing the two elements of the periphrastic construction, i.e., it attaches to the past participle. In the past perfect however, they attach within the auxiliary, in second position (see Example (12) *vs.* Example (14)). Thus, the two constructions seem to be different in nature. Instead of simplifying the analysis, assuming a periphrastic construction for the perfect tense forms would require stipulating separate rules for the two perfect tenses, the present perfect following the enclitic placement rule of VP second position enclisis, while the past perfect would follow the verb internal endoclitic placement rules.

2. The periphrastic analysis does correctly account for most perfect tense forms. However, it does not explain the reversed order of the MPMs and PE3, respectively PE2, for the specific cases due to person placement (see Examples (18) and (19)). Another compact operation would be needed to prevent the MPM to attach to the participle in (18) or to the auxiliary stem in (19) directly. But if we consider the forms as periphrastic, compact would not simply operate over morphs but over words or word-like elements, such as the participle in (18). Yet, this is problematic in an approach where we assume autonomous morphology (Aronoff, 1994).

(18) [$_{V}$ xward-û]=[$_{AUX}$ n]=**î**
[$_{V}$ eat.STEM-*û*]=[$_{AUX}$ AUX.P3PL]=**MPM.P3SG**
"He has eaten them." (PRESENT PERFECT)

(19) [$_{V}$ xward] [$_{AUX}$ bû=n]=**î**
[$_{V}$ eat.STEM] [$_{AUX}$ AUX.STEM =AUX.P3PL=**MPM.P3SG**]
"He had eaten them". (PAST PERFECT)

3. A lexicalist approach to morphosyntax entails that morphology and syntax do not submit to rules of the same nature. They are necessarily distinct modules of language. It is hence difficult to even explain the status of the STEM2 within the past perfect constructions. Elements treated within syntax do not undergo the same rules as morphs, they have to be words or word-like elements such as clitics. A bare STEM, such as *xward* in the past perfect tenses, makes for an extremely unlikely syntactic element.[7]

4. Moreover, when perfect tense forms bear a negative polarity feature, MPMs are inserted in internal second position,[8] and not within the alleged auxiliary. This fact alone strongly favours a synthetic analysis.

(20) na=**man**=xward-û=a
NEG=**MPM.1PL**= eat.STEM-PART=P3SG
"We have not eaten it."

(21) na=**man**=xward-bû=in
NEG=**MPM.1PL**=eat.STEM-BÛ=P3SG
"We had not eaten them."

are typologically highly unlikely to be anything but periphrastic. Since the auxiliary *bûn* in the past tense seems distinguishable among the elements constituting the past perfect, positing a periphrastic past perfect seems quite reasonable.

[7] One might suggest the existence of apocoped infinitives in Sorani, like those existing in Persian that indeed consist of bare stems (STEM2) (Lazard *et al.,* 2006). Yet this analysis would again run into difficulties due to the different behaviour of the MPMs in the present and past perfect. One would still expect the MPMs to attach to either type of participle.

[8] Except, of course, for the specific placement of third person singular and first person singular markers.

5. Finally, in the past perfect, the STEM and *–bû* are never separated by any morph, even though MPMs may occur immediately before STEM2 and after *–bû.* Thus no morph ever inserts between the presumed predicative element and the auxiliary. This systematic closeness between the two elements does not fit into an account where they would be part of two syntactically independent elements.

5. Completing the formal account of Sorani Kurdish MPMs

5.1 Completing the inventory of compact operations

From the evidence of the preceding section, we can deduce that Sorani perfect tenses are better analysed as synthetic forms. Nevertheless, it appears that these synthetic forms exhibit internal structure, which explains MPM placement properties.

We follow Bonami and Samvelian (2008) in resorting to the morphological compact operation that creates morph-sequences opaque even to the otherwise endoclitic MPMs of Sorani Kurdish. These compact operations apply to STEM2-PE2 or STEM2-PE3 sequences, as shown by Bonami and Samvelian (2008), but also affect the STEM2–*û* and STEM2–*bû* sequences of the perfect tenses, thus creating internal structure.

This approach allows for a homogeneous treatment of all perfect tenses, which given the data seems much more satisfying.

$$\texttt{compact}\left(\left[\begin{array}{ll}\text{MORPHS} & \langle \textit{xward, û} \rangle \\ \text{FEATURES} & \left[\begin{array}{l}\text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{present}\end{array}\right]\end{array}\right]\right) \longrightarrow \left[\begin{array}{ll}\text{MORPHS} & \langle \textit{xwardû} \rangle \\ \text{FEATURES} & \left[\begin{array}{l}\text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{present}\end{array}\right]\end{array}\right]$$

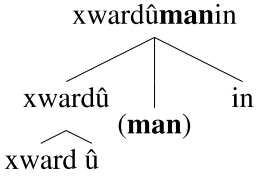
$$\texttt{compact}\left(\left[\begin{array}{ll}\text{MORPHS} & \langle \textit{xward, bû} \rangle \\ \text{FEATURES} & \left[\begin{array}{l}\text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{past}\end{array}\right]\end{array}\right]\right) \longrightarrow \left[\begin{array}{ll}\text{MORPHS} & \langle \textit{xwardbû} \rangle \\ \text{FEATURES} & \left[\begin{array}{l}\text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{past}\end{array}\right]\end{array}\right]$$

When confronted with compact units, an MPM cannot make out any internal structure and must thus attach on the outside. The formal representation of the complete verb forms corresponding to above quoted examples (12) to (15) are given below with their corresponding tree structure representations.[9] These trees show the extent of morphological structure created by the compact operations for each type of verb form. The structure available as frontiers for endoclitic insertion corresponds to the first nodes under the tree roots.

(22) xward-û=**man**=in
eat-PART=**P1PL**=P3PL
"We have eaten them." (PRESENT PERFECT)

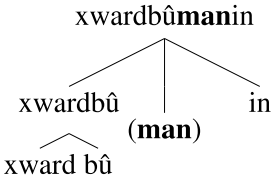
[9] The MPMs are typeset in bold print, as well as noted between brackets so as to express the fact that they insert in the final morph list as endoclitics.

$$\begin{bmatrix} \text{MORPHS} \langle \textit{xwardû} \rangle \oplus \textbf{\textit{<man>}} \oplus \langle \textit{in} \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{present} \end{bmatrix} \end{bmatrix}$$



(23) xward-bû=**man**=in
eat-PASTPERF=**P1PL**=P3PL
"We had eaten them." (PAST PERFECT)

$$\begin{bmatrix} \text{MORPHS} \langle \textit{xwardbû} \rangle \oplus \textbf{\textit{<man>}} \oplus \langle \textit{in} \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{past} \end{bmatrix} \end{bmatrix}$$

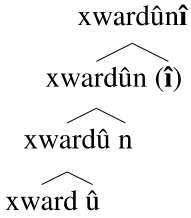


Whenever one of the specific rules applies, the corresponding compact operations described above are activated on top of the perfect compact operations.

(24) xward-û-n=**î**
eat.STEM-PART-P3PL=**P3SG**
"He has eaten them."

$$\texttt{compact}\left(\begin{bmatrix} \text{MORPHS} \langle \textit{xwardû, (i)n} \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{present} \\ \text{POLARITY } \textit{positive} \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{PERSON } \textit{3} \\ \text{NUMBER } \textit{sing} \end{bmatrix} \end{bmatrix}\right) \longrightarrow \begin{bmatrix} \text{MORPHS} \langle \textit{xwardûn} \rangle \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{present} \\ \text{POLARITY } \textit{positive} \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{PERSON } \textit{3} \\ \text{NUMBER } \textit{sing} \end{bmatrix} \end{bmatrix}$$

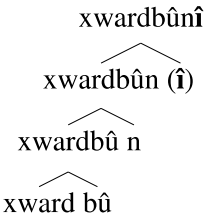
$$\begin{bmatrix} \text{MORPHS} \langle \textit{xwardûn} \rangle \oplus \textbf{\textit{<î>}} \\ \text{FEATURES} \begin{bmatrix} \text{ASPECT} + \textit{perfect} \\ \text{TENSE } \textit{present} \\ \text{POLARITY } \textit{positive} \end{bmatrix} \\ \text{SUBJECT} \begin{bmatrix} \text{PERSON } \textit{3} \\ \text{NUMBER } \textit{sing} \end{bmatrix} \end{bmatrix}$$



(25) xward-bû-n=**î**
eat.STEM-BÛ- P3PL=**P3SG**
"He had eaten them."

$$\texttt{compact}\left(\begin{bmatrix}\text{MORPHS}\left\langle \textit{xwardbû, (i)n}\right\rangle\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT + }\textit{perfect}\\ \text{TENSE }\textit{past}\\ \text{POLARITY }\textit{positive}\end{bmatrix}\\ \text{SUBJECT}\begin{bmatrix}\text{PERSON }\textit{3}\\ \text{NUMBER }\textit{sing}\end{bmatrix}\end{bmatrix}\right)\longrightarrow\begin{bmatrix}\text{MORPHS}\left\langle \textit{xwardbûn}\right\rangle\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT + }\textit{perfect}\\ \text{TENSE }\textit{past}\\ \text{POLARITY }\textit{positive}\end{bmatrix}\\ \text{SUBJECT}\begin{bmatrix}\text{PERSON }\textit{3}\\ \text{NUMBER }\textit{sing}\end{bmatrix}\end{bmatrix}$$

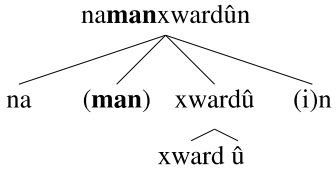
$$\begin{bmatrix}\text{MORPHS }\textit{<xwardbûn>}\oplus\textbf{\textit{<î>}}\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT + }\textit{perfect}\\ \text{TENSE }\textit{past}\\ \text{POLARITY }\textit{positive}\end{bmatrix}\\ \text{SUBJECT}\begin{bmatrix}\text{PERSON }\textit{3}\\ \text{NUMBER }\textit{sing}\end{bmatrix}\end{bmatrix}$$



Contrarily to the clustering occurring between the prefixes, the stem and the PE2, respectively PE3 person markers, the perfect tenses do not cluster by default with the prefixes. Thus formalisation of perfect tense negative verb forms, where the MPM occurs between the negative prefix and the stem, for example, runs as follows.

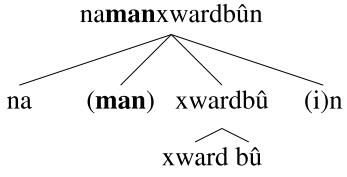
(26) na=**man**=xward-û=n
NEG= **P1PL**= eat.STEM-PART= P3PL
"We have not eaten them."

$$\begin{bmatrix}\text{MORPHS }\textit{<na>}\oplus\textbf{\textit{<man>}}\oplus\textit{<xwardû>}\oplus\textit{<(i)n>}\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT + }\textit{perfect}\\ \text{TENSE }\textit{present}\\ \text{POLARITY }\textit{negative}\end{bmatrix}\end{bmatrix}$$



(27) na=**man**=xward-bû=n
NEG= **P1PL**= eat.STEM-BÛ= **P3PL**
"We had not eaten them."

$$\begin{bmatrix}\text{MORPHS }\textit{<na>}\oplus\textbf{\textit{<man>}}\oplus\textit{<xwardbû>}\oplus\textit{<(i)n>}\\ \text{FEATURES}\begin{bmatrix}\text{ASPECT + }\textit{perfect}\\ \text{TENSE }\textit{past}\\ \text{POLARITY }\textit{negative}\end{bmatrix}\end{bmatrix}$$



Perfect clusters containing object affixes, however, do cluster with the prefixes, just as their non-perfect counterparts.[10] This clustering is simply triggered by the operations occurring when constraints for *b)* or *c)* are met (cf. Section 4).

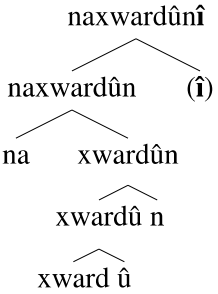
$$\texttt{compact}\left(\left[\begin{array}{l}\text{MORPHS}\left\langle \textit{na, xwardû, (i)n}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT } \textit{+ perfect}\\ \text{TENSE } \textit{present}\\ \text{POLARITY } \textit{negative}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON } \textit{3}\\ \text{NUMBER } \textit{sing}\end{array}\right]\end{array}\right]\right) \longrightarrow \left[\begin{array}{l}\text{MORPHS}\left\langle \textit{naxwardûn}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT } \textit{+ perfect}\\ \text{TENSE } \textit{present}\\ \text{POLARITY } \textit{negative}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON } \textit{3}\\ \text{NUMBER } \textit{sing}\end{array}\right]\end{array}\right]$$

$$\texttt{compact}\left(\left[\begin{array}{l}\text{MORPHS}\left\langle \textit{na, xwardbû, (i)n}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT } \textit{+ perfect}\\ \text{TENSE } \textit{past}\\ \text{POLARITY } \textit{negative}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON } \textit{3}\\ \text{NUMBER } \textit{sing}\end{array}\right]\end{array}\right]\right) \longrightarrow \left[\begin{array}{l}\text{MORPHS}\left\langle \textit{naxwardbûn}\right\rangle\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT } \textit{+ perfect}\\ \text{TENSE } \textit{past}\\ \text{POLARITY } \textit{negative}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON } \textit{3}\\ \text{NUMBER } \textit{sing}\end{array}\right]\end{array}\right]$$

Thus the formal representation for negative perfect forms with morphological stems containing object affixes is as follows.

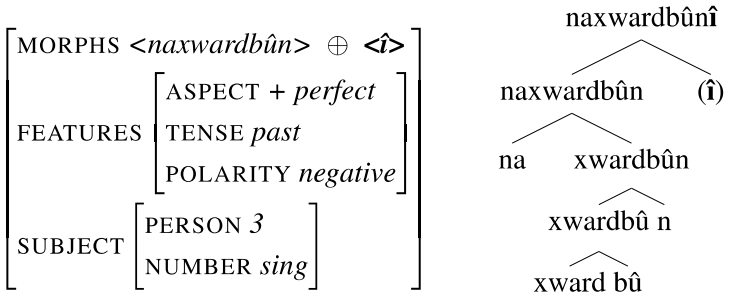
(28) na-xward-û-n=**î**
NEG- eat.STEM-PART- P3PL=**P3SG**
"He has not eaten them."

$$\left[\begin{array}{l}\text{MORPHS } \textit{<naxwardûn>} \oplus \textbf{<î>}\\ \text{FEATURES}\left[\begin{array}{l}\text{ASPECT } \textit{+ perfect}\\ \text{TENSE } \textit{present}\\ \text{POLARITY } \textit{negative}\end{array}\right]\\ \text{SUBJECT}\left[\begin{array}{l}\text{PERSON } \textit{3}\\ \text{NUMBER } \textit{sing}\end{array}\right]\end{array}\right]$$



(29) na-xward-bû-n=**î**
NEG- eat.STEM-PART-P3PL=**P3SG**
"He had not eaten them."

[10] The examples given here are for the third person singular subject rule *b).* The same kind of clustering would of course also occur with the first person singular subject or object *vs.* plural person marker, rule *c).*



5.2. Summary and respective ordering of the compact operations

The observations and analyses presented above indicate that Sorani clitic MPMs present the following placement properties.

1. If VP is not verb initial, the MPM is an enclitic attaching in second position within the sentence (Samvelian, 2007).
2. If a VP is verb initial, the MPM attaches in verb-internal second position, according to the following constraints.
   a) By default, the MPM attaches after the first element of the morph list, i.e. either the first prefix, or, if there are no prefixes, after the stem.
   b) For the perfect tenses, the stem and the perfect marker (be it present perfect *–û* or past perfect *–bû*) are clustered into one element in the morph-list. They are called morphological stems by Bonami and Samvelian (2008). This constraint is formally encoded by the compact operation A below.
   c) The third singular person marker (be it standard or MPM) always follows the object marker. This constraint is formally encoded by the compact operation B below.
   d) Be it a subject or object marker, the first singular person marker always precedes any plural marker. This constraint is also formally handled by the compact operation B below.
   e) Whenever a verb form contains prefixes and undergoes the compact operation described in *c)* or *d)* above, the prefixes are also clustered to the stem and the person marker during compact operation B.

The two* compact *operations

A. STEM + *û* → < STEM*û* >
   STEM + *bû* → < STEM*bû* >

After applying compact A, <STEM*û* >, respectively <STEM*û* >, are considered morphological stems. In particular they count as STEM in compact operation B below.

B. STEM + PE → <STEM-PE >
   da + STEM + PE → <*da*STEM-PE >
   bi + STEM + PE → <*bi*STEM-PE >
   nû + STEM + PE → <*nû*STEM-PE >
   na + STEM + PE → <*na*STEM-PE >
   na +da + STEM + PE → <*nada*STEM-PE >
   bi + STEM + PE → <*bi*STEM-PE >

Ordering of the placement constraints

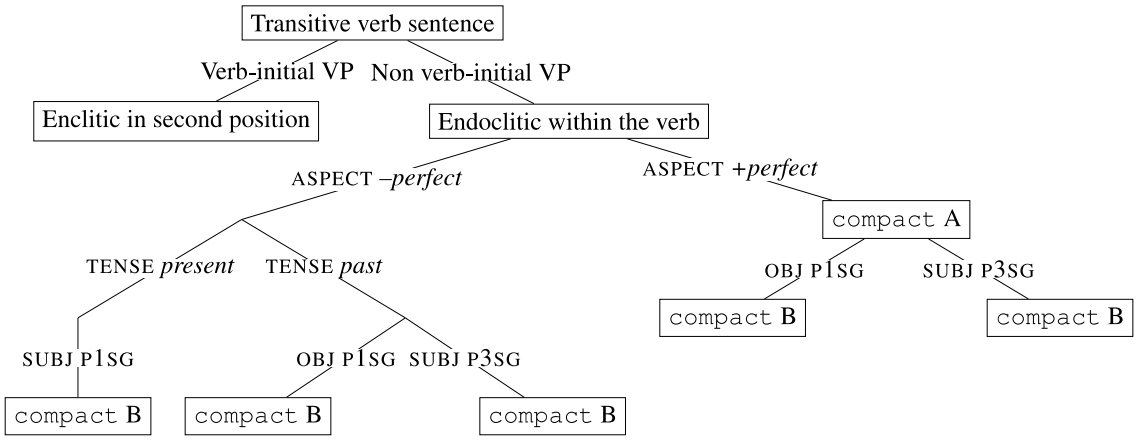
Figure 1 below gives a summary representation the possible combinations of the four compact operations depending on the criteria described above.

Given a sentence containing a transitive verb, the VP can be either verb initial or not verb initial.

If the VP is not verb-initial, the clitic placement follows the pattern in the left first level node: it attaches as an enclitic in second position within the VP.

If the VP is verb initial, the clitic is an endoclitic within the verb and follows the endoclitic placement constraints on the right part of the tree.

If any of the contraints indicated on the tree branches are met, the verb will undergo one of the corresponding nodes' compact operations.



For example, within a verb-initial VP, a past perfect verb having positive polarity, a third person singular subject and a third person plural object, such as *xwardbûinî* 'he will eat them', will follow the following path:

- The initial unclustered morphs are the following: *xward, –bû–, –in–, =î.*
- The verb *xwardbûinî* is part of verb-initial VP. Thus the first path to the right applies.
- It has perfect aspect, thus compact A applies.
  The first clustering outputs *xwardbû, –in–, =î.*
- It has a third person singular subject, thus compact B applies.
  The second clustering outputs *xwardbûin.* To this output the endoclitic *=î* attaches as *xwardbûin=î.*

Note that a past perfect verb having positive polarity does not contain any prefixes. Thus compact B does only affect the stem and the third person singular object affix.

Within a verb-initial VP, a present verb with negative polarity and a first person singular subject and third person plural object, such as *nâxwardimyân* 'he will eat them', will follow the following path:

- The initial unclustered morphs are the following: *nâ–, xward, –im–, =yân.*
- The verb *nâxwardimyân* is part of verb-initial VP. Thus the first path to the right applies.
- It has no perfect aspect, thus compact A does not apply. We follow the second path to the right.
- It has a first person singular subject, thus compact B applies. Compact B here clusters the negative prefix *nâ–*, the stem and the first person singular subject affix
  –im–. The endoclitic *=yân* attaches in the thus new available second position.

The clustered output thus contains *nâxwardim.* Thus the endoclitic *=î* attaches to it as *nâxwardim=yân.*

6. Concluding remarks

We have presented an enhanced formal analysis of Sorani Kurdish endoclitic placement. This analysis extends the analysis and formalisation given by Bonami and Samvelian (2008) by taking into account an additional constraint that had been missed and by including present and past perfect verb forms that had not been studied so far.

In our formalisation, the compact operation used by Bonami and Samvelian (2008) is not limited to clusters containing the negative prefix and/or third person singular subject markers. It also systematically applies to first person singular person marking and to *û/bû* elements, i.e., to clusters containing the perfect and past perfect markers.

Our analysis allows for a non-periphrastic and homogeneous treatment of all Sorani perfect tenses that correctly accounts for the intricate MPM placement phenomena left unexplained until now.

We have shown that a structured representation is needed to account for the clitic placement of MPMs and the morphology-syntax interface in Sorani. Thus, the properties of Sorani MPMs formalised in this paper also argue in favour of partial internal morphological structure.

References

Anderson, Stephen R. 1992. *A-morphous Morphology.* Cambridge, UK: Cambridge University Press.
Anderson, Stephen R. 2005. *Aspects of the Theory of Clitics.* Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford.
Aronoff, Mark. 1994. *Morphology by Itself.* MIT Press.
Baerman, Matthew. 2007. Morphological reversals. *Journal of Linguistics,* 43, 33–61.
Baker, Mark C. 1988. *Incorporation.* Chicago, USA: University of Chicago Press.
Beard, Robert. 1995. *Lexeme-Morpheme Base Morphology.* Albany, N.Y.: State University of New York Press.
Bedir Khan, Emir Djeladet and Roger Lescot. 1970. *Grammaire kurde (Dialecte Kurmandji).* Paris, France: Librairie d'Amérique et d'Orient.
Beidar, Paul. 1926. *Grammaire kurde.* Paris, France: Librairie Orientaliste P. Garthner.
Blau, Joyce. 1980. *Manuel de kurde, dialecte sorani: grammaire, textes de lecture, vocabulaire kurde-français et français-kurde.* Paris, France: C. Klincksieck.
Blau, Joyce. 2000. *Méthode de kurde sorani.* Paris, France: L'Harmattan.
Bonami, Olivier and Pollet Samvelian. 2008. Sorani Kurdish person markers and the typology of agreement. In *13th International Morphology Meeting,* Vienna, Austria.
Bresnan, Joan. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language and Language Theory.*
Brown, Dunstan, Marina Chumakina, Greville G. Corbett, Gergana Popova, and Andrew Spencer. 2012. Defining 'periphrasis': key notions. To be published.
Brown, Dunstan and Andrew Hippisley. 2012. *Network Morphology: A Defaults-based Theory of Word Structure.* Cambridge University Press.
Corbett, Greville G. and Norman Fraser. 1993. Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics,* 29, 113–142.
Crysmann, Berthold. 2002. *Constraint-Based Coanalysis: Portuguese Cliticisation and Morphology-Syntax Interaction in HPSG.* Number 15 in Saarbrücken Dissertations in Computational Linguistics and Language Technology. Saarbrücken: Deutsches Forschungszentrum für Künstliche Intelligenz and Universität des Saarlandes.
Crysmann, Berthold. 2006. Floating affixes in Polish. In S. Müller, Ed., *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar,* p. 123–139, Stanford: CSLI Publications.
Di Sciullo, Anna-Maria and Edwin Williams. 1987. *On the Definition of Word.* Cambridge, MA, USA: MIT Press.
Garzoni, Maurizio. 1787. *Grammatica e Vocabulario della Lingua Kurda.* Rome, Italy: Sacra Congregazione di Propaganda Fide.
Haig, Geoffrey. 2010. *Alignment change in Iranian languages: a Construction Grammar approach.* De Gruyter.
Harris, Alice. 2002. *Endoclitics and the Origins of Udi Morphosyntax.* Oxford, UK: Oxford University Press.
Justi, Ferdinand. 1880. *Kurdische Grammatik.* St. Pertersburg, Russian Empire: Commissionäre der Kaiserlichen Akademie der Wissenschaften, Eggers & C.U.J. Glasunow.
Kupść, Anna and Jesse Tseng. 2005. A new HPSG approach to Polish auxiliary constructions. In S. Müller, ed., *The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar, Department of Informatics, University of Lisbon, August 23-24, 2005,* pages 253–273. Stanford: CSLI Publications.
Lazard, Gilbert, Yann Richard, Rokhsareh Hechmati, and Pollet Samvelian. 2006. *Grammaire du persan contemporain.* Teheran, Iran: Institut Français de Recherche en Iran & Farhang Moaser Edition.
Lieber, Rochelle. 1992. *Deconstructing Morphology: Word Formation in Syntactic Theory.* Chicago, USA: University of Chicago Press.
Lieber, Rochelle. 2006. Syntax of words. In *The Encyclopedia of Language and Linguistics,* p. 405–408. Elsevier.
MacKenzie, David N. 1961. *Kurdish dialect studies.* London Oriental Series. London, UK: Oxford Univ. Press.
McCarus, Ernest N. 1958. *A Kurdish Grammar: descriptive analysis of the Kurdish of Sulaimaniya, Iraq.* PhD thesis, American Council of Learned Societies, New-York, USA.

Samvelian, Pollet. 2007. What Sorani Kurdish absolute prepositions tell us about cliticization. In F. Hoyt, N. Seifert and A. Teodorescu, Eds., *Texas Linguistics Society IX: The Morphosyntaxe of Understudied Languages,* Stanford, USA: CSLI online publications.

Soane, E.B. 1913. *Kurdish Grammar (Grammar of the Kurmandj or Kurdish Dialect).* Luzac & co.

Stump, G. T. 2001. *Inflectional Morphology. A Theory of Paradigm Structure.* Cambridge, UK: Cambridge University Press.

Thackston, W.M. 2006. Sorani Kurdish: A reference grammar with selected readings. Published online.

Wahby, Towfiq. 1956. *The Kurdish Grammar,* volume 1. Baghdad, Iraq: Librairie d'Amérique et d'Orient.

Walther, Géraldine. 2011. A derivational account for Sorani Kurdish passives. Communication at the 4th International Conference on Iranian Linguistics (ICIL4). June 17th-19th 2011. Uppsala, Sweden.

Zwicky, Arnold M. 1977. On Clitics. Indiana University Linguistics Club, 1977.

Zwicky, Arnold M. 1985. How to describe inflection. In M. Niepokuj, M. V. Clay, V. Nikiforidou and D. Feder, Eds., *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistic Society,* p. 372–386, Berkeley,USA: Berkeley Linguistic Society.

Empirical and theoretical aspects of phrasal compounds: against the "syntax explains it all" attitude

Carola Trips
University of Mannheim
ctrips@rumms.uni-mannheim.de

1. Introduction

In the literature, some attention has been paid to phrasal compounds (PC) especially since they seem to pose problems for a theory of morphology that is based on a classical generative, syntactocentric framework. If instances of Noun+Noun compounding (NNC) are compared with instances of PCs (examples in (1) vs. (2)) structural differences become evident immediately:

(1) a. peanut butter
b. love letter

(2) a. She also knows that the media tendency to lump together women singer-songwriters in a *"gee whiz, gosh, women are now making it'' syndrome* is patronising, if not pernicious. (BNC, A7S190)
b. Bombay-based Anil put India's failure to exploit its manpower and mind power and its lack of excellence in sport, economics and the arts down to a *"Learn what is there and don't question it'' attitude* (BNC, HAE4088)

What makes these compounds so special is that the left-hand member is a complex, maximal phrase: as in the examples given above, it can be a whole sentence like an IP (or CP depending on the analysis), which clearly sets them apart from NNCs, the left-hand member of which is non-phrasal and thus an entity on the word level.

Concerning the theoretical analysis of PCs, it has been shown that they are problematic in a generative framework, which assumes that on the one hand language can be clearly divided into a lexicon and a grammar and that on the other hand syntax determines the combinatorial properties of phonology and semantics (Chomsky:1965, Chomsky:1981). *The No Phrase* constraint (Botha:1981) and the *Lexical Integrity Hypothesis* (Lapointe:1980) are a result of these assumptions on the level of morphology which means that PCs will always violate these constraints as they show that syntactic phrases do have access to word structure. I will show that an analysis based on the semantics of PCs and allowing for a parallel production and processing of semantic, phonological and semantic structure like Jackendoff's model of Parallel Architecture (PA, Jackendoff 1997, 2002, Culicover & Jackendoff 2005) fares much better in this respect than analyses hitherto proposed (e.g. Botha 1981, Gallmann 1990, Lieber 1992, Wiese 1996, Ackema & Neleman 2004, Štekauer & Lieber 2009).

The aim of the paper is twofold: on the one hand I am going to fill a gap in the literature and present a qualitative and quantitative analysis of PCs found in the *British National Corpus* (BNC) which shows which types of PCs actually occur in English. This also includes an analysis of the distribution patterns of PCs across demographic and textual features such as mode, register (e.g. academic prose, spoken conversation) and text domain. This analysis will allow us to gain further insights into the question of why PCs are built at all by speakers/writers and why they are sometimes preferred over other options. Since in my opinion this relates to their morphopragmatic character, or if you will, their expressive flavour, I will discuss Meibauer's (2007) empirical study in this context and compare his results with mine. I will further discuss his approach based on his study of German PCs

and show that it should be integrated into a new approach, which I will introduce in the last part of the paper. As mentioned above, my analysis will be based on generative principles but part company with standard generative approaches to PCs since in my opinion they can be better explained if a conceptual-semantic structure as basis is assumed.

The outline of the paper is as follows: Section two describes the PCs found in the BNC and classifies the different types. In section three the distribution of PCs is investigated in terms of the demographic and textual features mentioned above. Meibauer's (2003, 2007) interesting observations concerning the morphopragmatic character of PCs will be taken into account here. Section four introduces a new analysis in the framework of Jackendoff's model of Parallel Architecture (PA). Section five concludes.

2. A qualitative analysis of phrasal compounds in English

The collection of PCs was gathered by exploiting the BNC. I chose to use the *BNCweb* via the Lancaster interface since it allows the user to work with the *cqp* (corpus query processor) language in a convenient way. Further, it provides statistical information like the distribution of the phenomenon across categories, e.g. mode, text type, age of author etc. Before defining the corpus query, it was necessary to check how PCs are actually spelt, i.e., how the phrasal part of these compounds is represented in written speech (see section 3). I found that in English, or at least in the BNC, the phrasal part of PCs predominantly occurs in the form of a quotation. Although I also found PCs where the phrasal parts are separated by dashes, they were not as frequent as those presented in quotations. Since at present I could not find a technical solution to the problem of covering all hits for PCs with dashes, I confined the search to quotations. I defined the query as follows: "search for any string that ends in an N and is preceded by 3 to 10 words which are embraced by quotation marks". The results gained by this query search was then manually checked, non-PCs were weeded out, and true PCs were categorised according to the word categories of their non-head and head using *grep* (global regular expression print) and other unix shell tools.

In the following, I will classify the total amount of instances that can be called (true) PCs into two broad classes, non-verbal and verbal PCs. Non-verbal PCs are those that do not contain verbal material in their phrasal non-head, whereas verbal PCs are those that do. I decided to opt for these two classes since, as I will show below, there is a relevant semantic difference between PCs containing a verb and PCs that lack verbs. In section four I will provide theoretical support for this assumption.

Within these two broad classes a number of different patterns occur, I will describe these patterns in the tables below and give some examples[1]. What will become evident is that in English there is a plethora of possibilities to modify nominal heads in PCs. This does not come as a surprise and has been found in other Germanic languages as well as e.g. in German (Meibauer 2003, 2007).

In Table 1 and 2 the patterns for the non-verbal type are presented. In the left-hand column the patterns, presented by the part of speech tags of the BNC, are given, in the right-hand column relevant examples of that pattern are given. Table 1 and 2 present the patterns with the highest frequencies in the corpus, Table 3 further below presents marginal, less frequent, patterns.

The first pattern is the *Nominal-Preposition-Nominal* (Nom-Prep-Nom) one, also including other types of nominals: The first and/or the second noun can also be modified by determiners and adjectives, and the noun can also be a proper noun. Preposition stands for any preposition except of the preposition *of*, which is part of a pattern that builds its

1 All examples appear in the original graphic representation of the BNC.

own subclass (see below). As can be seen from the second column, the examples found could again be classified into several types, e.g., lexicalised (in the sense of having been stored in the mental lexicon as a unit), and non-lexicalised ones, and the former type again into titles, idioms, cliches, etc. For the moment however, I will refrain from doing so since it presupposes an interpretation which I would like to provide later on. Apart from the nominals occurring in first position in these phrasal non-heads, adverbs can occur as for example *the "away from oil" policy.* Prepositions can also be found in the rather marginal pattern *Adj-Prep-Adj* as in *a "medium to high" probability*, and the pattern *Adj-Prep-Nom* as in *these "kind to hair" curlers* (so in short in AdjPs as non-heads of PCs).

As mentioned above, another highly frequent pattern is the *Nominal-of-Nominal* one containing the preposition *of*. Here again, subtypes can be identified: the first and/or the second noun can be modified by a determiner or adjective, and either noun can occur as proper noun. Other phrases can contain the preposition of too (see the example *the "out of touch" policy* where the adverb is the head of the phrase).

Table 1: Patterns of non-verbal PCs in the BNC, part I (most frequent patterns)

Nom-Prep-Nom	a "cost per case" basis, the "Brothers in Arms World", "Highway to Hell" album, the "toe in the water" stage, "milk for Spain" appeals, the "sex in shiny packets" literature, this "gentle rain from heaven" process, the "famous for fifteen minutes" type, a "medium to high" probability
Num-Prep-Num	the "first in last out" policy, a "ten to two" position, the "two for the price of one" sales
Adv-Prep-Nom	the "away from oil" policy, the "straight for English" policy, a "once upon a time" approach, the "always on the top" option
Adj-Prep-Nom	these "kind to hair" curlers, the "tieless in Soho" fashion, the "Famous for fifteen minutes" type, a "free for life" card, the "ready for action" look
Prep-NomP	a "with a run" proviso, this "at a glance" guide, the "In such a night" speech, a "by appointment only" notice, the "around the next corner" syndrome, the "of no consequence" category
Nom-*of*-Nom	the "language of thought" thesis, the "Prince of Thieves" film, a "state of the nation" novel, a "sword of Damocles" hanging, the "representational theory of mind" sense, the "Dream of Blue Turtles" album, a "top end of the market" service, a "bunch of sixes" opportunity
Adv-*of*-Nom	the "out of touch" policy

Table 2: Patterns of non-verbal PCs in the BNC, part II (most frequent patterns)

Nom-*and/or*-Nom	a "chicken and egg" situation, the "year and a day" rule, the "Jekyll and Hyde" behaviour, the "Beowulf and brooches" approach, the "Copyright and performing Right" article, the "Race and Poll Tax" workshop, the "warts and all" school, the "peace and goodwill to all men" attitude, a "sun, sea, sex and sangria" story
Pron-*and/or*-Pron	the "them and us" syndrome, the "I and Thou" relationship
Adv-*and/or*-Adv	the "little and often" principle, the "then and now" variety
Prep-*and/or*-Prep	a "before and after" basis

The next pattern showing a high frequency of occurrence is the *Nominal-and/or-Nominal* pattern where two nominals are conjoined by *and/or*. These nominals either have the shape of nouns or proper nouns, the former of which can again be modified by determiners and/or adjectives, or occur as personal pronouns as in *the "them and us" syndrome*. Examples like *the "peace and goodwill to all men" attitude* and *a "sun, sea, sex and sangria" story* further show that postnominal modification by a PP and the listing of nouns is also found (and hence possible). Other subtypes are phrases where two adverbs (*the "then and now" variety*) or prepositions (*a "before and after" basis*) are conjoined.

In Table 3, marginal patterns are presented:

Table 3: Patterns of non-verbal PCs in the BNC (marginal patterns)

List of NPs (with and without punctuation)	the "no alcohol, no limits" campaign, a "first author, last author" citation, a "one member, one vote" system, a "no pain no gain" position
NP-*only*	a "one attempt only" scenario, the "Black women only" space, a "personal use only" agreement
Phrases showing ellipses of words (indicated by ...)	the "one--pause--two" sequences, the "if ... only" form, the "if ... then" technique
Det-Adj-Nom	the "not enough memory" line, those "Three Blind Mice" exercises, a "no first use" pledge

Here we find the pattern *Nominal-as-Nominal* as for example *the "little and often" principle*, phrases of PCs which contain a list either indicated by commas or not as in *the "no alcohol, no limits" campaign* or in *a "no pain no gain" position*, the *NP-only* pattern where a full NP is modified by a postnominal only denoting 'no one except one particular thing or person' as in *a "personal use only" agreement*, and the pattern where the phrasal part shows ellipsis indicated by "..." as in *the "if ... then" technique*. Interestingly, the pattern where a full NP consisting of a determiner, adjective and noun instantiates the phrasal part of the PC is rarely found, contrary to intuition as for example in *the "not enough memory" line*. That is why it is also found among the marginal patterns here.

Next, I am going to discuss the verbal type of PCs in the BNC. I will start out with the VPs as phrasal heads, the relevant patterns of which are given in Table 4:

Table 4: Patterns of verbal PCs in the BNC: VPs

Verb(+Obj)-*and/or*-Verb(+Obj)	a "wait and see" mentality, a "make or break" time, a "Bring and Buy" coffee, a "take it or leave it" attitude, a "Come and Try It" day, the "see and be seen" generation, the "don't drink and drive" message
Verb-Object (including phrasal verbs)	a "Guess the Weight" cake, the "Cut the Crap" LP, a "wait a minute" gesture, a "think of a number" game, a "search personal history" category, the "show the shirt" routine, the "Try your strength" machine, a "follow my leader" type
Verb-prep-NP	the "Reach for the Sky" Appeal, a "smack in the face" offer, a "wait for mail" situation
Verb-Object-Modifier	a "slow it down" mode, the "get your hands dirty" philosophy, the "recycle at all costs" policy, "could try harder" comment
V-ing-Object	that "powdering my nose" act, the "Retaining the angle" exercises, the "Rocking The Forest" mini-LP
Verb-*to*-Inf	a "return to learn" entitlement, a "need to know" basis
List of VPs (with and without commas)	the "hear no evil, see no evil" brigade, the "won't pay, can't pay" campaign, a "first come, first served" basis

The first and most frequent pattern is the coordination of two (transitive) verbs with or without object, as for example in *a "take it or leave it" attitude* or *a "wait and see" mentality*.

The pattern that is about as frequent as this one is just one VP, which is a transitive verb and its object. Examples found are *a "wait a minute" gesture* or *the "Cut the Crap" LP*, where the VP refers to the title of an LP (by the Clash). It also includes phrasal verbs, and full NP objects that are modified by determiners and/or adjectives, or prenominal possessives.

We have seen above that the most frequent pattern for the non-verbal type is the one where a nominal is followed by a preposition and a nominal. Concerning the verbal pattern, we find verbs that are modified by prepositional phrases (PP) with the function of modifier adjunct or prepositional object[2]. Some examples are *a "smack in the face" offer* and *a "wait for mail" situation*. A similar pattern are VPs with implicit or explicit object that show adverbial modification as in *a "could try harder" comment* or verbs in resultative constructions where the adjective denotes a final state resulting from the action of the verb as *the "get your hands dirty" philosophy*.

Marginal patterns are first of all VPs expressed by a verb in the progressive form and its object (e.g. *that "powdering my nose" act*), a verb followed by a *to*-infinitive (*a "need to know" basis*), a list of VPs that appear with and without commas as in *the "hear no evil, see no evil" brigade*, and participles modified by PPs as in *the "made in Japan" tag*. In the latter case, most of the time the head noun is an object like a tag or sticker which is the bearer of the utterance "made in X" and hence seems to be a quotation.

Apart from VPs and their different patterns, the majority of the verbal PCs are those that contain a full sentence, including the subject. Again, the different patterns of this type are given in a Table:

2 As concerns the type *Verb-prep-NP* type where the forms *reach*, *smack*, etc. could potentially also have the status of nouns, I decided to apply the semantic criterion of conversion and assigned verbal status to all occurrences.

Table 5: Patterns of verbal PCs in the BNC: sentences (part I)

Full subject NP + predicate	the "dog eat dog" view, the "expected price equals marginal cost" principle, , a "gee whiz, gosh, women are now making it" syndrome, the "little old lady who's lost her ticket" routine, the "straw that broke the camel's back" syndrome, the "polluter must pay" argument
Subject pronoun+predicate	the "I knew as much" smirk, the "I cannot read that" instruction, the "I want to lick you all over" variety, "I hate Kylie Minogue" T-shirts, the "we know best" philosophy, the "I'll go away and think about it" response, the "you had to be there to totally appreciate it" vein, the "I've been attacked" routine, the "Oh yeah man, we was stoned" book, the "I am not going to miss out on the fun" brigade, a "we are marching and that's that" position, the "Whoops, sorry, we forgot you" Oscars, a "once we've sold it, we forget it" attitude
Copula construction	the "small is beautiful" brigade, a "Meat Is Murder" placard, the "blood is thicker than water" explanation, this "Steffi is Great" attitude, the "less is more" effect, a "Weather hot, cricket wonderful" postcard, the "slippery, when wet" floors
Directives	the "Look At Me" category, the "Go for Gold" scheme, a "kick me please" type, a "Learn what is there and don't question it" attitude, a "NOT TO BE TAKEN AWAY" sticker, that "see me privately, or not at all" ultimatum, the "Say No To Strangers" campaign, the "do and be rewarded" training, a "don't touch me" statement, a "catch me if you can" game, the "Now Print!" mechanism
Directives with *let*	a "let it happen" attitude, the "let's get away from it all" kind, a "let's get a sunlounger and lie on the sand" sort of resort

The first two patterns of this type are declarative sentences with either a full NP subject or a subject pronoun. These in turn can either have the shape of simple sentences (subject-verb-object) where none of the units are modified as in *the "dog eat dog" view*, or they can show modification of its elements, as for example in *the "little old lady who's lost her ticket" routine*, where the "little old lady" is modified by a relative clause. We even find interjections like "gee whiz", "gosh" or "whoops" as part of the PCs as in *a "gee whiz, gosh, women are now making it" syndrome* or *"Whoops, sorry, we forgot you" Oscars*. These interjections clearly give PCs a flavour of informal spoken speech. We will come back to this point in the discussion of the sociolinguistic factors that determine the distribution of PCs in section 3.

Complex sentences showing coordination or subordination are also found, as for example in *the "I'll go away and think about it" response* or *a "once we've sold it, we forget it" attitude*. Most of the time, the form of the subject pronoun is first person singular or plural, there are some sporadic examples with second person singular (*the "you had to be there to totally appreciate it" vein*). In all of these cases, full verbs, modals and auxiliaries can occur specified for all tenses (present, past, perfect, future), predominantly in indicative mood and active voice.

Another quite prominent pattern are copula constructions where the copula is *be* and the subject complement can either be an adjective (*the "small is beautiful" brigade*, *this "Steffi is Great" attitude*) or a noun (*a "Meat is Murder" placard*). There are examples as *a "Weather hot, cricket wonderful" postcard* where the copula has been left out ("weather is hot, cricket is wonderful"). This also applies to the PC *the "slippery, when wet" floors* ("floors are slippery when wet).

The next type I would like to discuss are directives occurring in the form of an imperative sentence, i.e., it lacks a subject and generally has a verb in the base form[3]. Again, all kinds of sentences are possible, simple ones as in *the "Look At Me" category*, complex ones as in *a "Learn what is there and don't question it" attitude*, negative imperatives as in *a "don't touch me" statement* and directives with *let* as in *"let's get a sunlounger and lie on the sand" sort of resort*. Questions are also quite frequent as phrasal parts of PCs as Table 6 shows:

Table 6: Patterns of verbal PCs in the BNC: sentences (part II)

Yes/No-questions with *be* or *have*	the "Is West Belfast Working?" Conference, the "isn't it a nice day?" stage, the "Are you sure" field, the "Is this a dagger?" speech, the "are ya doin' awright?" approach, the "has he or hasn't he?" riddle, the "Have you heard the Good News" routine, the "but my, aren't they dangerous?" tag
Yes/No-questions with *modals* or *do*	that "Would you like to sit on my knee?" nonsense, the "Shall we go to the pub?" variety, the "does he take sugar?" approach, a "Did you know" section, the "Do you like housework?" question, a "gee-whiz, would-you-believe-it?" fashion
Wh-question	a "What's on" column, the "what is whisky?" debate, the "What should we do now?" variety, the "Where are you now?" Directory, the "who duz wot" bit, the "who's that dying on the runway" stuff, the "Who The Hell Did This One, Then?" party, the "how did I ever live without it" variety
Echo question, *what* in *so* expression	those "he did what?" examples, the "So what?" syndrome

Here we find on the one hand Yes/No questions with *be* (*the "isn't is a nice day" stage*), *have* (*the "Have you heard the Good News" routine*) as well as with modals (*that "Would you like to sit on my knee?" nonsense*) and *do* (*the "Do you like housework?" question*). On the other hand *wh*-questions occur where arguments and adjuncts are questioned as for example in *the "What should we do now?" variety*, *the "who duz wot" bit*, *the "Where are you now?" Directory* and *the "how did I ever live without it" variety*. Note that the question is not always graphically marked as a question (the question mark is sometimes lacking). Echo questions are a marginal type (*those "he did what?" examples*) as well as *so* expressions introduced by *what*. Although being marginal, i.e., occurring only sporadically, they nevertheless are indicative of the informal spoken flavour that has been attributed to PCs. Table 7 summarises the types of PCs found in the BNC (numbers given are absolute frequencies):

Table 7: Types of PCs found in the BNC

NP+N	694
VP+N	650
AdvP+N	32
PP+N	15
AdjP+N	6

3 Verbal phrases are categorised as directives if it was clear that the omitted subject is the 2nd person (this can be demonstrated by adding a tag question, e.g., "look at me, will you?").

Out of a total of 1397 PCs in the *BNCweb* (where the phrasal part is marked by quotation marks), the most frequent type is the nominal type with a total of 694 occurrences (i.e. tokens). The next highest frequency is found for the verbal type, including sentences, questions, etc. with a total of 650 occurrences. Furthermore, PCs with AdvPs (32 occurrences), PPs (15 occurrences) and AdjPs (6 occurrences) were found, but their frequency is much smaller than the nominal and verbal type. In the following section, it will be investigated if the frequencies of the NP+N and VP+N type correlates with features like mode, register, text type, age and gender.

3. Demographic and textual features determining the distribution of PCs and the expressive flavour of PCs

Since the BNC is a mixed general-purpose corpus of (British) English it should represent written and spoken language in a balanced way. What we find if we survey the composition of the corpus is however, that the spoken component constitutes approximately 10 per cent (10 million words) whereas the written component constitutes 90 per cent (90 million words), which means that we do not have equal proportions here. Although users can yield valuable empirical statistical data for the spoken and the written part, numbers showing overall results and their interpretation have to be treated with caution.

In the following, results concerning the distribution of PCs across demographic and textual features will be presented in a number of tables. The number of words, the number of hits and the frequency per million words (pmw) will be given and serve as a basis for the interpretation of the data. The definitions of the features to be discussed can be found in the *Reference Guide for the British National Corpus* (http://www.natcorp.ox.ac.uk/docs/URG/index.html).

Let us start with the distribution of PCs with respect to mode. Table 8 provides the results gained:

Table 8: Distribution of PCs across mode of speech

Spoken or Written:			
Category	**No, of words**	**No, of hits**	**Frequency per million words**
Written	87,903,571	1394	15.86
Spoken	10,409,858	3	0.29
total	98,313,429	1397	14.21

Out of a total of 1397 instances of PCs, 1394 were found in written speech, only 3 in spoken speech. At first sight, this seems to be a clear result, but a number of methodological problems that might have led to it, must be addressed here. First of all, as noted above, the written component constitutes 90 per cent of the whole of the BNC, thus it is not surprising that many more occurrences of PCs are found there as opposed to the few instances found in the spoken component. However, the frequency count (per millions words, pmw) confirms this result: 15.86 of PCs occur in written speech, only 0.29 occur in spoken speech. Second, at the beginning of section two the process of finding the "right" corpus query was described, and it was said that the most frequent pattern of PCs in the corpus was the one where the phrasal part is marked by quotations. Nevertheless, I also found instances where the phrasal parts were separated by dashes, e.g. "... the tired, *out-of-touch* grandfather in the Elysée" (BNC, HXU 1689). So although a large number of PCs are indeed marked by quotation marks and not by hyphens, it is methodologically extremely hard to exclude that there are other possibilities, for example no marking at all. To account

for the latter case, I cross-checked highly frequent patterns like e.g. a determiner followed by the personal pronoun referring to first person singular as in

(3) We are left with the fun loving (overgrown kids) and the *"I am not going to miss out on the fun'' brigade.* (HP61079)

but without quotation marks. The result was that I did find sporadic examples as

(4) In fact, the day before *the I Love Lucy St Patrick's Day*, she'd been told of a good job with the British Council in the Gilbert Islands. (A0L 1497)
(5) Van den Berghs' marketing director Bill Young took the UK marketer of the year trophy in recognition of his deft handling of *the I Can't Believe It's Not Butter controversy*, which showed that even the monolithic Unilever can be quick on its feet when its brands are under threat. (BNH 916)

but they are quite rare. However, to fully account for all the patterns in a corpus like the BNC one would have to find a way to retrieve all of the occurrences which definitely is a task that should be taken on in future research. What can be said at the moment is that the majority of PCs are indeed marked by quotation marks so that we get quite a good picture about the patterns that occur, albeit not a comprehensive, conclusive one.

Third, if the three occurrences from the spoken part of the BNC are investigated, we see that they are represented just like the ones in the written part:

(6) The other kind of camp on is erm *the "ring no reply'' camp on*. (KS6 508 PS6KK)
(7) Which is why it's a very good idea when you next go into your offices this afternoon, to rock the receiver, like this, because any camp ons put on the wrong extension on *the "ring no reply '' camp on* will be matured, on a first come first served basis, so that's how people get to talk to you. (KS6 513 PS6KK)

(8) It's full of *the "So what ?'' syndrome*. (KRP 986 PS62R)

The first two examples which illustrate the same type of PC - *the "ring no reply" camp on* - are from a dialogue during a telephone system training, the example of *the "So what?" syndrome* was uttered during a dialogue at the Environmental Health Officers' conference. Again, I cross-checked the most frequent patterns, but this time I retained no hits, which is probably a matter of coincidence. For spoken speech, what one would intuitively expect to occur is actually a pause after the determiner and probably also before the head noun to indicate a complex morphological unit in the flow of speech, for example:

(9) The other kind of camp is <pause> *the "ring no reply" camp on* <pause> which can be selected by the costumer.

Although pauses are encoded in the corpus, they cannot be queried since they are not tagged. What would be interesting to see is if the assumption that speakers indeed embrace PCs by pauses in the flow of speech is borne out, again I have to leave this aspect for further research.

Having discussed these problems, let us return to the frequencies found for PCs in spoken and written speech. If we apply a statistical hypothesis test to establish the significance of a comparison of the frequencies given in Table 8, we gain the result of $x^2 = 151.70368$ which means that the difference between the two different types of mode is

significant at $p < .001$[4]. To pursue this result in more detail, next, I would like to discuss the results gained concerning the distribution of PCs across derived text type[5]. Since only three occurrences of PCs were found in spoken speech, the following discussion will predominantly refer to the 1.394 hits in written speech.

The results presented in Table 9 show that the highest number of hits (427) are among the text type labelled "Other published written material" which is an exclusion of all the other categories relating to written material given in the table. It also represents the highest frequency (23,82) of PCs within these text types. The next highest number of hits (398) is found in the text type labelled "Non-academic prose and biography" with a frequency of 16.46, followed by 255 hits (16.16) in the text type "Academic prose". The next highest frequency occurs in the text type "Unpublished written material" with 20.82 pmw. As concerns the three examples in spoken speech, all three occurred in the derived text type labelled "Other spoken material".

Table 9: Distribution of PCs across derived text type

Derived text type:			
Category	**No, of words**	**No, of hits**	**Frequency per million words**
Other published written material	17,924,109	427	23.82
Unpublished written material	4,466,673	93	20.82
Academic prose	15,778,028	255	16.16
Newspapers	9,412,174	163	17.32
Non-academic prose and biography	24,178,674	398	16.46
Fiction and verse	16,143,913	58	3.59
Other spoken material	6,175,896	3	0.49
Spoken conversation	4,233,962	0	0.00
total	98,313,429	1397	14.21

A look at the distribution of PCs across text type given in Table 10 reveals that the highest number of hits (1232) occurs in "Written books and periodicals" followed by 159 hits in the category of "Written miscellaneous".

Table 10: Distribution of PCs across text type

Text type:			
Category	**No, of words**	**No, of hits**	**Frequency per million words**
Written miscellaneous	7,437,161	159	21.38
Written books and periodicals	79,187,792	1232	15.56
Written-to-be-spoken	1,278,618	3	2.35
Context-governed	6,175,896	3	0.49
Demographically sampled	4,233,962	0	0.00
total	98,313,429	1397	14.21

[4] I have used the *Corpus Frequency Test Wizard* online available at sigil.collocations.de/wizard.html.

[5] "Derived text type" refers to larger units of genre which have been defined by David Lee, one of the designers of the corpus.

Since the designers of the corpus aimed at proportions of 60 per cent from books and 30 per cent of periodicals (the latter of which include about 250 issues of newspapers), and only 10 per cent from miscellaneous sources (published, unpublished, and written-to-be spoken), the text type "Written books and periodicals" with 79,187,792 words shows a lower frequency of PCs (15.56) than the type "Written miscellaneous" (21.38) with a total of 7,437,161 words. The three examples in spoken speech occurred in the context-governed part, which is the part that consists of more formal encounters as for example meetings, lectures and the like, complementing the demographic component of the corpus which represents more informal encounters defined by age, sex, social class and geographic region. By looking at the examples, we have seen that all of them were produced in a rather formal context, which might be a surprising finding.

As an interim conclusion, we can say that PCs are a written phenomenon, they predominantly occur in books and periodicals, but also in publicity leaflets, brochures, fact sheets, school and university essays or letters. For the latter case (i.e. written miscellaneous) I provide some examples below:

(10) By day it's very much *a "let's get a sunlounger and lie on the sand" sort of resort* and the sandy beach is well equipped with everything you'll need in the way of bars, restaurants and shade, and there are several places to waterski, windsurf and parascend or try your hand on the wet bikes (AM0 1208, Club 18-30 summer holiday brochure 1990).

(11) It is a real link, not just *a "we'll be thinking of you from time to time" relationship* (CC1 205, Queen's Park Baptist Church Magazines)

(12) We are left with the fun loving (overgrown kids) and *the "I am not going to miss out on the fun" brigade* (HP6 1079, Scottish Amicable Newsletter)

To get further insights into the motivation to express something by means of a PC, next I would like to turn to the distribution of PCs across text domain. In the BNC, a general distinction has been made between imaginative and informative. Since again representativeness of the corpus was one of the main goals of the designers so that the corpus could be regarded as "a microcosm of current British English in its entirety"[6], with respect to the difference between imaginative and informative they chose to draw 25 per cent from imaginative texts and 75 per cent from informative texts.

Table 11: Distribution of PCs across text domain

Text Domain:			
Category	**No, of words**	**No, of hits**	**Frequency per million words**
Informative: Arts	6,574,853	208	31.64
Informative: Commerce and finance	7,341,009	180	24.52
Informative: Social science	14,025,538	267	19.04
Informative: Applied science	7,173,003	141	19.66
Informative: Belief and thought	3,037,532	57	18.77
Informative: Leisure	12,191,902	229	18.78
Informative: World affairs	17,244,523	209	12.12
Informative: Natural and pure sciences	3,818,803	41	10.74
Imaginative prose	16,496,408	62	3.76
total	87,903,571	1394	15.86

[6] Source of quotation: http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html.

The former category refers to fictional texts and texts perceived to be literary or creative. They are not classified according to field of subject. All other texts are labelled informative and are classified according to the eight domains listed in Table 11. The highest number of hits occurs in the domain of "Social science" (267), followed by "Leisure" (229) and "World Affairs" (209). A look at the frequency per million reveals that they actually occur most frequently in the domain of "Arts" (31.64), hence it is not the number of hits but the frequency pmw that we should take into account here. Apart from this result, the table also shows that PCs only amount to 3.76 of the total in imaginative prose. Again, I give some examples to provide the contexts in which these PCs occur.

Examples from imaginative prose:

(13) They can't fool me with *that "powdering my nose" act* (A0D 1728, from the book "A classic English crime").

(14) I decided to try *the "little old lady who's lost her ticket" routine*, but I was rumbled immediately and directed politely but firmly to the station manager's office (A0F 1448, from the book "Part of the furniture).

Examples from informative prose:

(15) Relatively weak description of *the "he was very glad of my arrival" sort* is cut, and we are left with the histrionic handclasp of Stepan Verkhovensky the actor manqué whom no reader of The Possessed will ever forget (A18 1440, from the book "Dostojewski")

(16) She also knows that the media tendency to lump together women singer-songwriters in *a "gee whiz, gosh, women are now making it" syndrome* is patronising, if not pernicious (A7S 190, from "The Guardian", electronic edition of 1989-11-08: Arts section)

(17) Most eventually got honorary Lifetime Achievement Awards — alias *the "Whoops, sorry, we forgot you" Oscars*, or even "Whoops, sorry, we didn't know you were still around", as happened to Sophia Loren in January, thirty years after she won Best Actress for Two Women (ABS 2601, from the "Esquire")

By looking at the results we have gained so far, it seems that PCs predominantly occur in periodicals and magazines and in informative text domains like "Arts" and "Commerce and finance". The question is if there is a correlation between the medium, i.e. periodicals, in which PCs are most frequently found and the text domains mentioned. Although a number of papers have dealt with PCs from a qualitative point of view (cf. Botha 1980, Lieber 1988, Lawrenz 1996), only one study of German PCs has particularly dealt with the question of why this type of compound is produced in the first place. Meibauer (2007) sees PCs as a marked phenomenon that can be explained if their expressive character is taken into account. Let us consider example (16) from above: the sentence shows a high degree of lexical density, expressed by compounds like *media tendency* and *women singer-songwriters*, and finally the PC *"gee whiz, gosh, women are now making it" syndrome*. If we pondered over an alternative for the PC, we would probably come up with an NNC like *superwoman syndrome* which would come quite close to the semantics of the PC but would nevertheless be less expressive or witty. The properties of expressivity or wittiness have been attributed to (marginal) morphology, e.g., by Zwicky & Pullum (1987:335) who have stated that a derivation like *laserteria,* meaning "a specialist retail outlet for laser equipment", could not be uttered in a business meeting without raising chuckles, and that

these words are "... whimsical coinages, carefully contrived for dubbing commercial enterprises, and carry an effect lacking in plain derivational morphology", which could also be defined as pragmatic effect (see also Bauer 1997, 2002). According to Meibauer the expressivity of PCs is caused by a conflict between two pragmatic principles belonging to the theory of generalised conversational implicatures, the I(nformativeness)-Principle and the Q(uantity)-Principle (cf. Levinson 2000). As long as a speaker produces an NNC like *superwoman syndrome* he or she adheres to the I-principle since minimal linguistic information is produced with the result that the recipient has to infer the underspecified information from the context of the utterance. As soon as a speaker however produces a PC like *"gee whiz, gosh, women are now making it" syndrome*, for him or her adhering to the Q-principle is more important since he or she has provided the strongest statement possible in that context, i.e., a statement that is more informative than one expressed by a NNC. Meibauer assumes that this is the case because PCs include sentences which contain propositions, have a set of entailments, and are bearers of illocutions. The conflict that arises between the two principles is stated by Meibauer as follows:

(18) *Expressivity in CP phrasal compounds*
Expressivity of phrasal compounds stems from a conflict between a principle that requires enrichment of a minimal and underdetermined structure in normal compounds (e.g. the I principle) and a principle that requires maximal informativity (e.g. the Q principle) and leads to the integration of a phrase into word structure.

(Meibauer, 2007, 248)

To find proof for his assumptions, Meibauer conducted a number of experiments with students who had to evaluate a PC as well as a number of alternatives concerning their understandability and wittiness. Meibauer defined understandability as the case when the effort of enrichment is too big, and wittiness as the case when incongruity on the word level occurs, which means that the integration of a phrasal meaning into a word meaning is surprising for recipients. The latter aspect implies that a PC is wittier than an NNC. For his study he used the following material:

A non-lexicalised PC in context:

(19) Während diese Zeilen entstehen, werden mehrere hundert laminierte **"Kaufe-Ihr-Auto-Kärtchen"** hinter die Hubscheibenwischer alter Mittelklasse-Mercedes geklemmt. Dabei würden deren Besitzer viel lieber an den freundlichen jungen Mann verkaufen, der sich so rührend um seine anderen alten Autos kümmert.
[Youngtimer 2/06,S.55] 'While these lines are written, several hundreds of laminated buy-your-car cards are stuck behind the lift windscreen wipers of old middle class Mercedes. Yet their owners would prefer to buy their cars to the friendly young man who is so very solicitous towards his other old cars.'

(Meibauer, 2007, 250)

Alternatives to the PC:

(20) a. Autokärtchen
car cardDIM
b. Kaufkärtchen
buyV/N cardDIM
c. Kaufe-Ihr-Auto Kärtchen
buy1.PS.SG-your-car cardDIM
d. Kärtchen "Kaufe Ihr Auto"

cardDIM “buy1.PS.SG your car"

e. Kärtchen mit der Aufschrift “Kaufe Ihr Auto"
cardDIM with the writing “buy1.PS.SG your car"

f. Kärtchen, auf denen “Kaufe Ihr Auto"
cardDIM on which “buy1.PS.SG your car" is written
(Meibauer, 2007, 250)

The task of the informants was to rate the PC and its rivals in context and in isolation in terms on understandability and wittiness on a 5 point scale (high vs. low degree of understandability/wittiness). As predicted, the PC reached the highest value for understandability and wittiness in context. Since it was not clear if it was the whole context that was perceived as being witty or the PC itself, the same experiment was conducted with the PC and its rivals in isolation. The results were similar but the values for both understandability and wittiness are lower than in task one. This means that the context indeed contributes to the overall interpretation in terms of these two features.

If we adopt Meibauer's assumptions about the pragmatic effect PCs have, and especially the role understandability and wittiness play, we might explain why they are predominantly found in text types like periodicals and magazines but also in leaflets and brochures. On the one hand, writers of these media do not have as much space as writers of books, so they have to write in a compact, yet entertaining and appealing fashion. If they produce PCs (instead of NNCs) they are likely to attain an effect of wittiness and at the same time a high degree of understandability, which is much more important than in other text types like e.g. fiction and verse. As concerns the distinction between imaginative and informative, it seems to be plausible that in the latter domain at least a high degree of understandability is more important than in the former domain (compare the examples given above). The high(er) number of hits in the subdomains "Commerce and finance" can then be explained along the same lines: since in selling something it is extremely important to attract the potential buyer, to achieve a high degree of understandability by using a PC is a good strategy. And since wittiness is also an attractive feature, PCs reach both effects automatically. Thus, Meibauer's explanation of the occurrence of PCs based on pragmatic grounds serves quite well to (partly) explain the quantitative results of the study presented. Nevertheless, this effect could also be due to the fact that different types of language situations lead to differences in the distribution of word classes: in their reference grammar, Biber et al (1999) point out that nouns are much more common in newspapers and academic texts than they are in fiction and conversation. Since PCs are nouns, such a distribution would be expected. As concerns the reason of why there is a considerable difference between written and spoken speech, we could assume that it is probably a matter of processing but this is only an ad hoc guess and of course would have to be investigated in depth.

As mentioned above, the aim of this study was to provide a qualitative, and to a limited degree, quantitative basis to explore the occurrence of PCs and their distribution across demographic and textual features. We have seen that the latter type of features do indeed, at least to some degree, determine their occurrence. Now, to complete the picture, I will present the results gained by looking at the distribution of PCs across age of author:

Table 12: Distribution of PCs across age of author

Age of Author:			
Category	**No, of words**	**No, of hits**	**Frequency per million words**
0-14	59,559	3	50.37
15-24	542,578	5	9.22
25-34	2,267,024	44	19.41
35-44	6,726,929	71	10.55
45-59	7,230,584	57	7.88
60+	5,126,298	31	6.05
total	21,952,972	211	9.61

As concerns the highest number of hits, authors between 35 and 44 produced most of the PCs in the corpus (71), followed by the age group 45-59, which is again followed by the age group 25-34 years (44). In relation to the total number of words, the age group between 0 and 14 years of age with a frequency of only three hits has the highest frequency pmw (50.37) but it is obvious that a result based on three cases only is not very reliable. If we wanted to find out whether the difference between the age categories 1 (0-14) and 3 (25-34) is really a significant difference, and not just an accident, again we would have to apply a statistical hypothesis test. The result is that the difference is not significant ($x^2 = 1.43453$). However, if the frequencies of age category 3 and 4 (71 hits) are compared, we get the result that in this case the difference found is significant at $p < .01$ ($G^2 = 9.48800$). The same applies to the comparison of frequency of age category 3 and 5 (57 hits), it is significant at $p < .001$ ($G^2 = 18.818885$).

If we finally take into account the distribution of PCs across gender of author presented in Table 13, we see that with respect to the number of hits male authors used 424 PCs whereas female authors only 111 (I will leave out the interpretation of mixed authors here with mixed being defined as more than one author of different gender), so there is a difference in frequency between 13.83 and 7.61 pmw. The statistical hypothesis test is applied again to establish the significance of a comparison of the frequencies given in the table with the result of $x^2 = 31.81616$ which means that the difference between male and female producers of PCs is significant at $p < .001$. Both results, concerning age and gender, could maybe explained with the same factor, namely that more men between the age 25 to 34 write newspaper articles because they are more often hired than women.

Table 13: Distribution of PCs across gender of author

Gender of Author:			
Category	**No, of words**	**No, of hits**	**Frequency per million words**
Male	30,662,031	424	13.83
Mixed	6,538,929	97	14.83
Female	14,588,254	111	7.61
total	51,789,214	632	12.20

Before we turn to the theoretical part of the paper in section 4, I would like to discuss the frequency breakdown of PCs in the BNC, which actually brings me back to what I have said at the beginning of section 2 as concerns the difference between lexicalised and non-lexicalised forms. It has been assumed by a number of authors (cf. e.g. Baayen 1993, Plag 2003) that words with a high frequency can be correlated with their being stored (as whole words) in the mental lexicon, i.e. with their status of being lexicalised. Words with a low frequency, on the other hand, are not likely to be stored, i.e., they do not have an entry

in the mental lexicon. According to this line of reasoning, *Hapax legomena* (items occurring only once in a corpus) are a good indicator to define which words are stored (being based on non-productive rules) and which words are not stored and hence being built productively. What we would expect to find for PCs is that the preponderance of them shows a low frequency since most of them are of the verbal type which is non-lexicalised and built on the fly. We further expect to find a number of PCs with higher frequency numbers because they include established concepts of world knowledge like titles, clichés, etc. The results presented in Table 14 confirm our expectations: first of all, none of the PCs in the BNC occurs with a high frequency, the highest frequency found is 11 for the phrasal non-head *"harm to interest"*, followed by *"Is West Belfast Working?"* with 8 occurrences, followed by the phrasal non-heads *"Reach for the Sky"* (7), *"first come, first served"* (6), *"What's On"* (5), *"What's Happening"* (5), *"response to injury"* (4) and *"law and order"*, which is a total of 50 cases and thus 3.6% of all PCs. Six different types of phrasal non-heads occur three times (total of 18 cases), two of which are given in the table. 59 different types of phrasal non-heads occur two times in the corpus (total of 118 cases), five of these are given in the table. The rest occurs only once, which means that 86.7% of all PCs in the corpus are *Hapax legomena* (1211 cases), six of which are given in Table 14. Of course, it could be due to mere chance that these PCs occur only once in the corpus, but two points speak against this conclusion: first, the BNC is a huge corpus and the likelihood that the result is due to chance is very small. Second, we could use the observation that those PCs with higher frequencies show more than one different type of nominal head as an indicator of lexicalisation. This assumption seems to be borne out, most of the PCs with the highest frequencies in the corpus (between 11 and 2 tokens) do indeed occur with different nominal heads: for example *"harm to interest"*, the phrasal non-head with the highest frequency, occurs with *theory*, *theorist*, and *principle*. There are even cases like the phrase *"law and order"* that shows five different nominal heads: *cases*, *campaign*, *language*, *attorney*, and *rhetoric*. The only two cases that speak against this assumption are the PCs with the second and third highest frequencies. If these PCs are searched for in the corpus, it becomes obvious why they show deviations from the "rule": both have been mentioned in one text several times and therefore should probably only be counted once (*the "Reach for the Sky" Appeal* in the RAFA journal (A67), and *the "Is West Belfast Working?" conference* in the EFD periodical (EFD)). This small investigation then shows that it is not the assumption that is flawed but the quantitative statistics used (which is a general problem, of course, and not particular to this study).

Table 14: Frequency breakdown of PCs in the BNC

Frequency breakdown of PCs in the BNC		
	phrasal non-heads	**N-heads**
11	"harm to interests"	theory, theorist, principle
8	"Is West Belfast Working?"	Conference
7	"Reach for the Sky"	Appeal
6	"first come, first served"	basis, principle, stands
5	"What's On''	leaflet, section
5	"What's Happening''	listings, section, pages
4	"response to injury"	hypothesis
4	"law and order"	cases, campaign, language, attorney, rhetoric
3	"wait and see"	mentality, group, attitude
3	"small is beautiful "	brigade, centre, rule
2	"workshop of the world "	tag, type
2	"warts and all"	school, closeness
2	"them and us"	syndrome, attitude
2	"take it or leave it"	attitude, basis
2	"sword of Damocles"	hanging
1	"women suffer a great deal through their husbands"	sensuality
1	"women speak more standard"	rule
1	"sell me your shares or shoot yourselves in the foot"	type
1	"sell cheap, the future looks bright"	technique
1	"gee-whiz, would-you-believe-it?"	fashion
1	"gee whiz, gosh, women are now making it"	syndrome

Taking into account the structure of PCs, a correlation between a higher frequency of occurrence and the non-verbal type, and a lower frequency of occurrence and the verbal type can be assumed. The 50 cases with the highest frequencies are predominantly of the non-verbal type including patterns like *Noun-and-Noun* or *Noun-prep-Noun*. The majority of hapaxes , however, are of the verbal type that are expressed by different types of complete sentences (see Tables 4 and 5 again). All of these hapaxes are not listed in any dictionary which supports the claim that they are a good indicator of the productivity of this type of PC. Thus, the quantitative study presented in this section has shown that the most productive type of PC is the verbal type, which is, as I have stated above, also most interesting from a theoretical perspective. It is that perspective we will turn to in the following.

4. A new approach

After having set the empirical basis in the previous section, in this section, I am going to sketch an analysis of PCs based on the model of Parallel Architecture (e.g. Jackendoff 1997, 2002, Culicover & Jackendoff 2005) for verbal PCs.

In the introduction, the motivation to provide a more satisfying analysis of these types of PCs was spelled out, and it was said that the hitherto proposed analyses have all run into problems because they are based on the following properties defining the classical generative framework: (i) they are syntactocentric, (ii) all derivational processes always start from syntax, (iii) there is a strict division between lexicon and grammar. Regardless of whether these analyses reflect a strictly lexicalist, word syntactic or mixed-model point

of view, all of them have run into serious problems explaining the "peculiar" properties of PCs because they have to allow the integration of a syntactic phrase into a morphological phrase, i.e. into word structure, although the generative system based on properties (i) to (iii) does not. From my point of view, these proposals have dealt with PCs in a one-sided way by only looking at the formal properties and neglecting the semantic aspects. I am going to discuss a different approach and will hopefully provide a more balanced analysis of this phenomenon by doing justice to both the formal and semantic properties of PCs.

So far we have seen that about half of the PCs found in the corpus are of the verbal type, and that almost all cases of that type are hapaxes being produced on the fly and thus non-lexicalised. Since this type of PC poses more problems for a formal analysis than the non-verbal type, because it really integrates a transparent syntactic structure into a word structure which cannot be said to be a whole, lexicalised, unit, in the following we will predominantly deal with this type. From the examples of the verbal type presented in Tables 4 and 5 we see that a wide variety of verbal phrases occurs, i.e., all kinds of verbs with all possible morphological inflections, with arguments and adjuncts, the elision of the verb, verbs in declarative main clauses as well as in questions, and sentences introduced by interjections as in spontaneous, authentic speech. In Table 15 the nominal heads of these PCs are classified in terms of their conceptual-semantic properties following Jackendoff 1995 and Meibauer 2003:

Table 15: Conceptual semantic classification
of the nominal head of verbal PCs (non-exhaustive)

INDIVIDUAL	variety, category, brigade, community, people, writer, attorney, team, guardian, author, searcher, teenager, theorists, Greek-Cypriot, type, group, man, wife, whiner, watchdog, starfish, sir, searcher, psychologist, prisoner, player, person, junior, guru, gang, foe, fan, expert, crew, corporation, coalition, candidate, campaigner, party
PROPERTY	image, quality, style, look, smirk, style, nature, feeling, touch, sensuality, quality, power, face
CONCEPTUAL ENTITY	idea, approach, regime, experience, theory, basis, principle
ATTITUDE	philosophy, attitude, line, position, policy, ideology, syndrome, viewpoint, vein, standpoint, statement
ACTION	series, act, routine, tactics, strategy, scheme, campaign, smokescreen, action, activity, event, exhibition, programme, conference, lunch
UTTERANCE, MEDIUM CONVEYING UTTERANCE	argument, message, gesture, story, speech, song, phrase, sound, chant, response, record, slogan, comment, report, refrain, proverb, sign, sticker, postcard, newspaper, banner, button, reader, book, letter, prospectus, chapter, section, album, LP, column, T-shirt, magazine, leaflet, guide, command, card, rhetoric, riddle, question, error, appeal, compilation, tag, box
TIME	heyday, holiday, day, session, time, era, moment, episode, phase, week(end), period, stage, situation,
THING	jacket, Oscar, machine, sweetener

In (21) a number of examples for each of the semantic concepts of the nominal heads are given:

(21) a. INDIVIDUAL

We are left with the fun loving (overgrown kids) and the *"I am not going to miss out on the fun" brigade*. (HP6 1079)

b. PROPERTY
Martinho was watching, with that *"I've got nothing to do with this" look* that he put on when he'd fucked things up good. (H9N 1983)

c. CONCEPTUAL ENTITY
Please try to avoid *the "does he take sugar?" approach*, ask the person in the chair directly "Would you like a push?" rather than ask their companion if they have one. (CHK 1298)

d. ATTITUDE
He claimed that he was sick of *this "Steffi is Great" attitude* and he accused you of showing favour towards Steffi. (A0V 485)

e. ACTION
They can't fool me with *that "powdering my nose'' act*. (A0D 1728)

f. UTTERANCE
If you are being pressurised by someone, use this tactic; it's *the "I'm just looking, thank you"* or *the "I'll go away and think about it" response* to the pushy salesperson. (CEF 1025)

g. MEDIUM CONVEYING UTTERANCE
FREED from a lift in his Harare hotel, the Bearded Wonder sends us a *"Weather hot, cricket wonderful" postcard* from Zimbabwe. (K52 2291)

h. TIME
Radio brought the main news from the outside world; nuclear tests in the Pacific, civil rights marches in America, the coronation of Queen Elizabeth and *the "never had it so good" era* in Britain. (H7E 1024)

i. THING
Most eventually got honorary Lifetime Achievement Awards -- alias *the "Whoops, sorry, we forgot you'' Oscars*, or even "Whoops, sorry, we did n't know you were still around", as happened to Sophia Loren in January, thirty years after she won Best Actress for Two Women .(ABS 2601)

We see that the non-heads of all these examples are sentences, which contain a proposition that is based on truth values. This applies to all the examples found in the different semantic categories, also to those where the copula verb be has been elided as in *a "Weather hot, cricket wonderful" postcard* (also note that the example in f. contains gapping which is generally possible in compounds, e.g. *word and sentence structure*). Interestingly, only few cases occur where the head noun denotes a concrete thing as in (21) i. (the other examples found are *a "cut and sew" jacket*, *the "Try your strength" machine*, and *the "We're not going heavy" sweetener*), the preponderance of occurrences must be classified as MEDIUM CONVEYING UTTERANCE as can be seen from the table above. I assume that this observation can be attributed to the nature of verbal PCs (see below).

If we try to apply one of the basic functions that can fill out the function *F* in NNCs as proposed by Jackendoff (2009, 2010a), the one that seems to come closest to the semantics of these PCs is BE (Y,X) meaning 'Y is (also) an X' since it is based on a predication relation. Jackendoff provides the following examples of NNCs for this relation:

(22) a. *boy king* (dvandva compound)
b. *witch doctor* (objects that are a mixture of N1 and N2)
c. *tractor-trailer* (objects composed of N1 and N2) (Jackendoff 2010a: 437f)

However, if the data in (22) are compared with verbal PCs in (21), we find that the relation is not the same. For example the PC *the "Steffi is Great" attitude* (21 d.) does not denote

"'Steffi is Great" is (also) an attitude'. The same holds for the basic function KIND (X,Y), 'an X of kind Y', denoting a relation among kinds. So whereas it is true that 'a puppy is a kind of dog' (puppy dog) this relation does not underlie the PC *"Steffi is Great" attitude*: it is not true that "'Steffi is Great" is a kind of attitude'. If these basic functions were applied to all the examples in (21) the result would be that none of these underlie the PCs. To understand why this is not possible let us compare an NNC with a PC in more general terms: the former type is based on the function $F(X_1,Y_2)$ yielding the meaning of [N_1 N_2]. In the literature, it has been extensively discussed that in isolation this relation is very hard to determine due to the semantic underspecification of compounds (see e.g. Fanselow, 1981, Meyer, 1993). Thus, a *puppy dog* could be interpreted according to the KIND relation but it could also be interpreted as e.g. "a dog who eats puppies". Although PCs are also compounds, they do not share this property with non-phrasal compounds: in our example, *the "Steffi is Great" attitude*, the relationship between the phrasal non-head and the non-phrasal head is much less underspecified, i.e., more clearly defined, namely that the utterance "Steffi is Great" expresses an attitude. This assumption is corroborated by Meibauer's experiment we have dealt with above, since he took the semantic underspecification of compounds into account and investigated PCs in isolation and in context. What he found was that his informants evaluated PCs as equally informative, understandable and witty. Thus, NNCs are much more underspecified and context-dependent than verbal PCs. What I then claim for the interpretation of verbal PCs is that this type of compound is based on the IS-A relation because they contain a proposition:

(23) [State IS-INSTANCE-OF ([x; TOKEN], [y; TYPE])] (Jackendoff 2010a:13)

Under this assumption, it is possible to provide a general account for the semantic interpretation of all PCs of this type (which also includes their form as I will show below), and to explain why NNCs differ from CPs semantically.

Based on this conceptual structure, I further claim that two types of verbal PCs must be distinguished: a) the type where the utterance refers to the concept of THING, and b) the type where the utterance refers to the concept of EVENTUALITY (cf. Varzi, 2002). Figure 1 graphically represents this classification:

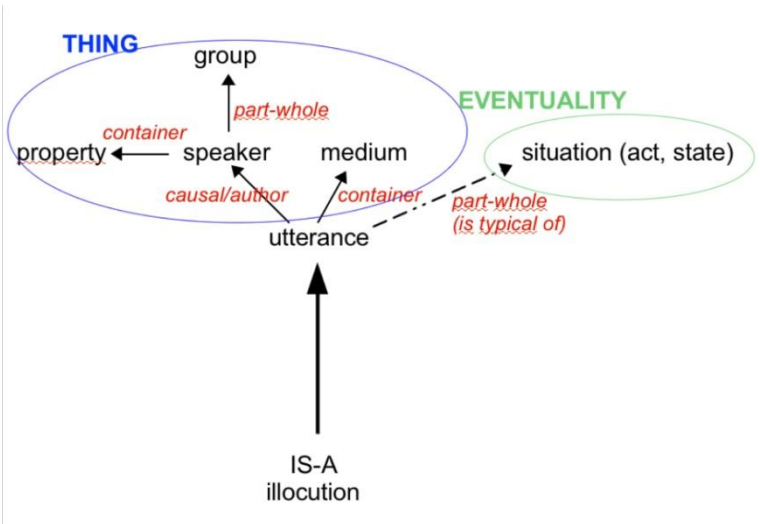


Figure 1: Conceptual semantic relations of verbal PCs

As can be seen from Figure 1 metonymic coercion, indicated by the arrows and metonymic shifts in italics, plays a crucial role in defining the properties of verbal PCs. I assume that

metonymy is a conceptual phenomenon which can be defined as proposed by Panther & Thornburg (2007: 242):

(24) a. Conceptual metonymy is a cognitive process where a source content provides access to a target content within one cognitive domain.
b. The relation between source content and target content is contingent (conceptually nonnecessary), i.e., in principle defeasible.
c. The target content is foregrounded, and the source content is back-grounded.
d. The strength of the metonymic link between source and target content may vary depending, among other things, on the conceptual distance between source and target and the salience of the metonymic source.

Type a) PCs differ from type b) PCs in that the former type refers to the cognitive domain of THINGS and almost always shows type mismatches which are resolved by metonymic coercion (for the notion of type shift or coercion see Pustejovsky, 1995, ch. 7). The only exception here are heads of the type UTTERANCE like *response*, *argument*, etc. (see Table 15) where the relation between source and target content is direct: [State IS-INSTANCE-OF ([x;I'LL GO AWAY AND THINK ABOUT IT], [y;RESPONSE])]. In all other cases of this type the relation between source and target content is indirect, for example, in the case of MEDIUM CONVEYING UTTERANCE, an example of which would be *a "Weather hot, cricket wonderful" postcard,* a type shift from UTTERANCE to MEDIUM CONVEYING UTTERANCE is assumed :

(25) "Weather hot, cricket wonderful"$_1$ postcard$_2$ = [state IS-INSTANCE-OF (UTTERANCE$^{\alpha}$ [WEATHER HOT, CRICKET WONDERFUL]$_1$, MEDIUM CONVEYING UTTERANCE$^{\alpha}$ [POSTCARD]$_2$)]

In this case, the content of the coerced function *F* is filled out by material from the proper function of *postcard*. In Jackendoff's framework, the process of cocomposition (i.e. the coercion of extra functions into the structure by filling them with either basic functions or with internal semantic structure from the first and second part of a NNC; cf. Jackendoff 2009: 120) is invoked here.

Type b) PCs including the concepts of ACTION and STATE are instances of processes of the concept of EVENT(UALITIES). Here, only one metonymic coercion occurs, *is typical of* (*part-whole*), which shows a strict association between UTTERANCE and EVENT(UALITIES). No such restriction is found for type a) PCs because they lack strict associations between speech acts and nominal heads denoting things. For illustration, let us take a closer look at the interpretation of the following two PCs: *this "powdering my nose" act* and *this "Steffi is Great" attitude*. The former can be paraphrased as 'the utterance "(I am) powdering my nose" is typical of an act', thus it includes the IS-A relation (the strict association between the utterance and the concept denoted by the nominal head is indicated by the concept of ACTION[7]):

(26) "powdering my nose"$_1$ act$_2$ = [state IS-INSTANCE-OF (ACTION$^{\alpha}$ [POWDERING MY NOSE]$_1$, ACTION$^{\alpha}$ [ACT]$_2$)]

In this case and in other cases, it is not just an act which is clearly defined by the phrasal non-head but actually an act seen as a stereotype, i.e., the phrase is used as periphrasis to refer to a salient, conventionalised piece of information in one cognitive domain, which may even lead to using it as a euphemism:

[7] In the following, all relevant concepts (ACTION, PSYCHOLOGICAL STATE, etc.) are defined according to the classification of WordNet-3.1).

(27) I'll use your bathroom. To powder my nose, as nice girls say.
(L. P. Davies *What did I do Tomorrow?* 1972, p. 72; OED online)

In the latter case, the relation can be paraphrased as 'the utterance "Steffi is Great" is typical of an attitude', again the IS-A relation holds and the strict association between the utterance and the concept of attitude (state) is indicated in the conceptual structure as follows:

(28) "Steffi is Great"$_1$ attitude$_2$ = [state IS-INSTANCE-OF (PSYCHOLOGICAL FEATURE$^{\alpha}$ [STEFFI IS GREAT]$_1$, PSYCHOLOGICAL FEATURE$^{\alpha}$ [ATTITUDE]$_2$)]

Whereas for type a) PCs a number of metonymic shifts are possible, for type b) PCs only one shift is. Note however, that generally these types of shifts are constrained, since a shift from source to target content is dependent on its cognitive domain. In the case of *the "I am not going to miss out on the fun" brigade,* the utterance "I am not going to miss out on the fun" conveys an attitude which is attributed to a group of people (brigade is used here to denote a social group). Here we find a more indirect metonymic shift than in the other cases above: first, a shift from the utterance to an individual (*causal/author*), and second, a shift from individual to group (of individuals) via the *part/whole* relation. Since an attitude is conveyed, the structure of the PC thus resembles the one proposed for the PC *"Steffi is Great" attitude*.

(29) "I am not going to miss out on the fun"$_1$ brigade$_2$ = [state IS-INSTANCE-OF (PSYCHOLOGICAL FEATURE$^{\alpha}$ [I AM NOT GOING TO MISS OUT ON THE FUN]$_1$, PSYCHOLOGICAL FEATURE$^{\alpha}$ [BRIGADE]$_2$)]

Generally, it could be concluded that the difference between the interpretation of NNCs and PCs lies in the fact in the former case, the semantic relation between the non-head and head is much more underspecified because it is based on a one-to-many relation. PCs on the other hand do not share this property because they are based on the IS-A relation and rather specified instances of metonymic shifts or coercions, which have also been called 'rules of construal' (Nunberg 1979).

We have further seen that metonymic coercion plays a crucial role in explaining the nature of verbal PCs. Based on the observations made above, it could be assumed that the expressive flavour of PCs discussed in section 3 results from the strength of the metonymic link between source and target content. To put it simpler, the more indirect a metonymic link is between the source and target content, the wittier a PC is. Cases where no metonymic coercion occurs would therefore be perceived as being less witty. This seems to be borne out for type a) PCs:

(31) a. If you are being pressurised by someone, use this tactic; it's *the "I'm just looking, thank you"* or *the "I'll go away and think about it" response* to the pushy salesperson. (CEF 1025)
=> no metonymic coercion, less witty

b. Most eventually got honorary Lifetime Achievement Awards -- alias *the "Whoops, sorry, we forgot you'' Oscars*, or even "Whoops, sorry, we did n't know you were still around", as happened to Sophia Loren in January, thirty years after she won Best Actress for Two Women .(ABS 2601)
=> metonymic coercion, more witty

For type b) PCs where a strict association between source and target content was assumed this explanation does not hold. Rather, I would assume that in this case the observation

that an act is seen as a stereotype, is relevant (apart from the fact that the unexpected occurrence of a sentence within a word must also have an effect). According to Levinson, lexical items have the potential to implicate stereotypical default readings, so "What is expressed simply is stereotypically exemplified" (2000:37). It is the interplay of these rules that can account for the properties of PCs.

In section 3, Meibauer's definition of the expressivity of PCs was discussed. Meibauer claimed that enrichment and informativity (based on a conflict between Levinson's I and Q principle) on the structural level are critical for explaining this property. But perhaps it is not the structural level but the conceptual level which plays the decisive role. Papafragou (1996) defines two communicative reasons for using metonymies: on the one hand metonymies cause extra processing effort which is "levelled out" by a gain in contextual effects (additional implicatures). On the other hand, the processing effort may be smaller than that for a literal expression of the metonymic sense. If we applied the latter communicative reason to Meibauer's assumptions and to the production of PCs, we could say that producing this type of compound leads to enrichment via metonymic coercion, i.e, to additional contextual effects, which is not possible with NNCs. What we automatically gain is a maximum of informativeness, so from this point of view, it is not a conflict that arises. Although the cognitive effort is greater, it is still the most economical way to get to enriched conceptual information, thus a PC will be preferred over an NNC because it is wittier (distance between target and source content) and more understandable (more enriched, more transparent).

5. Conclusion

This paper has dealt with phrasal compounds from an empirical and a theoretical point of view. The motivation to study this phenomenon in depth was on the one hand the lack of a qualitative and quantitative empirical study and on the other hand a satisfying analysis. In sections 2 and 3 the empirical corpus study based on the BNC was discussed, and the PCs found were classified into two main categories, verbal and non-verbal PCs. The verbal type occurred about as frequent as the non-verbal type, and it was shown that although a number of patterns do determine the occurrence of phrasal non-heads (e.g. *Nom-prep-Nom*), this mainly applies to the non-verbal type. For the verbal type, it was shown that all kinds of sentences are allowed, simple as well as complex. It was assumed that the latter type is the more interesting type since (i) it is the more productive type, (ii) the sentential non-head includes a number of entailments, (iii) a proposition based on truth values is always included, (iv) as an utterance the phrasal non-head includes illocutions. Further, apart from the qualitative analysis in section 3 a quantitative analysis in terms of the distribution of PCs across textual and demographic features was provided. The main results were that PCs are a phenomenon of written speech (significant correlation), more precisely informative prose, and that they predominantly occur in newspapers and periodicals. It was also shown that the distribution across age categories and differences between male and female producers of PCs were significant. It was said that one explanation for these results could be explained with the same factor, namely that more men between the age 25 to 34 write newspaper articles because they are more often hired than women.

Concerning the question of why PCs are produced at all, Meibauer's interesting observations regarding PCs in German were discussed and applied to the English data presented here. It was concluded that his definitions of understandability and wittiness could explain the findings, and in more general terms, the morphopragmatic character of PCs.

Based on this empirical study, in section 4 a sketch of an analysis along the lines of Jackendoff's Parallel Architecture was provided. The fact that verbal PCs contain

propositions invited a conceptual-semantic analysis based on the IS-A predication relation which was applied to a number of PCs to show that it generally holds, and that metonymic coercion are needed to account for the facts. Thus, I hope to have shown that an analysis based on Conceptual Semantics is quite promising and might be able to account better for the phenomenon than any other analysis that has been proposed so far in syntactocentric, derivational models of generative grammar.

Bibliography

Ackema, P. and Neeleman, A. 2004. *Beyond Morphology*. Oxford: Oxford University Press.

Baayen, H. 1993. "On frequency, transparency and productivity", In G. Booij and J. van Marle *Yearbook of Morphology 1993*, 181-208. Dordrecht: Kluwer.

Bauer, L. 1997. "Evaluative morphology: in search of universals". *Studies in Language* (21): 533–75.

Bauer, L. 2002. *English Word-Formation*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman Addison Wesley.

Borer, H. 1988. "On the morphological parallelism between compounds and constructs". In G. Booij and J. van Marle (eds), *Yearbook of Morphology* 45–65. Dordrecht: Kluwer.

Botha, R. 1981. "A base rule theory of Afrikaans synthetic compounds". In M. Moortgat, H. van der Hulst and T. Hoekstra (eds), *The scope of lexical rules,* 1–77. Dordrecht: Foris.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambrige Mass.: MIT.

Chomsky, N. 1981. *Lectures on Government and Binding*. Studies in generative grammar 9. Dordrecht: Foris.

Culicover, P. and Jackendoff, R. 2005. *Simpler Syntax*. Oxford: Oxford University Press.

Fanselow, G. 1981. *Zur Syntax und Semantik der Nominalkomposition: ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung im Deutschen*. Linguistische Arbeiten 107. Tübingen: Niemeyer.

Gallmann, P. 1990. *Kategoriell komplexe Wortformen*. Tübingen: Niemeyer.

Jackendoff, R. 1975. "Morphological and semantic regularities in the lexicon". *Language* (51): 639–71.

Hoffmann, S., S. Evert, N. Smith, D. Lee, Y. Berglund Prytz. 2008. *Corpus Linguistics with BNCweb - a Practical Guide*. Frankfurt: Peter Lang.

Jackendoff, R. 1983. Semantics and Cognition. Cambrige Mass.: MIT Press.

Jackendoff, R. 1990. Semantic Structures. Cambrige Mass.: MIT Press.

Jackendoff, R. 1995. "The Boundaries of the Lexicon". In M. Everaert, E.-J. van der Linden, A. Schenk and R. Schreuder (eds), *Idioms. Structural and Psychological Perspectives*, 133–65. Hillsdale, N.J.: Erlbaum.

Jackendoff, R. 1997. The Architecture of the Language Faculty. Cambrige Mass.: MIT Press.

Jackendoff, R. 2002. Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford: Oxford University Press.

Jackendoff, R. 2007. "A parallel architecture perspective on language processing". Brain research (1146): 2–22.

Jackendoff, R. 2008. "Alternative minimalist visions of language". In Chicago Linguistic Society 41: The Panels, R. Edwards, P. Midtlyng, K. Stensrud and C. Sprague (eds), 189–226. Chicago: Chicago Linguistic Society.

Jackendoff, R. 2009. "Compounding in the Parallel Architecture and Conceptual Semantics". In R. Lieber and P. Štekauer (eds), *The Oxford Handbook of Compounding*, 105–129. Oxford: Oxford University Press.

Jackendoff, R. 2010a. Meaning and the Lexicon. Oxford: Oxford University Press.

Jackendoff, R. 2010b. "The parallel architecture and its place in cognitive science". In B. Heine and H. Narrog (eds), *The Oxford handbook of linguistic analysis*, chapter 23, 583–605. Oxford University Press.

Lapointe, S. G. 1980. *A theory of grammatical agreement*. Ph.D. thesis, University of Massachusetts, Amherst.

Lawrenz, B. 1996. "Der Zwischen-den-Mahlzeiten-Imbiß un der Herren-der-Welt-Größenwahn: Aspekte der Struktur und Bildungsweise von Phrasenkomposita im Deutschen". In: *Zeitschrift für Germanistische Linguistik* 24, 1-15.

Levinson, S. C. 2000. *Presumptive Meanings*. The Theory of Generalized Conversational Implicature. Cambridge, Mass.: MIT Press.

Lieber, R. 1988. "Phrasal compounds and the morphology-syntax interface". Chicago Linguistic Society II Parasession on agreement in grammatical the-ory (24): 202–222.

Lieber, R. 1992. Deconstructing Morphology. Chicago: The University of Chicago Press.

Meibauer, J. 2003. "Phrasenkomposita zwischen Wortsyntax und Lexikon". *Zeitschrift für Sprachwissenschaft* 22 (2): 153–188.

Meibauer, J. 2007. "How marginal are phrasal compounds? Generalized insertion, expressivity, and I/Q-interaction". *Morphology* (17): 233–259.

Meyer, R. 1993. *Compound Comprehension in isolation and context*. Linguistische Arbeiten 299.

Nunberg, G. 1979. "The nonuniqueness of semantic solutions: Polysemy". *Linguistics and Philosophy* 3, 143-184.

Panther, K.-U. and Thornburg, L. 2007. "Metonymy". In D. Geeraerts and H. Cuyckens (eds), The Oxford Handbook of Cognitive Linguistics, 236-264. Oxford: Oxford University Press.

Papafragou, Anna. 1996. "Figurative language and the semantics-pragmatics distinction." *Language and Literature* 5: 179–93.

Plag, I. 2003. *Word-Formation in English*. Cambridge: Cambridge University Press.

Pustejovsky, J. 1995. The Generative Lexicon. A Theory of Computational Lexical Semantics. Cambrige Mass.: MIT Press.

Štekauer, P. and Lieber, R., (eds). 2009. *The Oxford Handbook of Compounding*. Oxford: Oxford University Press.

Trips, C. 2009. Lexical Semantics and diachronic morphology: the development of the derivational suffixes -hood, -dom and -ship in the history of English. Tübingen: Niemeyer.

Trips, C. Submitted. "The relevance of phrasal compounds for the architecture of grammar".

Varzi, A. C. 2002. "Events, Truth and Indeterminacy". *The Dialogue* 2: 241–264.

Wiese, R. 1996. "Phrasal compounds and the theory of word syntax". *Linguistic Inquiry* (27): 183–193. Zwicky, A. and Pullum, G. 1987. "Plain morphology and expressive morphology". In J. Aske et al. (eds), *Proceedings of the thirteenth annual Meeting of the Berkeley linguistic society*, 330–340. Berkeley: Berkeley Linguistic Society.