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) $[[\textit{high}N]_{AP}[N]_{N'}]_{N'}$
(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.