# 2 *Compounds and Words*

## 2.1 Introduction

Perhaps the basic assumption underlying compounds is that they are words – they even are called ‘compound words’ on some occasions and are often defined as being words whose elements are words. There are two major problems with this insight. The first is that sequences of words are usually syntactic structures, not morphological ones, so we need to justify this conclusion, not merely accept it as an assumption. The second is that we have no generally accepted definition of a word. Not only is there no definition of word, we have a number of distinct elements which are viewed as being words of different kinds, and definitions of words of any of these kinds prove to be difficult or controversial.

In this chapter the notion of word will be considered, and criteria which might seem to distinguish single words (which may be morphological structures) from sequences of words (which are syntactic structures) will be analysed. It should be borne in mind that the canonical word, at least in English, is morphologically simple: *car*, for instance. In other languages the canonical word may involve inflectional morphology as of necessity: it is impossible to cite the Latin word *carrus* ‘cart’ without providing a case, gender, number marker for it. While inflection will prove important in some respects later, we are not primarily concerned with inflection here except as a guide to something else. And what we are primarily concerned with is words which, if we ignore inflectional marking, are not morphologically simple but are constructions.

## 2.2 Words, Words, Words

The distinction between a lexeme such as ARRIVE and a word-form such as *arrived* is well-established in the literature, as is the Lyonsian (Lyons 1968) notation employed here. The only point to underscore is that *arrive* is also one of the word-forms which can realised the lexeme ARRIVE: the full set being, on most accounts, *arrive, arrives, arrived, arriving*. Where the distinction between lexeme and word-form is not crucial, the typographical distinction will not be maintained, and italics will be used.

The notion of word-form is partly (perhaps largely) derived from the notion of inflection. The word-form represents not only the relevant lexeme but also all the relevant inflectional categories. For Lyons this implies that items that do not mark inflectional categories cannot be lexemes. Following Bauer et al. (2013: 8) this view is not accepted here: there can be a lexeme FROM or a lexeme LIKEWISE, but these lexemes are only ever represented by a single word-form.

Another way of viewing the word-form is that it is the realisation or representation of the morphosyntactic word (for the terminology see Bauer et al. 2013: 10), that is, including all relevant inflectional material. On this basis, the word-forms *arrive* and *arrived* in the list given just above are each interpretable as more than one morphosyntactic word: *arrive* could be the infinitive, the imperative or a non-3SG present tense; *arrived* could be a past tense or a past participle. Again, the morphosyntactic word will not be the primary focus in what follows. What is important here is that the word-form may not always be a simple concept.

When we are dealing with compounds, the general assumption is that we are dealing with items which are words in the sense that they are lexemes (see Bauer 2004a, sv *compound*). This does not imply that the citation form of the constituent "words" must be included in the form of the compound. While an English compound like *carpark* contains two elements each of which could be used as the citation form of a lexeme, the Latin form *agri·col·a* 'field·cultivate·er = farmer' does not contain the citation form of *ager* 'field' or *colere* 'to cultivate' (Oniga 2014: 167). It is for this reason that compounds are sometimes said to be made up of elements which are stems or roots: not because the implications are fundamentally different, but because stems and roots are required minima for lexemes, and some representation of a lexeme is required in each element of a compound. It would be true to say that *carpark* contains two roots or two stems as well as to say that it contains two lexemes. In some languages, calling the realisation of the lexeme in a compound a root or a stem may be more precise than saying that it is a lexeme: a stem is a distinct form of a lexeme to which affixes are added, the root is the smallest part of the word that realizes the lexeme. In English, the distinction rarely matters.

## 2.3 Orthographic Words

There is a general assumption that orthography reflects native-speaker intuition about wordhood, and that orthographic unity is a sign of being a word.

Correspondingly, lack of orthographic unity is taken to indicate that a construction is syntactic. There is plenty of evidence to support such a viewpoint, but also evidence against it. In any case, the evidence from unity and the evidence from lack of unity are not equivalent.

A word like *altogether* began life as a phrasal expression *all together* (*OED*); the two may now be distinguished from each other. The phrase *in so far as* is thus spelt in the *OED*, though it is now frequently found as <insofar as>. The prepositions *into, onto* also started out as two orthographic words (*OED*). In all such cases it seems that frequent co-occurrence leads to univerbation. It is tempting to include things like *before-tax* (as in *before-tax profits*) in the same category, but the hyphen here may simply indicate that *before-tax* is to be read as a constituent acting as an attributive (in which use the hyphen is standardly employed: see Bauer et al. 2013: 56), rather than as an indication of wordhood. Be that as it may, there is evidence of a diachronic shift from multiple orthographic words to single orthographic word. Similar cases of vacillation in usage between one and two orthographic words can be found in Danish (Bauer 2000: 253, Dansk Sprognævn 2015), under similar circumstances.

The orthographic question can also affect items which contain obligatorily bound morphs. There is some evidence that prefixes are sometimes viewed as independent orthographic words. COCA (Davies 2008) provides examples such as those in (1).

(1) hyper activity, hyper efficient, maxi systems, mega success, micro engines, neo Nazi, pre diabetes, pre pregnancy BMI, super glue, super loud

What we may be seeing here is change in the system: *insofar* is in the process of becoming a word in English, as is *hyper*, although one shows a tendency towards larger words, the other a tendency towards smaller words. Such periods of instability in change can be long-lasting, though, so that the system as a whole is never stable and never quite allows a match between orthographic criteria and other criteria for wordhood.

We might also interpret the use of hyphens with some very productive suffixes as evidence of a similar feature. COCA provides examples such as those in (2).

(2) an almost **Christmas-y** lumination and loveliness
The very **design-y** backgrounds
The honey brings out the **perfume-y** herbs in the mix
**vintage-y** leopard-print
‘What time is it?’ # ‘**Dawn-ish**.’ # ‘Dawn-ish?’ # ‘Not quite dawn, no longer night.’

> Ridley Scott's **noir-ish** *Blade Runner*
> the **Warhol-ish** gig Jon Voight walked into in *Midnight Cowboy*.

In the case of -*ish*, the suffix can also be used in isolation: the example in (3) is again from COCA.

(3) It's 68, 68-ish. GIFFORD: Yeah. **Ish**. It – you know what, it's warmed up.

It is difficult to know how to interpret such examples, but one possibility is that orthographic wordhood is not quite as much a yes-no question as we tend to suppose: there are degrees of wordhood, and some prefixes, for example, are very near the boundary of being independent words.

Given that prefixes and suffixes may not be securely non-words, it is very difficult to know how to interpret the orthography of compounds. It is well-known that the orthography of N + N sequences is variable in English: we can write *coffeepot, coffee-pot* or *coffee pot, wordformation, word-formation* or *word formation*. Nevertheless, the orthography is not infinitely flexible: <universityadministration> or <rail way> would both be odd. But hyphenation, in particular, seems to be so subject to fashion, house-style and individual preference (not to mention such factors as line-ends and typographical exigencies) that it is hard to draw firm conclusions from it. That is not to say that hyphenation is entirely free, either. Where a hyphen connects two elements in a three-element constituent, it is virtually invariably the case that the hyphenated items are a constituent in the longer constituent. These comments are, of course, restricted to English. Hyphenation in other languages may behave differently. But wherever it is used, it has to be asked whether its function is joining two elements (cf. the German name of the hyphen, *Binde·strich* 'binding dash') or keeping them separate, as it might be argued to do in <co-ordination>).

If we look away from hyphenation, this dual nature of the representation is not relevant. But the fact we regularly write <schoolboy> but never write *<universitystudent>, or that <textbook> is found but not *<librarybook>, tells us that some of the spelling conventions derive from word length (possibly associated with etymology or morphology) rather than with grammatical or semantic factors. That is, writing some construction as two words may not indicate that it is – in any but an orthographic sense – two words, though writing things together does seem to imply that they are seen in some sense as a single word.

Precisely what that sense is may require some consideration. It is, or was until recently, normal practice when reading a web address such as

Cambridge University Press
978-1-108-41603-0 — Compounds and Compounding
Laurie Bauer
Excerpt
More Information

www.airnewzealand.co.nz out loud to say ‘Air New Zealand, all one word’. The “one word” in such cases is purely orthographic and has no other implications. To the extent that this is true, it may suggest that orthography is the prime determinant of what is considered to be a word in English – something which brings an unhealthy circularity into the question of definitions, since the normal expectation among linguists would be that orthography should reflect linguistic intuitions, orthography always being dependent upon speech.

In short, the relationship between orthography and word-status is rather more fraught than is generally recognised. Clearly, if we look across a language like English as a whole, there is a tendency for orthographic unity to coincide with wordhood as defined by other means; but there are sufficient mis-matches for it to be difficult to take orthography as criterial in defining a word in any other sense. If we look beyond Standard Average European, the mismatches, while different, become even more threatening to the idea of an equivalence between the orthographic word and the word in other senses.

## 2.4 Phonological Words

The whole reason that we have a notion of phonological word is that phonological words do not necessarily match other kinds of words. There is an added problem here, namely that there does not seem to be agreement on what constitutes a phonological word. In this section we will consider some potential segmental criteria and some potential prosodic criteria whose function, in general terms, is to delimit the word rather than to define a particular phonological sequence as being a word.

### *2.4.1 Segmental Criteria*

We can divide segmental criteria for wordhood into two major kinds: those that show where a word begins or ends, and those that show whether a break between significant elements is or is not a word-break (the alternative being that it is a break between morphemes or formatives which are word-internal). Suomi (1985) calls these positive and negative word boundary signals, respectively.

As an example of segmental boundary marking, consider terminal devoicing (*Auslautverhärtung*) in German. No monosyllabic stem acting as a word-form in German can end in a voiced obstruent: *Rad* ‘bicycle’ is homophonous with *Rat* ‘council’, even though the genitive forms, *Rades* /ʁaːdəs/ and *Rates* /ʁaːtəs/ respectively, are not homophonous. However, the devoicing also affects obstruents with following voiceless obstruents within the word-form (e.g. *gib·st* 'you$_{\text{SG}}$ give' is /gɪpst/) or with a following morpheme boundary within the word-form (*lieb·lich* 'adorable, lovely' is /liːpliç/) (Hall 1992: 28–29). Wiese (1996: 200–205) shows that this is better described as a syllable-structure constraint rather than a matter of morphological structure at all, but finds some instances which are awkward for that analysis. So, whatever this process is marking, it is not – or is not only – the word. The roughly comparable phenomenon in Russian (Cubberley 2002: 73–77), while it differs in details from the German, also fails to mark the word uniquely.

As an example of a segmental phenomenon showing that two elements belong together, consider Japanese rendaku. Rendaku is the voicing of an initial voiceless obstruent when that obstruent is the second element of a compound. Some examples are given in (4) from Itô and Mester (2005: 40–41).

(4) *Japanese rendaku*

| *1st element* | *gloss* | *2nd element* | *gloss* | *Compound* | *gloss* |
|---|---|---|---|---|---|
| iro | 'colour' | kami | 'paper' | irogami | 'coloured paper' |
| asa | 'morning' | kiri | 'mist' | asagiri | 'morning mist' |
| maki | 'rolled' | sushi | 'sushi' | makizushi | 'rolled sushi' |
| hana | 'nose' | chi | 'blood' | hanaji | 'nosebleed' |
| ike | 'arrange' | hana | 'flower' | ikebana | 'ikebana' |

Note that although the voicing rule is mostly regular, the voiced equivalent of /h/ is /b/, as in *ikebana*. Rendaku does not apply (a) if there is a voiced obstruent in the second element (*kami* 'divine' + *kaze* 'wind' > *kamikaze* 'kamikaze' because of the /z/), or (b) if the compound is a coordinative compound (*me* 'eye' + *hana* 'nose' > *mehana* 'eyes and nose' – see Section 4.6 for more detail on coordinative compounds), or (c) when non-native elements are involved (Itô & Mester 2005: 42–43). Despite these generalisations, and the productivity of rendaku, many authorities see the process as in principle unpredictable.

Here we see that, even if we look away from the cases which are totally unpredictable, rendaku marks only a subset of compounds, because only a subset have phonologically appropriate form to allow the marking.

Many languages have a feature of vowel harmony. Vowel harmony is the agreement of vowels for some phonetic/phonological feature within (typically) the word. Thus vowels within the word may agree as to frontness/backness, as to tenseness, as to roundedness and so on. Although there are languages in which compound words show such agreement (Chukchee is one such;

see Bogoras 1922: 892–894), in most (including Finnish, Turkish and Turkana) the elements of a compound but not the compound as a whole are the domain of vowel harmony.

### *2.4.2 Suprasegmental Criteria*

In many languages, stress is cited as a feature which defines the word. This is perhaps particularly so in those languages in which stress plays a demarcative function, and stress falls regularly on the first (Czech, Finnish), last (Hebrew, Turkish) or penultimate (Polish, Swahili) syllable of the word. Even in such languages, the phonological word defined by stress tends to differ from the word in other senses because unstressed function words and clitics have to be included within the stress-defined unit, but not within the word viewed as related to the lexeme.

In some tone languages, compounds may be marked tonally, either by patterns of tone sandhi or by other means. For example, Chao (1968, cited in Dixon & Aikhenvald 2002: 11) notes that in Wu dialects of Chinese, tone sandhi within a compound is different from tone sandhi between non-compounded words. In Bambara N-N compounds only the first syllable may be either high or low, all subsequent syllables being high tone, independent of the tone of the element in isolation (Creissels 2004), which is also the pattern shown in non-complex words.

### *2.4.3 Discussion*

What all these determinants of phonological words have in common is that they fail to match any lexico-grammatical sense in which a compound might be “a word”. In some cases it is difficult to see how they could: if terminal devoicing systematically marked the ending of every word, every word would have to end in a voiced obstruent. Under such a scenario, many final obstruents would appear devoiced so often that it would not be obvious that they had any voiced correlate (consider the history of the German directional adverb *weg* ‘away’ which is only ever pronounced with a final /k/, although it has an orthographic – and etymological – link to the noun *Weg* ‘way’ in which alternation between /g/ and /k/ can be heard).

It may be that there are cases where the phonological word and the word in other senses match much better. If that is the case, however, there is no need for both constructs. Discussion of a phonological word separate from a word implies a mismatch or a non-exhaustive analysis.

The discussion in this section has been oversimplified in that it has assumed that one phonetic or phonological factor might act as a defining criterion for the word. It is much more likely that some cluster of phonetic or phonological criteria would have this effect. In a language like English, such factors might include degree of aspiration, stress, vowel and consonant length, various phonotactic constraints (e.g. that /zf/ cannot occur within a monomorphemic word or that /ŋ/ cannot appear at the beginning of a word). It is not clear that such factors as a set can actually define the set of words in a way that matches other notions of word any better than the individual factors can. Suomi (1985) argues that even with vowel harmony and first-syllable stress, various phonological criteria do not uniquely identify non-phonological words in Finnish.

## 2.5 Listedness and Wordhood

Scrabble players will be very familiar with the challenge that something is not a "real" word, but something that has just been made up. "Real" words, in common parlance, are listed in dictionaries. Words like *patriation* (which is not listed in the *OED* with this meaning) in (5) are correspondingly not "real" words.

(5) '"Selective patriation"? What in God's name does that mean?' the prime minister looked from one to the other until his gaze settled on Derek Farmer. 'Is it even a word?'
'If it wasn't before, it is now, Geoff,' said Farmer. (McNab, Andy. 2015. *State of emergency*. London: Bantam, p. 68)

To phrase this differently, there is an expectation that words should be listed, while there is no expectation that phrases or clauses should be listed. The concept of listedness is not a simple one: listedness for the individual speaker may not be the same as listedness for the community; it is not clear to what extent listedness is a psychological matter (in which case *comes* may well be listed as a function of its high frequency) and to what extent it is a derived from unpredictability (in which case *cat* is listed, because its form cannot be predicted from its meaning, but *comes* is not listed, because its form can be predicted from the form of *come* and the meaning).

Di Sciullo and Williams (1987) spend some time in showing that there is no necessary link between wordhood and listedness. Not only are there words which are not listed (as illustrated by (5)), there are non-words which are listed. On the one hand, morphemes are presumably listed, especially if they can be used productively, since the meaning of -*ise* is as unpredictable as the meaning of *cat*. But there are also items which are larger than words which are listed. These include the types listed in (6).