Entering meaning

1 Lexicology to lexicography

Lexicology is used for the overall study of a language's vocabulary, usually including its history. *Lexical studies* has about the same area of application as, and is more traditional in the Anglo-Saxon world than, lexicology. It is distinguished from another linguistic pursuit, *lexicography*, the art and science of dictionary making, carried out by *lexicographers*. Accordingly, the practical business of lexicography may be seen as *applied lexicology*. The term *lexicologist* is even less widely used in the English-speaking world than *lexicology:* people investigating the vocabulary at large are usually considered (some kind of) *semanticists*.

This attempt at a short and simple definition (adapted from Crystal 2008) already shows that the boundaries of lexicology and lexicography, and their place within – or by the side of – (theoretical) linguistics can be approached in different ways.

The following text will give a taste of what goes on in these areas of language study, essentially without taking sides in matters, at times battles, of definition. When a term is used, it will, however, be adequately defined, even if somewhat informally. Wherever important differences between usages are found, that too will be pointed out, but this will be kept to the minimum. Where parallel terms are in currency, they will all be pointed out, since familiarity with (a sometimes too varied) terminology is necessary for further study. Of the theoretical underpinnings, however, only as much will be introduced as is needed for an understanding of the phenomena of the lexicon on the one hand, and the work of lexicographers on the other.

A lot of what will be said in the first, LEXICOLOGY part, will not surface in the LEXICOGRAPHY part[1] because it has no direct relevance for dictionary making or dictionaries in general. By contrast, because lexicographic decisions are not solely motivated by aspects of lexicology, the second, LEXICOGRAPHY part is not just informed by the LEXICOLOGY part. Note, however, that the insights of lexicology may be central for the lexicon and linguistic studies even if they do not find ready application in "applied lexicology". *Etymology*, for instance, while generally considered to belong within lexicology, has limited application in general dictionaries (and none at all in bilingual ones), and partly for this reason little will be said about it. Also, wherever *interfaces* between the lexicon and other compartments of language are mentioned, little space will be devoted to matters of phonology, which, it is hoped, is adequately covered in other courses.

You will have noticed that the title *Entering meaning* is a *pun*. The Cambridge Advanced Learner's Dictionary[2] defines **pun** thus: 'a humorous use of a word or phrase which has several meanings or which sounds like another word'. As a trained linguist with certainly more background than the general reader, you will surely be dissatisfied with this definition. If so, you are encouraged to check its definitions in more serious sources. But if you appreciate the pun (you don't have to laugh your head off, mind you, just see how it works), then you will realize that it is the meaning of **enter** that causes the *lexical ambiguity* underlying the pun.[3] So let's enter the meaning of **enter**. It may either be used in meaning ① 'go in' (as in **enter a room**), or ②

[1] To be added to this existing LEXICOLOGY part at a later.

[2] Dictionaries cited from or just mentioned will be referred to with an abbreviation only; this Cambridge dictionary, e.g., is CALD (2008). These are listed in the Bibliography.

[3] You may feel that not just the meaning of a word but also some aspect of grammar plays a role here. We can ignore that right now.

'record; register' (as in **enter a word in a dictionary**).[4] If you go with ①, this course will go into – look at – meaning, i.e. do lexicology (which comprises lexical semantics, remember). The other meaning, ② suggests that we will examine how this meaning gets treated in dictionaries.

1.1 Meaning and sense

Notice that so far only the first of meaning and sense, *meaning* has been used. These terms may be used synonymously, while in some authors they are differentiated: in these latter frameworks, *sense* signifies the central, core meaning of a lexical item, sometimes called the *cognitive meaning*. These authors, accordingly, will speak of other kinds of meaning, such as *connotative, stylistic* etc. Note that the modifier **cognitive** *collocates*, i.e. goes together or combines with **meaning**, but not with **sense**: **?cognitive sense** is an impossible *collocation*/combination. Similarly, **core** does not collocate with **sense**, for the same reason: if **sense** = **core meaning**, then that would be superfluous, or *redundant*. Many sources indeed use *sense* and *meaning* synonymously, and most of the time they will not be distinguished in this text either.

1.2 Road map

The LEXICOLOGY part focuses on the challenging topic of the status of the word as a linguistic construct and other word-like items, separating different types of words from affixes. This is its longest and pithiest section. It then discusses homonymy, polysemy, and its subtype, regular polysemy. Then word classes are dealt with at length. The second major section of this part looks at lexical units above the word: phrasal verbs, compounds and idioms. The notion of listeme is introduced. A summary of the various linguistic objects treated is provided at the end.

The LEXICOGRAPHY part introduces dictionaries by discussing their aims and providing a rough-and-ready typology. The subsection *Art and craft of lexicography*[5] takes a closer look (at a selection of topics) and describes some of the decisions taken in the production of a dictionary. Then some of the sources and methods used by lexicographers will be explored. The last subsection, about the future of dictionaries, looks at how much of what the LEXICOGRAPHY part offers is still valid among the mushrooming gadgetry that we are surrounded by in the 21st century.

2 Lexicology

Lexicology is sometimes defined as the study of the *lexicon*, or , *lexis*, or the *vocabulary*, or the *stock of words*, or *word stock* of a particular language; these five notions are roughly identical. Words are undoubtedly central to lexicology, but with two caveats: (i) the seemingly obvious notion of *word* must be sorted out first (and this we will do soon); (ii) given that there are by far not just words in a lexicon, but also things both below and above the word level, lexicology studies all of these, not just words. Let us call these *lexical items* or *lexical elements;* again, these two will be used synonymously. Below the words are situated the *morphemes* – it is common knowledge that these also belong within the lexicon.

Something more important should be stressed at this point: that, especially due to a syntactic bias not only in your own linguistic careers but in the history of the discipline, single words

[4] The Hungarian equivalents may be ① **belép** and ② **felvesz, rögzít**. Note that unlike Hungarian, English **enter** is transitive with both meanings.

[5] The title of Landau (2001).

have long been seen as making up the lexicon, at the expense of *multiword* sequences of different sorts. In much of modern syntax, usually *single-word items* are supposed to *"be inserted"*, which suggests a picture of the lexicon as a list of individual words (of their *lexical entries*). By contrast, many analysts today argue that knowledge of one's lexicon involves knowledge of (almost or at least) as many multiword lexical items as single-word ones. Just a foretaste of what is meant by these items in the lexicon: **mousetrap**, **flash drive**,[6] **shut down**, **download**, **give somebody a bell/tinkle/buzz/ring**;[7] **take advantage of smth; walk down the aisle**;[8] **what you see is what you get**;[9] **if it ain't broke**,[10] **don't fix it**. These various kinds of expressions are different in terms of structure, but they may be all lumped together under the heading *multiword expression*, *multiword element*, or *multiword item* (MWE or MWI) – these are all used. You may be familiar with *multiword* itself from descriptive grammar: **shut down** or **look into** are *multiword verbs*, MWV's. The term *phraseologism*, though less favoured in Anglophone linguistics, is also used to mean a MWE.

Some sources also use the term *multiword lexeme* in the sense of MWE. We, however, reserve "lexeme" for single-word items. The term *lexemic* (i.e. belonging to the lexicon) is sometimes used for them. Though MWE's undoubtedly are items of the lexicon, they are not lexemes:[11] a lexeme is an abstraction of a *single-word* element. It makes no sense to talk about the word forms of things that are not words: while **walk** is a lexical item and (the form of) a lexeme, **walk down the aisle** is (a lexical item but) not a lexeme. Only single-word items can have forms arranged in a *paradigm*, not multiword items: **come out in the wash** – another MWE, specifically an idiom – does not have a paradigm, only COME does.

The terms *lexicology*, *lexicography* and (*lexical* or *word*) *semantics* are often mentioned together. We will use a broad definition of *lexicology* that goes something like 'a branch of linguistics concerned with the meaning, use, structure, and history of words (or the vocabulary of a language)'. *Lexicography* will be defined as 'the principles/practice/process/profession of making dictionaries'. Note that thanks to the three slash marks, this is really as many as four definitions, they just get hidden; the slashes save space: ① 'the principles of making dictionaries' ② 'the practice of making dictionaries'… etc. This kind of space-saving is a standard feature of definitions in monolingual dictionaries. *Semantics*[12] in the broadest sense is a field of linguistics that investigates the meaning of expressions from morpheme to sentence (or possibly above); lexical semantics looks at the meanings of (morphemes and) words (and other lexical items), and thus may be considered as belonging to lexicology. If *pragmatics* (the study of the *use* of expressions in concrete *communicative situations*) is defined not as a neighbouring field of semantics, but as one that is included in it, then lexicology will automatically include pragmatics (this is what our definition above amounts to). Since lexicology investigates the *for-*

[6] This is **pendrive** (or **pendrájv**) in Hungarian (which, by the way, is less used in English). This situation can be compared to the German **handy** meaning 'mobile phone', and the Hungarian **boiler** (or **bojler** or even **boyler**) meaning 'water heater' – these don't at all exist in English with the same meaning (**boiler** = H. **kazán**).

[7] 'Phone smb'.

[8] 'Get married'.

[9] Not just in the IT domain, where it really means 'there is nothing hidden'.

[10] **Ain't broke** is **isn't broken**, of course. The informal saying means 'avoid attempting to correct, fix, or improve what is already sufficient. Often with an implication that the attempted improvement is risky and might backfire. (Adapted from http://en.wiktionary.org/wiki/if_it_ain%27t_broke,_don%27t_fix_it.)

[11] Compounds, exemplified here by **mousetrap** and **flash drive**, are special: they are arguably combinations of single-word elements, but they *are* words themselves, so they do have word forms. (The longer you expand a compound, however, the more perverse it becomes to talk about a *form of it*: cf. **flash *drive* chip**.)

[12] Semantics, with its compartments, is an even more complicated affair than lexicology or lexicography; that need not concern us here.

mal/structural aspect of words, *morphology* may also be seen as its component. To some analysts, the place of morphology is just here, within lexicology; to others, it is a self-contained branch; to yet others, it doesn't even exist, only as an ancillary (beside), or a compartment (within), of phonology and/or syntax. We need not take sides on this matter.

We will hardly have anything to say about the history of words; that is investigated by a rich, self-contained discipline. To be sure, the etymology of lexemes (whether they are native English or foreign) does influence their formal behaviour (e.g. the combination possibilities of *bases* and *affixes*); however interesting a subfield of morphology this may be, we will largely pass it over.

The relationship between Lexicology, Semantics and Pragmatics is *transitive:* if L includes S, and S includes P, then L includes P. Fig. 1 shows these relations.

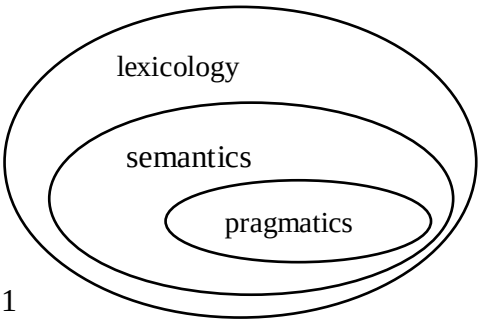


Fig. 1

2.1 What a word is

2.1.1 Lexeme, word form, grammatical word

It is a truism in linguistics that *word* is ambiguous, i.e. it is used in several, non-congruous senses. The ambiguity of *word* is all the more strange since lay people do not really sense it; moreover, words are the most easily accessible linguistic entities of which native speakers have intuitive awareness, greater than of any other object or structure of language. This is a well-known paradox in linguistics at large. When you discuss what this or that word means, you probably have one of the senses of *word* in mind, that of *lexeme:* if the form **eye** has a lexical meaning – 'the organ in your face which you see with',[13] or 'the organ of sight'[14] – then chances are that the same lexical meaning is present (completely present, and nothing else is present) in the form **eyes** as well. What you're after here, then, is the lexeme EYE (conveniently written in small capitals). If you want to explore how some word behaves in the language, however, you may want to look at different *word forms:* **eyes** may pattern, i.e. behave differently from **eye**. Consider the expressions **in someone's eyes**, **smb's eyes are out on stalks**,[15] **or cry your eyes out**: singular **eye** is not just odd but impossible here. By contrast, **a black eye**, **cast an/your eye over smth** or **the apple of smb's eye** are not good *idiomatic* English in the plural. There is a good sense in which **eye** and **eyes** are the same; and another, equally valid sense, in which they are different.

Lexeme is an abstract unit of the lexicon: you can't pronounce or write the lexeme EYE, just its word forms. This situation reminds one of the abstractness of the *phoneme* or *morpheme*;

[13] Adapted from CALD (2008). Definitions will mostly be adapted, i.e. slightly modified to focus on some aspect, or save space. Sources of definitions will not be given in all cases, so as not to clutter up the page with footnotes.
[14] Adapted from RHWUD (1999).
[15] Interestingly (but not really surprisingly if you think of cartoons), the image is similar to that in the Hungarian **kocsányon lóg a szeme** (with **szem** 'eye' in the singular rather, in my usage at least).

you can only utter *phones* and morphs, which are *variants* of the phonemes and morphemes. Word forms are thus *manifestations*, or *realizations*, or *instantiations* of lexemes (again, these three may be used synonymously).

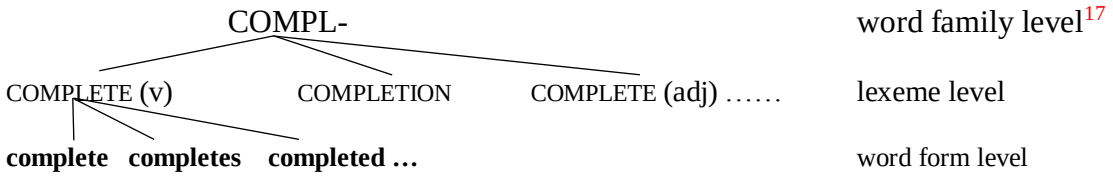
You may prefer a more *inclusive* definition of lexeme, whereby *all* types of word have lexemes: thus e.g. FOR, AND, TO, OFTEN are lexemes, although they do not have different forms realizing them. Or you may prefer a *narrow* definition, under which only words with actually different forms – that is, words with *paradigms* – count as lexemes; then only *open classes* (e.g. nouns and verbs) of words will have lexemes. This distinction, however, is not always made in treatments of the lexeme. The narrow definition would imply, e.g. that some adverbs (those with *synthetic comparatives* and *superlatives,* such as **soon**) do, while some (those with *analytic* comparatives and superlatives, such as **often**) do not have lexemes: in that system, **often** is not a lexeme while SOON is one. Also, if grammatical words did not have lexemes, the difference between the types of **that** would become more difficult to handle: the *relative* **that** (e.g. **the dog *that* I bought**) and the *demonstrative* **that** (e.g. ***that* is my dog**) have very different semantic, syntactic, and phonological properties, and this warrants talking about separate lexemes.

We will stipulate that items belonging to different word classes are automatically different lexemes, however close their meanings may be: the noun EYE is one lexeme, and the verb EYE 'look at with interest' is another. Note that other sources may approach this differently.

Another notion sometimes used to talk about lexically (formally *and* semantically) related lexemes is that of *word family:* COMPLETE (v), COMPLETE (adj), COMPLETION, INCOMPLETE (etc) are a word family.[16]

Each of these has its own concrete realizations, something like in Figure 2:

Fig. 2



The situation is even more complex than that, however. Consider the three occurrences, or "copies", of **told**[18] in the following sentences:

(i) **She *told* me that joke**
(ii) **She's *told* that joke before**
(iii) **That joke has been *told* a million times**

These are clearly *word forms* of the *lexeme* TELL that formally coincide: looked at from

[16] It must be stressed that a word family is not the same as the wider notion of *semantic field* or *lexical field*. Word families include both semantically and *formally* related items, while the members of a semantic or lexical field may be just related via meaning. Thus READY, PERFECT (v), FINISH etc do not belong to the COMPL- word family.

[17] BIG CAPITALS have been used to indicate that this level is even more abstract than the LEXEMES, but this is not standard.

[18] The verb **complete**, which featured above, could also be used (or indeed any verb whose Past Tense and Past Participle coincide), but more examples may be more helpful if only because this way the situation of one phenomenon being tied to just one example is avoided.

above, this is the same word form. But the copies of **told** in (i) – (iii) also differ: the first appears on its own, following the subject: this is the past tense. In (ii) and (iii) **told** is not alone: in (ii) it follows a form of the perfect HAVE (here in *contracted* form): this is a *past participle*, the perfect *(past) participle*. In (iii), however, **told** comes after the *passive* BE: this **told** is also a past participle: the *passive (past) participle*. Though these indeed are identical word *forms*, their difference in terms of grammar is no less important: these three are different *grammatical words* or *morphosyntactic words* (under the TELL umbrella and) immediately below the word form level. We then have four levels, rather than three, with that of grammatical words having been added. Note the four degrees of abstraction here:

Fig. 3

—word family—
———lexeme———
———word form————
———grammatical word————

You might disagree at this point if you feel that the notion of *grammatical word* is more abstract than *word form*, since grammatical word involves an additional *abstract* notion – that of grammar – while word form seems to suggest a *concrete* entity. That is just an illusion: in any *concrete utterance* it is always either this or the other word form, i.e. a particular word occurrence that appears (e.g. the past tense **told** or the participle **told**), even where it may not be clear which one. This disambiguating function of the context is true even where different lexemes are involved. To use the most-quoted textbook example: though it is clear from the real-life context most of the time, sometimes it may still not be clear which grammatical word **banks** (belonging under which BANK lexeme: the "river" bank or the "money" bank) actually occurs in a concrete utterance, but surely it can always be only *one* of them.

In (i), (ii), (iii) above, then, the (grammatical word) **told** is a kind of (word form) **told**, which is a kind of (lexeme) TELL. With TELL, the word family level has not been included, for in this case there is not much in the way of word families.

The same English word form is thus used in different grammatical functions: this is called *syncretism*. This will be further illustrated with a regular verb, INVITE, because (unlike in irregular verbs) in regular verbs you can actually find the affix that is responsible for the syncretism, i.e. is *syncretic:*

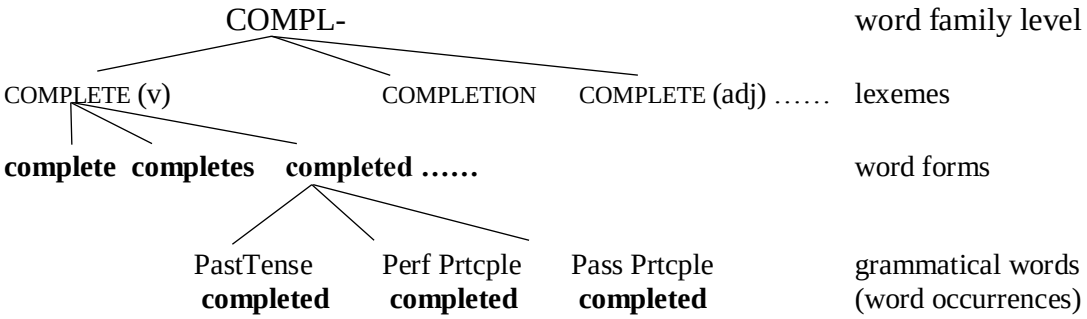
(i) **She *invited* me**
(ii) **She has *invited* me**
(iii) **I've been *invited***

It is the three **-ed** affixes that display syncretism: these are grammatically different, but have the same form; the syncretism of the whole word forms **invited** is because of that. Two enlightening ways of putting this are: (a) "the sentence structure and the word structure of a language do not mesh"; and (b) "the morphology lets down the syntax"[19] – that is, the morphology provides fewer forms than would be required by the syntax. If separate forms were available for the three grammatically/syntactically different categories (Past, Perfect Participle, and Passive Participle), there would not be syncretism. In the case of TAKE, by contrast, (i) and (ii) are different, resulting in just two forms coinciding, thus less syncretism.

Figure 4 adds the level of grammatical words to the three levels shown in Fig. 2.

[19] Both from Baerman, M, Brown, D & Corbett, G G (2005).

Fig. 4



This is fairly straightforward in English, which does not have much in the way of word forms. (Note that examples like these in (i)–(iii) do not work for *irregular* verbs with three differing "*parts*", *Present*, *Past*, and *Past Participle:* while (ii) and (iii) do have the same form,[20] (i) and (ii) are different).

In Hungarian, for example, where patterns are much richer, the differences between the three levels are more immediately obvious.

Lexemes	KÉSIK (v) 'be late'	KÉS (n) 'knife'
Word forms	——————**kések**——————	
Grammatical words	**kés-ek** 'I'm late'	**kés-ek** 'knives'

Here the lexemes themselves are not *homonymous*. Particular affixes, however, do display *homonymy:* the **-ek** on the left is a 1Sg verbal ending, while the **-ek** on the right signals the plural of a noun. This is also a kind of syncretism, i.e. grammatical homonymy, although not one involving the same paradigm. As a consequence of this grammatical, or *affixal homonymy*, the word forms themselves are homonymous.

or

Lexemes	VÁR$_1$ (v) 'wait'	VÁR$_2$ (n) 'castle'
Word forms	——————**várnak**——————	
Grammatical words	**vár-nak** 'they wait'	**vár-nak** 'to castle'

Here the lexemes themselves appear to be homonymous. This is the kind of homonymy – when lexemes, i.e. visible *citation forms* coincide – that is usually used to illustrate the phenomenon. The affixes are homonymous too: the **-nak** on the left is a 3rd person Plural ending, while the **-nak** on the right is Dative Singular case.

2.1.2 Type and token

So far we have seen that *word* really means three different things, and although some of the time *word* can safely be used with all these senses, precision may require usage of separate terms. Consider the following sentence:

He *waited*$_{(i)}$ and *waited*$_{(ii)}$, but nothing happened; he soon grew tired of *waiting*$_{(iii)}$

How many words are there in it? Thirteen is a possible answer, the one that you most probably get from most people. There are thirteen *running words* on this sentence. Is (iii) the same word as (i) and (ii)? It is visibly not the same *word form*, and consequently cannot be the same *grammatical word* i.e. word occurrence either. But is it the same *lexeme*? This is a rather difficult question for **ing**-forms, so let us focus on (i) and (ii) instead.

[20] In spelling at least, for the verb BEAR, there is strictly speaking a difference between **-borne** and **-born**. The variant **borne** is the form of the perfect past participle, while **born**, of the passive one: **She had *borne* a girl the previous year** vs **She was *born* in 1933**.
The **-borne** form, however, is also used in adjectives such as **waterborne (bacteria)** and **airborne (troops)** – and these are passive rather than perfect.

This issue is simpler: are (i) and (ii) the same word? If your answer was thirteen, then you counted these two as different words. When you count "copies", then (i) and (ii) will be different: these "copies" are termed *tokens* (and your *word count* will be the count of running words). By contrast, when you count *unique* words called *types*, these two count as the same: the sentence contains one *type* and two *tokens* of the word form **waited**. The sentence contains one type word **he** and two tokens of it; of all the other words, there is just one type and token each – to put it simply, they occur in one copy, i.e. once. Consequently, there are eleven *type words*, or *word types* in the sentence.

Note that we have taken both the type and the tokens from the *word form* level: we did not go deeper because we are not interested which of the two grammatically different **waited** words these two tokens are. (They happen to be the same grammatical word.)

Type vs token makes sense at all the three levels of abstraction: lexemes, word forms, and grammatical words, i.e. *word occurrences*. Here is a text[21] to illustrate how type and token counts can be done for these different kinds of *word*.

1 OUP has one of the world's largest and most wide-ranging language research programmes.
2 Our most important resources are the Oxford English Corpus and the Oxford Reading Programme.
3 The Corpus consists of entire documents largely from the World Wide Web, while the Reading
4 Programme is an electronic collection of sentences or short extracts that have been drawn
5 from a huge variety of writing, from song lyrics and popular fiction to scientific journals. It is
6 based on the contributions of an international network of readers who have been on the lookout
7 for instances of new words and meanings and will hopefully be doing just that in the future.

	lexemes	word forms		word occurrences
type	PROGRAMME	**programme**	==	**programme**
tokens	3x: (lines 1,2,4)	2x: (lines 2,4)		2x: (lines 2,4)
type		**programmes**	==	**programmes**
tokens		1x: (line 1)		1x: (line 1)
type	THAT$_{rel}$[22]	**that**$_{rel}$	==	**that**$_{rel}$
token	1x: (line 4)			
type	THAT$_{dem}$[23]	**that**$_{dem}$	==	**that**$_{dem}$
token	1x: (line 7)			

Note that for the purpose of this demonstration, word forms above beginning with upper case and lower case letters do not qualify as different word forms: **programme** and **Programme** are the same. The term *capitonym* is usefully sometimes used, however, for pairs like **polish** vs **Polish**, **job** vs **Job**; **march** vs **March**, **may** vs **May** and **china** vs **China**. Note that if you disregard the written language, the first two of these are minimal pairs – /ˈpɒlɪʃ/ vs /ˈpoʊlɪʃ/ and /dʒɒb/ vs /dʒoʊb/, while the rest are simply homonyms.

[21] Adapted to suit our purposes, the text is from http://www.oxforddictionaries.com/words/how-a-new-word-enters-an-oxford-dictionary.
[22] The relative word THAT (meaning 'which').
[23] The demonstrative THAT (opposite of *this*).

2.1.3 Headword, lemma, lexeme, citation form

The item we encounter in dictionaries at the head of an entry is not the lexeme itself. It is the form by which the lexemes can be represented and referred to, called the *citation form*.

Citation forms differ (a) for different parts of speech (b) for different languages; in English, verbs are *cited* by their base form, coinciding with the "bare" infinitive: **wait**; in Hungarian, their 3sg (indefinite) form: **vár**. (In other languages, still other forms may be used.) The citation form of nouns (both in English and Hungarian dictionaries) is the same: the *Nominative Singular*. Adjectives in both languages are cited by their *positive degree*; for adverbs, there is not much choice since they do not have formal variants. "Citation form" for the rest of the word classes does not make much sense if they are *nonvariant*.

This headword at the beginning of a dictionary entry is sometimes called the *lemma*. Used like this, a lemma is very much a concrete, physical (necessarily written) object, which is very different from the abstract lexeme. Even worse, *lemma* is also used in another way, basically to mean *lexeme,* the abstract notion. The term *lemma* is thus treacherously ambiguous, but also fortunately easily avoidable if *lexeme* and *headword* are used instead. The term *lemma* cannot be completely discarded, however, because *lemmatization* is a useful notion in lexicography, and this word is formed from the word *lemma* – not from *lexeme* or *headword* – and actually has to do with both of them.

2.1.3.1 "Words" in dictionaries: what is in a dictionary entry – type, token, lemma, lexeme, word form, grammatical word?

Here's a slightly modified passage from Pinker (1994):

"The computational linguist […] compiled all the *words* used in the 44 million *words* of text from Associated Press news stories beginning in mid-February 1988. Up through December 30, the list contained 300,000 *words,* about as many as [the number of *words*] in a good unabridged dictionary." (*italics mine*)

We know that those 44 million *words* are *running words*, or *tokens*, or *words of text*[24]. The 300,000 *words*, on the other hand, are unique strings of letters, i.e *unique words*, i.e. *distinct words*, i.e. (word) *types*. Neither of those figures gives us an idea about the number of either the lemmas/lexemes or the word forms that they cover. Neither do those two figures tell you the number of grammatical words in the Associated Press news corpus. But what are those seemingly simple and easy-to-count entities of which there are about 300,000 in a (rather large) dictionary, i.e the entries?

The *Merriam-Webster's Unabridged Third New International Dictionary* claims to contain over 450,000 *entries*. The *Random House Dictionary of the English Language, Second Edition, Unabridged* claims 315,000 entries – so this kind of figure is meant by the statement that a "good unabridged dictionary" contains about 300,000 words. But what do these figures mean?

British and American monolingual dictionaries typically include in their entry counts anything that is printed in *bold face*. This of course includes not just single words (of which there is probably the largest number) but also multiword lexical items – compounds and idioms – and bound morphemes – prefixes and suffixes – as well as *combining forms* (see

[24] The Hungarian term is *szövegszó*.

2.1.8.3). The *RHDEL2* and many other dictionaries also include entries for *proper nouns*, e.g. famous people and places; after all, these are meaningful and independently combining semantic units of English, elements of shared knowledge on which speakers rely when they communicate. On the other hand, dictionaries do not, as a rule, include *regular* inflected forms of a lemma, while irregulars are included, and count as entries.

While e.g. **boldface** is one word (one type) in a corpus, **bold face** would be two words; **unbolden** is obviously just one word (although **un-** and **bold** and **-en** may be included as entries in many larger dictionaries). On the other hand, in a corpus we would count in our list of types each and every form of each lemma, whether regular or not, not just the base form. In addition, in corpora, *all* proper names (places, people, companies, products, brands, etc. and not just names of famous people and places.) are counted.

According to *part-of-speech tagged* corpuses of newspaper texts, where proper nouns are tagged separately, i.e. differently from common nouns, proper nouns make up about one-fifth of the total token count. This means that the number of words in unabridged dictionaries is actually a lot larger in some respects, but lots smaller in others, than the number of words in one year of the Associated Press news corpus.

2.1.3.2 Are lexicons finite sets?

It is easy to find statements to the effect that the number of well-formed sentences in any language is *infinite*, while the number of lexical items is *finite* (as the number of rules that generate those sentences is also finite). In the light of corpus evidence, however, the dogma that the lexicon is finite has been chellenged.

The passage from Pinker (1994) mentioned in section 2.1.3.1 above continues:

"The number of possible words in a language, like the number of sentences, is infnite. […] On December 31 [1988], he found no fewer than 35 new forms, including **instrumenting**, **counterprograms**, **armhole**, **part-Vulcan**, **fuzzier**, **groveled**, **boulderlike**, **mega-lizard**, **traumatological**, and **ex-critters**.

This is the harvest of just one day. It seems clear, then, that the lexicon is *dynamic*: new words are being created – *coined* – all the time.

The lexicon of a language is an *unbounded set*; new lexical items are constantly being created. The lexicon is therefore not finite, for it is always possible to add one more item. However, it is not infinite in the sense that the number of items can be doubled, trebled, and multiplied indefnitely without affecting the nature of the set. This latter definition of infinite is the one that applies (e.g.) to the set of all numbers. It does not apply to the set of all words in any language, and it is questionable whether it applies to the set of all sentences in a language. We may say, therefore, that the set of all lexical items in a language is a *small infinite set*, while the set of all sentences is a *huge infinite set*.

2.1.3.3 Frequencies: Zipf's law

It has been observed that the most common item in any corpus of any language has twice as many occurrences (is twice as frequent) as the second most common; three times as many as the third most common; a hundred times as many as the hundredth; a thousand times as many as the thousandth; and a million times as many as the millionth. This is called *Zipfian*

distribution after the American linguist who noticed that word frequencies followed this pattern. To put it technically: the frequency of a type will be *inversely proportional* to its rank in a table of frequencies of the types in that corpus. This has implications for word counts: there will be great disparities between frequencies of more and less common words, and the most common words of all – in English, **the**, **be** (in all its forms, i.e. the lexeme **BE**), **of**, **and**, **a** – are several *orders of magnitude* more frequent than most other words. The word **the** is the commonest, accounting for one in fifteen, with a count of 6.2 million in the British National Corpus (BNC). Thus **the** is a thousand times more common than a common word like **district**, which in turn is a hundred times more common than a word like **sunburn** (which is by no means obscure).

In the 100-mn-token BNC, as we have seen, the most frequent type, the determiner **the**, had 6.2 million tokens, accounting for 6% of the total number of tokens. The second most frequent type, the preposition **of**, has 3.5 million tokens. The third most frequent type, the conjunction **and**, has 2.2 million. Just ten types – **the**, **of**, **and**, **to**, **a**, **in**, ***is***, **for**, **it**, ***was*** – account for one-quarter of all the tokens in the corpus, i.e. 25 million tokens. (Note that two of these types are word forms of the lexeme **BE**.)

2.1.3.4 Type–token ratio in different texts

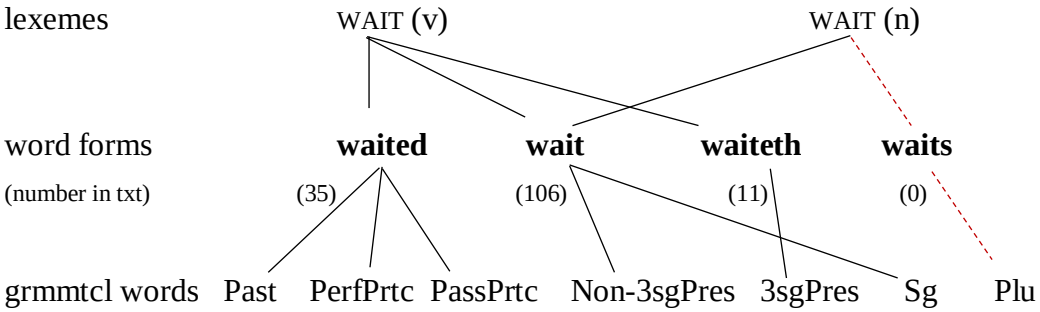
The proportion of types and tokens is characteristic of different types of text. If a passage contains 110 words (i.e. *running words*, *all the tokens* in the text), some of those *recur*, i.e. occur repeatedly. If the number of *different* words is counted, we get the number of *types*, say 60. The relationship between the number of different word forms, i.e. types, and the number of running words, i.e. *tokens*, is the type–token ratio, the TTR. TTR as a percentage can be calculated thus: (types/tokens) × 100. Accordingly, in our passage the TTR is (60:110)×100 = 54%. The lower the TTR, the more repetitions there are, i.e. the text is "looser", less condensed. A literary text has a much higher TTR than a conversation passage. TTR is low in conversations because they are less concerned with the transmission of (a lot of) information than writing. In fiction, which focusses on (elegance of) expression, the TTR is high; in news, the high TTR reflects the high density of nominal elements (people, places, objects etc).

2.1.4 Words in text

It is time to cast our net even wider. The word form **waited** appears 35 times in the King James version of the Bible (1611), a huge text indeed. **Wait** occurs 106 times. The word form **waits** figures 0 times, because this version of the Bible uses the archaic **waiteth**, of which there are 11 occurrences. **Waited** can only be a verb, so it is a word form of the *verbal lexeme* WAIT. Remember though that **waited** may be preceded either by some form of HAVE or some form of BE, in which case it is not the Past form but a different grammatical word: a *Perfect* Participle (e.g. in **have *waited***) or a *Passive* Participle (e.g. in **is *waited* for**). Moreover, **wait** may be a noun too: more technically, a word form of the *nominal lexeme* WAIT. **Waits** – absent from the text altogether – may also be a word form (plural) of the noun lexeme, or a form (3sg) of the verbal one.

These relations are shown in Figure 5:

Fig. 5



It is relatively easy to count the word forms in any text: all it takes is finding sequences of characters. Using a piece of software (which more will be said about in the LEXICOGRAPHY part), we have seen that **waited** turns up 35 times in our text.[25] If, however, you also want to find out how this word form is distributed between different grammatical words, i.e. how many of these word forms are Past Tenses and how many Participles (and of the latter, how many are Perfect and how many Passive Participles), then a simple character-string-based search is no help. Even more importantly, you can't even find out how many of the **wait** and the **waits** forms are verbal and how many nominal. And should you want to go *above* word forms and find all manifestations of the verbal WAIT lexeme, you also need to search for the word forms **waiting**, and then add up all the verbal occurrences of **wait**, **waited**, **waits** and **waiting**. That gives us the total for verbal WAIT.

The royal way to do that is by using sophisticated software that does this job for you. Grammar-sensitive counts – grammatically sensitive searches – like that are possible in (pre-)parsed and (pre)-tagged texts, i.e. texts that have been syntactically analyzed (= parsed) automatically and then labelled (= tagged) for word classes, or parts of speech.[26] That involves someone else also with an interest in corpuses (or corpora), who has prepared these large bodies of text for you. In this way you can conveniently specify any search criterion, resulting in fine-tuned searches such as: "find the plural **waits** occurrences", or "find the *nominal gerund* **waiting** forms (e.g. **everyone hates waiting**).

2.1.5 Word types: take two

A problem to do with *wordhood* different from all the previous ones comes up if we consider the following sentence:

> **He bought a**(i) **peach and an**(ii) **avocado**
> **He ate the**(iii) **peach but gave the**(iv) **avocado to the cat**

How are (i) and (ii) related? It can be argued that these are the same lexeme – the *indefinite article* – their grammar/function being exactly the same. But are they different word *forms* and different *grammatical* words? That is tricky: they are surely different in terms of form – both spelling and pronunciation – but here, their difference is not caused, or *conditioned* by the grammar/syntax, as in the case of **waited** and **waits**: their difference is caused by phonology.

---

[25] Even without a dedicated application, using the replace option of a word processor will also give you the figure you need: replacing a word with itself, the programme tells you how many replacements you had.

[26] How that is done falls outside our scope.

The difference between (i) and (ii) is *conditioned by phonology* (i.e. whether the phoneme following them is a consonant or a vowel).

How are (iii) and (iv) related? Here, too, they can be argued to be the *same lexeme*: their function (*definite article*) is the same. Whether their *form* is the same is a trifle more complicated: to (simple, *character string based*, i.e. *un-parsed* and *un-tagged*) searches, their form would be the same; they very much differ, however, in pronunciation, so their forms *are* different (the spelling just does not recognize this). The spoken difference, as before, is not conditioned by the syntax, but the phonology. Moreover, in the same fashion as above: whether a consonant or a vowel follows. It may thus be worth giving examples in (partial) transcription:

He bought **/ə/**$_{(i)}$ peach and **/ən/**$_{(ii)}$ avocado
He ate **/ðə/**$_{(iii)}$ peach but gave **/ðɪ/**$_{(iv)}$ avocado to the cat

In cases like these, i.e. with *phonologically* (and not grammatically) *conditioned* variants, we usually don't speak about different *word forms*, but refer the problem to the domain of morphology: /ə/ and /ən/ on the one hand, and /ðə/ and /ðɪ/ on the other, are *alternants* of the same *morpheme*. Another way of looking at this is that the articles are not considered to be lexemes realized by word forms, but morphemes realized by allomorphs.

This, of course, takes it for granted that words are also morphemes: *independent*, *free*, *self-contained*, *standalone* ones. We might as well say that /ə/ and /ən/ are variants of the same lexeme. The reason that many analysts do not is exactly that they reserve *lexeme* for those morphemes that do not show just phonological variants, but also grammatical variation, i.e. that have a *paradigm*, a full set of standard forms differing in terms of *grammatical features* or *grammatical categories:* the Hungarian **vár** 'wait', **vársz** 'you wait', **vártál** 'you waited' etc; **vár** 'castle', **várakba** 'into castles', **váram** 'my castle', etc; **eszik** 'eat', **ettek** 'they ate' etc; **ész** 'brain', **eszem** 'my brain', etc; and English **invite**, **invited** etc; (nominal or verbal) **wait, waits** etc. In that sense, **a** and **an** and the two types of **the** are less than words – or *less central* members of the category "word" – for they lack a paradigm. We do not recognize lexemes, then, when a lexical item shows no formal variation, either grammatical or phonological: we will not say, e.g. that **and** or **absolute** are lexemes – even though it would perhaps make sense to claim that these are lexemes that are instantiated by just one word form each.

2.1.6 Syntactic, orthographic, and phonological word

Another, no less challenging problem concerning wordhood is illustrated by the following sentences, the formal-sounding (i) and the neutral (ii) below. (The reason for (i) sounding formal is that it has an *uncontracted* **do not**.)

(i) **People do not sell their umbrellas**

The items **do** and **not** seem to have clear word status (a) *phonologically*, (b) *syntactically*, (c) *lexically*, and one might add, (d) *orthographically;* **do** is a form of one lexeme, **not** is another lexical item:

(a) Both have a vowel, both are a self-contained syllable.
(b) **Do** can leave its position in questions; it *moves*, so it is *positionally mobile:*
Do people — not sell their umbrellas?
(c) They are clearly distinct lexical items with an identifiable meaning each (even though the meaning of **do** is not *lexical*).
(d) There is a *space* between **do** and **not**.

(ii) **People don't sell their umbrellas**

There is a sense in which there simply is no **do** and no **not** here; there is just *one* element, **don't**; this, however, seems to have word status:

(a) **Don't** has a vowel, it constitutes a *syllable*. **Not** "has lost" its vowel, so it can't form a syllable; the original shape of /duː/ has changed beyond recognition, to /doʊ/. This is a completely different word, /doʊnt/.

(b) **Don't** is positionally mobile, but it cannot be split into its parts. The "do" here cannot leave its position in questions; it is not positionally mobile.[27]

(c) Whether **don't** is a *lexeme* is a tricky issue. On the one hand, it has been stuck together from two distinct (sometimes independently occurring) *lexemes*; on the other, speakers seem to *store* this item in their *mental lexicon* just like any lexeme; it's not that they glue these two elements together *"online"*.[28]

(d) The item **don't** is one *orthographic word*, with no space. Note that e.g. **no-man's-land** has no space, so it is an orthographic word, but **no man's land**, arguably the same expression in every respect, is three orthographic words. Either way, it is the same lexical item – /ˈnoʊmænzlænd/ – in both cases. But the line must be drawn somewhere, and "orthographic word" is defined just by reference to spaces. This nicely demonstrates the uselessness of the *orthographic criterion* for wordhood even in English.

There is then, a difference between orthographic,[29] syntactic, and phonological words: there are spaces between items that have the status of orthographic words; mobility characterizes those that have the status of syntactic words; and phonological independence is a feature of those with the status of phonological words.

Some syntactic words are too small, some are too large to be phonological words: **a** and **the**, e.g. are too small (they usually do not even have stress on their own); **re-randomization** and **dishwasher repairman** are too large (they have more than one stress).

Whatever is true of **don't** vs **do not** here is also valid for **mustn't** vs **must not** etc. **Don't** is one phonological word, and so is **mustn't**; both contain two grammatical (syntactic) words: **do not**; **must not**. The second item, **nt**, is a *clitic* – not a word phonologically see 2.1.9.

Represented by brackets (PW = phon. word; GV = gramm. word): $[_{PW}\,[_{GV}$ **must**$][_{GV}$**nt**$]]$

The domain for vowel harmony in Hungarian, for example, is the phonological word. This means that the noun **Buda**, which has back vowels only, can only have back vowels in sufixes too: **Budá-*ra*** 'to Buda'. The noun **Pest**, with a front vowel, can only have front suffixes, cf. **Pest-*re*** 'to Pest'. Does the name **Budapest**, with its two back and one front vowel, violate vowel harmony? No, if we maintain that the domain of vowel harmony is the phonological word, not the grammatical word. When **Budapest** is suffixed, it is a suffix for the whole compound, which is one morpho-syntactic (grammatical) word. But the suffix will have a front vowel: **Budapest-*re***, proving that the second member of the compound is a separate phonological word.

morphological structure: $[_{N}\,[_{N}\,[_{N}$ **Buda**$][_{N}$ **Pest**$]]$ re$]$[30]

prosodic structure: $[_{PW}$ **Buda**$]\,[_{PW}$ **Pestre**$]$

[27] If it did move, the **n't** chunk – incapable of occurring on its own for it has no wovel – would be left behind: ***Do people __ n't sell umbrellas?**

[28] This issue of mental storage and retrieval, which has only a relatively recent tradition in psycholinguistics, is very challenging. See the Bibliography.

[29] It is usually added at this point that this only holds for languages with a certain type of writing system.

[30] Or, alternatively, a PP (with the postposition **-nAk** as head and the N(P) Budapest as its complement).

2.1.7 Minimum free form; positional mobility, internal stability

The term *syntactic word* has been defined with reference to this property of mobility with some success for almost a century.[31] Along with *positional mobility*, usually the property of *internal stability* is mentioned, and these two characteristics of words are allegedly recognized by linguistically untrained, "naive" speakers. This strictly structural/formal definition, which makes use of the properties of *mobility* and *internal stability* (or *coherence* or *internal cohesion* or *cohesiveness* or *uninterruptibility*), basically holds that a word has just these two features, which complement each other. Recall that any definition based on *non-structural* or *non-formal* criteria is of little value. To put the definition using these formidable words simply: a word is a *minimum free form* (capable of occurring as an independent utterance), one that is stable internally, and capable of moving in larger structures. So: if it can form an utterance by itself, it is a word.

Q. **Which country is he from?**	but	Q. **Was he interested or uninterested?**
A. **Austria.**		A. ***Un.**

As the asterisk shows, this is impossible as **un-** is not a word; it is not a free form; it is a *bound form*, a *bound morpheme*. This is no news: even the lay view is that morphemes cannot stand in isolation, words can. (Put more technically: bound morphemes can't, while free morphemes can – and words are free morphemes.)

The internal stability of a word in this, syntactic sense means this: once it has been created, it is closed for the syntax; no syntactic rule can change the contents of "the brackets around a word"; its internal structure is a no-go area for the syntax. The order of word elements is fixed and cannot be altered: **un-interest-ed** but neither ***interest-ed-un** nor ***un-ed-interest**.

2.1.7.1 Use vs mention

Caution is needed here because as a citation of a grammatical form, *any* lexical item may be used as an utterance, even the morpheme **un-** we saw above: the following dialogues are both grammatical. That, however, is not what is wanted: for a real *diagnostic test* of independent wordhood, we need words *used*, not *mentioned*.

Q. **Which article did you mean?**	and	Q. **Which prefix was used?**
A. **"An".**		A. **"Un-".**

Here, the elements **an** and **un** are mentioned, i.e. "quoted" from the English language, i.e. they are *metalinguistic:* they are equivalent to saying "the article **an**" and "the prefix **un-**".

In the following sentences the nouns are used (i–ii) and mentioned (i'–ii'). In writing, *mention* is specially marked usually by italics, underline, boldface, (or quotes – but this is discouraged); in this text, examples are bold anyway, so we made an exception in (ii').

(i) **Dogs are cute**	vs	(i') ***Dogs* has 4 sounds**	or **Dogs has 4 sounds**
(ii) **Men are physically stronger**	vs	(ii') **Men** is irregular	or **"Men" is irregular**

The fact that **an** (and many others whose word status we would like to "save") cannot appear as an independent utterance certainly changes the picture of words that we have had so far. Many of the lexical items so far are words under this definition, but apparently **the** and **a(n)** are not. Many similarly short determiner-like items, by contrast, do occur on their own, so they are words: **some** and **this**, for example, can constitute utterances.

[31] Associated with the name of the American linguist Leonard Bloomfield.

2.1.7.2 Central and peripheral members of categories

One solution to the problematic status of **a** and **the** is to say that "all words are equal, but some words are less equal than others". There would be *central* and *peripheral* members of the category *word*. Note, however, on the one hand that there is no disputing the fact that **a** and **the** (and e.g. **though**, **and**, **because**) are less equal/free than **beer** and **dogs**; and on the other, that categories – not just in language but "out there" – do indeed seem to be like that: *fuzzy*, or *hazy*, or *indeterminate*, with both (more) central and (more) peripheral members. This challenging categorization problem will come up in the discussion of word classes and then of idioms again (and is now surfacing in linguistics more often than before).

2.1.8 Between words and affixes

One way of avoiding the fuzziness of the "some words are more equal" situation, that is, of retaining the crisp categories desirable to many linguists, is to set up subcategories within or below the word level. Setting them up does not mean *inventing* them but actually *finding* them: because *they're there.* The claim then is that there is a domain between undisputable words at the upper end of some scale, and obvious affixes at the bottom. The lay person, unfortunately, is totally unfamiliar with this domain, which – alas – is populated by diverse kinds of elements.[32]

At the top, a lexical element that satisfies the above wordhood conditions – free form, showing uninterruptibility (internal stability, grammatical cohesion) and positional mobility – is [1] an *autonomous* (or *independent*) *word* – **beer, dogs**. An element that is bound (i.e. cannot occur on its own) but can be separated from another (bound or free) element by an autonomous word is a [2] *dependent word*. This way we can go on calling the forms **a(n)**, **the**, **and**, **though** etc *words:* we have saved their word status, which is intuitively felt anyway. Even less autonomous are [3] the *semiwords*, which are the initial and final constituents of *compounds* that cannot occur outside of compounds (i.e. are not free), but can undergo *coordination deletion* both forward and backward (see 2.1.8.2). Finally, [4] genuine *affixes* are bound, and also disallow coordination deletion in either direction. [1]–[2] are words, strictly speaking, while [3]–[4] are non-words.

Note that in this system *free form* and *bound form* are different from the usual approaches: here, boundness reaches up to the (dependent) word level.

We will look at [2] and [3] more closely: what is meant, in [2] by an element being bound but being able to be separated from another (bound or free) element by an autonomous word? And what is forward and backward coordination deletion mentioned in [3]?

2.1.8.1 Dependent words

The article **the** in (a) and the preposition **on** in (b) below are dependent words because (although they cannot constitute an utterance in themselves) they can be separated from another (free or bound) form by an autonomous word. In (a2), the article **the** – a dependent word – is preceded by the italicized ***on***, and followed by the italicized ***same***; in (b2) the preposition **on** – a dependent word – is preceded by the italicized ***lustily***, and followed by ***many***. So, autonomous material can be inserted between the dependent words and the forms preceding or following them.

[32] This analysis is based on Kenesei (2007).

(a1) **near** **the** **premises**
(a2) **near** *or on* **the** *same* **premises**

(b1) **singing** **on** **roofs**
(b2) **singing** *lustily* **on** *many* **roofs**

2.1.8.2 Semiwords

Is the /laɪk/ in **beer-like** (or **beerlike**) an affix? Or is it more independent, i.e. more word-like?[33] Note that this is not the **like** in **I *like*(i) beer** or the **like** in **This is *like*(ii) beer**. The meanings of **like**(i) and **like**(ii) contrast starkly: **like**(i) is 'be fond of'; **like**(ii) is 'similar to'. The meaning in **like**(iii) is rather close to **like**(ii) – similarity – but grammatically, **like**(iii) is clearly different. While **like**(ii) is followed by its complement, **-like**(iii) is preceded by it: **like X** vs **X-like**. There may be a hyphen in **like**(iii), which suggests a difference. Note also the Hungarian translations, which suggest that all of these are different items.[34] Let us now see how the "hyphenated"[35] **-like** behaves. We will first see how compounds behave.

In *coordinated compounds* – such as the coordination of **wine bottles** and **beer bottles** – it is possible to *delete* the second constituent in the earlier, first *conjunct*, while keeping the identical constituent in the later, second conjunct:

[**wine bottles**] **and** [**beer bottles**] → [**wine** Ø] **and** [**beer bottles**]

[**beer bottles**] **and** [**wine bottles**] → [**beer** Ø] **and** [**wine bottles**]

Wine and **beer** are autonomous words. Although the **-like** that we are investigating is not autonomous, its deletion under the same conditions is still possible:

[**wine-like**] **and** [**beer-like**] → [**wine-**Ø] **and** [**beer-like**]

[**beer-like**] **and** [**wine-like**] → [**beer-**Ø] **and** [**wine-like**]

While the item **-like** allows this, nothing *below this* level does: *no affix can be deleted in this way*. It is impossible to delete the second constituent from the first conjunct, while keeping the identical constituent in the second conjunct if they are affixes. The following strings are ungrammatical, showing that **-ing** and **-ed** mere affixes:

[**laugh-ing**] **and** [**sing-ing**] → *[**laugh-**Ø] **and** [**sing-ing**]
[**sing-ing**] **and** [**laugh-ing**] → *[**sing-**Ø] **and** [**laugh-ing**]

[**laugh-ed**] **and** [**shout-ed**] → *[**laugh-**Ø] **and** [**shout-ed**]
[**shout-ed**] **and** [**laugh-ed**] → *[**shout-**Ø] **and** [**laugh-ed**]

---

[33] Pun unintended.

[34] In Hungarian, (i) is ***Szereti* a sört**, (ii) **Ez olyan, *mint* a sör**, and (iii) is **sör-*szerű***. The hyphen highlights the same difference between (iii) and (i)–(ii) as in English.

[35] The scare quotes are needed because (i) in speech, there obviously is nothing there; (ii) the hyphen is optional anyway.

Hungarian also shows proof of this: you cannot delete, e.g., the Accusative Affix in similar conditions:

[**bor-t**] **és** [**sör-t**] 'wine and beer Acc' → *[**bor-**Ø] **és** [**sör-t**] 'wine and beer'
[**sör-t**]**és** [**bor-t**] 'beer and wine Acc' → *[**sör-**Ø]**és** [**bor-t**] 'beer and wine'

Just as in English, of course, you can delete the semiword **-szerű** '-like':

[**bor-szerű**] **vagy** [**sör-szerű**] 'wine-like or beer-like' → [**bor-**Ø] **vagy** [**sör-szerű**]
[**sör-szerű**] **vagy** [**bor-szerű**] 'beer- or wine-like' → [**sör-**Ø] **vagy** [**bor-szerű**]

This element, then, is a semiword: less than a dependent word but more than an affix.

Dependent words, as we have seen, include various grammatical words such as determiners, prepositions, conjunctions, and clitics. Recall, however, that not all members of these classes are dependent: **the** is a dependent word, while **this**, e.g., is autonomous. The most challenging kind of dependent words are the *clitics*.[36]

To sum up this section: the system of lexical items "downward" from and including the autonomous word looks like the following:

Fig 6

————————autonomous *word*————	e.g. **wine**, **drink**, ***some***
——————dependent *word*————	**a**, **the**, **'ve**, **'s**
————semi*word*————	**-like**
——affix——	**-ed**, **-ing**

2.1.8.3 Neo-classical compounding

Typical examples of forms that seem to be halfway between free and bound forms are the *combining forms*, which produce a type of compound. While normal compounds are produced from free forms (free root morphemes), these neoclassical or *Greek-type compounds* are special in that their two members are bound. There are *initial* members and *final* combining members. Examples include **astro-**, **biblio-**, **bio**, **geo-** and **xeno-** for initial combining members, and **-naut**, **-graphy**, **-logy**, **-phobia** for final members, which give you the lexemes ASTRONAUT, BIBLIOGRAPHY, BIOGRAPHY, BIOLOGY and XENOPHOBIA. The hyphens indicate that these lexical items cannot occur as free forms. A few of them may be both initial and final combining forms, e.g. **morph-** & **-morph**; **phil-** & **-phile** as in *MORPH*OLOGY and POLY*MORPH*; *PHIL*OSOPHY and ANGLO*PHILE*.

These combining forms – Latin and Greek roots – cannot be affixes, since affixes cannot combine with each other; thus they are best analyzed as roots, albeit unusual ones. Boundness, then, unites affixes (which are by definition bound) and these *bound roots*. Note that many of the actual compounds are English: they never existed in the source languages.

Combining forms can not only combine with (i) each other (as the examples above), but also with (ii) words and (iii) bound morphemes: the word *MORPH*OSYNTAX exemplifies the initial combining form **morpho-** linking up with a word; SCIENTOLOGY is a final combining form, **-logy** combined with a *bound root*. The **-o-** element is best seen as a linking vowel.

Combining forms, treated and termed variously in different sources, are semiwords under the classification introduced here.

[36] The Hungarian term for clitic – *simulószó* – suggests that it *is* a word, a "clinging word".

2.1.9 Clitics

Consider the following **'s** genitives:

(i) [$_{DP}$ ***Mom***]**'s car**
(ii) [$_{DP}$ **the girl from *Italy*]'s car**
(iii) [$_{DP}$ **the girl we talked *about*]'s car**

In these constructions, the genitive **'s** that is attached to its phonological *host* (italicized here), which is – obviously – always the last *word*, is a kind of *clitic*. Syntactically, this genitive **'s** belongs to the whole *phrase*[37] (bracketed here) – that is why the **'s** is not a genitive *case suffix:* affixes do not behave like this.

A genuine case ending belongs to the *head*, not to any old word that happens to be last one in the phrase. In (iii) above it is even more evident that the **'s** cannot be a suffix, because prepositions cannot be suffixed.

To better see how this works, we may go to Hungarian again, which has Case suffixes. The nominal phrase in [**a lány-t Olaszországból**] 'the girl-Acc from Italy' has the Accusative suffix on the head noun **girl**. The ***[a lány Olaszországból-t**] version is impossible.

When a clitic follows the host, like in (i)–(iii), it is an *enclitic*. When it attaches to the beginning, it is a *proclitic*. Clitics that form a *prosodic unit* with a host on their *left* are enclitics; clitics forming a unit to their *right* are proclitics.

Another, very different, classification of clitics is into *simple* and *special clitics:* a simple clitic can be replaced with some *more independent* word of which it is a kind of *contraction*. The genitive **'s** is not the contraction of any self-standing lexical item; it is a *special clitic*.[38]

Consider, however, the following **'s** forms: these are contractions of the word form **is**. These are *simple clitics*.

(i') [**Mom**]**'s driving fast**
(ii') [**The girl from Italy**]**'s driving fast**
(iii') [**The girl we talked about**]**'s driving fast**

The various forms of the auxiliaries HAVE, BE, CAN, WOULD, MIGHT, SHOULD etc when contracted, can attach as simple enclitics – **'ve**, **'s; 'm, 're** and **'s** – to their host, i.e. the last word in the Subject DP immediately preceding them:

they've eaten	**she's got it**	**I'm through**	**bag's are being emptied**
ˈðeɪv ˋiːtən	ˈʃiːz ˋgɒtɪt	ˈaɪm ˋθruː	ˈbægz ə (ˈ)biːɪŋ ˋemptɪd

Thus, reduced auxiliaries (and **n't**) are enclitics. Articles and prepositions are usually proclitics (but may also be stressed, when they are not, cf. /eɪ/, not /ðiː/ Gates; / ˈfɔːmɪ/, / ˈtuːjə/).

Beside e.g. **wouldn't**, **would've**, **we'd** and **we've**, more complex *cliticization* also happens:
We *wouldn't've* – (= would not have) – /wʊntəv/ **thought this was a problem**
We'd've – (= we would have) /wɪdəv/ **– helped you if you had asked us**
but not in combination: ***We'dn't've** (= we would not have) ***/wɪdəntəv/

[37] The genitive phrases (DPs themselves) are usually claimed to be in the specifier positions of these larger DPs.
[38] It is sometimes termed a *phrasal affix*, which, unfortunately, suggests that it *is* an affix.

2.1.9.1 Four types of English clitics

The table summarizes the dual classification of clitics.

	Clitics	
	"variant"	"self-contained"
Proclitics	**d'** as in **d'you see?** (< **do**)[39] **y'** as in **y'all**[40] (< **you**) **'t** as in **'twas** (< **it**) etc	Prepositions Articles
Enclitics	Reduced and contracted auxiliaries **I'd** [< **had**] **known** **I'd** [< **would**] **know** etc	Genitive **'s**

The genitive **'s** is the least wordlike of all these clitics; its status is actually closest to affixes – but as we have seen, it is not one. Certainly, while (i) prepositions, articles and auxiliaries are felt to be *words* by speakers, and (ii) simple clitics are variants of such words, (iii) the genitive **'s** is different, for it has no word status for the naive speaker.

Note that generally, the apostrophe's function is to signal that something is missing, but that is not what is does with the genitive **'s**.

It is not the *direction* of cliticization that is signalled by the apostrophe, but the *omitted* part of a word (i.e. in those cases where it signals omission). The reason that the **'d** (with the **'** on the left) is an enclitic is *not* that the host is on the left, but that material is missing from the left: only the **'d** remains of ~~**would**~~ or ~~**had**~~. And the reason that the **d'** (with the **'** on the right) is a proclitic is *not* that the host is on their right, but that here, omission is from the right: just the **d'** remains of ~~**do**~~. The case of **'twas** shows this clearly: though the **'t** is as a proclitic, the apostrophe is on the left because the vowel of **it** is missing here.

Apparently, in a more subtle analysis, there is a zone of discrete points – or there is a *continuum* – between affixes and words. If this is so, it means that there are clear-cut subcategories – or there is a shady area – between derivation and compounding, since in derivation it is affixes that take part, while the constituents of compounds are words. The phenomenon of dependent words and semiwords, and – within the dependent words – clitics, a very heterogeneous group anyway,[41] suggests that the demarcation is not as straightforward as it may seem.

Problems of demarcation come up not just between affix and word, but also between phrase and word. This will be the topic of the *Above the word* section.

[39] Or (< **did**) as in **did you see?** – but (< **do**) is more frequent.
[40] The form **y'all** /yɔːl/ is a 2Pl personal pronoun mainly in Southern AmE.
[41] Even more than is suggested by the above discussion.

2.2 Homophony, homonymy, homography and polysemy

Recall that *syncretism* was aptly characterized[42] as the case of the morphology letting down the syntax: when there is just one morphological shape serving distinct syntactic functions. *Homonymy* may no less pertinently be called the situation of "the lexicon letting down the semantics": the case of there being just one lexical form for two distinct – different, unrelated – meanings. (Let us state right away that this act of "letting down" is not harmful in either case, should the image suggest that.) In *polysemy*, there is also one form with several meanings, but these are related, and usually felt to be so by speakers. Note that etymology as a criterion has not been used: it happens that senses that are *not* felt to be related still go back to the same source etymologically, and the other way round. Because speakers have no etymological information, the history of words will be irrelevant for our purposes.

The traditional definition of the three *"homo-X"* terms runs something like the following.

Homo*phony* is the case of two (or more[43]) words pronounced identically (though spelt differently). Examples are:

/nɒt/ **not** = **knot**	/tuː/ **two** = **too**	/reɪz/ **raise** = **rays**
/eɪt/ **ate** = **eight**	/biː/ **be** = **bee**	/hɜːts/ **hertz** = **hurts**
/wʌn/ **one** = **won**	/soʊl/ **soul** = **sole** etc.	

/raɪt/ **rite** = **write** = **right** (= **Wright**) /sent/ **scent** = **cent** = **sent**

Homo*nymy* is the situation of two (or more) words pronounced identically and *also spelt* identically, e.g.:

stalk$_1$ 'part of plant' vs **stalk**$_2$ 'follow/harass', both are /stɔːk/;
left$_1$ 'Past of *leave*' vs **left**$_2$ 'opposite of *right*', both are /left/;
bear$_1$ 'animal' vs **bear**$_2$ 'carry', both are /beə/;
sole$_1$ 'bottom of shoe' vs **sole**$_2$ 'only' vs **sole**$_3$ 'type of flatfish', all are /soʊl/.

Homo*nymy*, under this definition, is a subtype of homo*phony*, with the written forms also falling together. If you recognize – some would say fetishize – the primacy of spoken language,[44] however, then you might want to have just one category, homo*nymy*; then homo*graphy* will at best be a peculiarity, just a footnote within homo*nymy*. In that system, homo*graphy* – as a cultural but non-linguistic, or at least a secondary linguistic phenomenon – simply disappears.

Homo*graphy* is different from both: it is the case of two (or more) words pronounced differently (a normal/expected case, after all they're different!) and *still* spelt identically, e.g.:

tear$_1$ /tɪə/ 'drop of liquid' vs **tear**$_2$ /teə/ 'pull apart';
row$_1$ /raʊ/ 'quarrel' vs **row**$_2$ /roʊ/ 'cause boat to move';
bow$_1$ /boʊ/ 'weapon' vs **bow**$_2$ /bau/ 'bend the head';
read$_1$ /riːd/ [Present] vs **read**$_2$ /red/ [Past].[45]

---

[42] Baerman, M, Brown, D & Corbett, G G (2005).

[43] This is by far less typical.

[44] Recall that speech is primary both in the life of an individual and the history of languages (and there are both individuals and languages that lack writing).

[45] Note that to the learner, these may present *reading* problems, for it is difficult to dissociate the two pronunciations, but are unproblematic (after a while) in *spoken* production, where you know which of the two you want to retrieve and use.

Note that the same item can occur in more pairs: **sole** contrasts with **soul** as a pair of homo*phones*, and features among the homo*nyms*, too.

Advanced

While homo*nymy* is not infrequent in Hungarian, the following pairs illustrate homo*phony*, i.e. homonymy with different spellings. In the verbs on the right, the consonant (underlined) is the result of the blending of /t/+/ʃ/, as seen in the spelling <ts>, while in the nouns on the left, it is "organic", spelt <cs>.

rikkan<u>cs</u> 'newsboy'	**rikkan<u>ts</u>** 'shriek!'
ron<u>cs</u>uk 'their ruin'	**ron<u>ts</u>uk** 'we should ruin it'

Polysemy is usually defined as a case of *lexical ambiguity*[46], the *multiple*[47] *meaning of a word*. The point of course is that it is *one word* which has several meanings, while in the case of the *homo-X* terms above, it is always *two (or more) words* that are involved. The meanings are obviously related – this is why we speak of *one* word rather than many.[48] Typical examples include words with one basic and one (or more) *transferred*, or *metaphorical*, or more abstract senses, e.g.:

the **mouth** ① of a person ② of a cave ③ of a river;
the **legs** ① of a person ② of a chair;
the **foot** ① of a person ② of a hill.

These are usually based on (different types of) similarities, e.g. **mouse** ① animal ② a shy, quiet person ③ computer peripheral. The problem is that when a definition mentions the notion *metaphor*, it does a great disservice to (the student of) linguistics, because *metaphor* suggests something special, out-of-the-ordinary, literary – which means that polysemy will be thought of in similar terms. Polysemy, however, is typical, frequent, an everyday phenomenon happening all the time in all languages – cf. the open (as opposed to closed) word classes: the stock of polysemous words is continuously expanding. This is what differentiates them from homonyms: homonyms are special, sometimes surprising and funny, and relatively rare.

2.2.1 Regular polysemy

Polysemy in language often comes in regular sets. In English, for example, it is often the case that the same word doubles up as the name of trees (which is a *countable*, *count*, or *unit* noun) and their wood (which is an *uncountable*, *non-count*, or *mass* noun): e.g. **an oak** vs **made of oak**. Another like situation is with (count) animals and their (non-count) meat: e.g. **two chickens** vs **some chicken**. There are many regular polysemy classes in English.

In *regular polysemy*, words have a particular relationship with other words in their *lexical set* in such a way that several of their meanings parallel each other. This means that if *Word1*

[46] Recall that *ambiguity* is of two types, *lexical* and *structural*; the latter may be exemplified by **long nights and days**, which means either (i) 'long nights and *long* days', or (ii) 'long nights, plus days of indeterminate length'.
[47] Note that "multiple" rather than "double" is typical here; cf. Note 43.
[48] The Hungarian terms for homonymy and polysemy are more than unhelpful: azonos alakú *szavak* '<u>words</u> with identical form' and többjelentésű *szavak* '<u>words</u> with several meanings', with their plural *szavak* does not make it obvious that with homonyms, it is *several* words, while with and polysemes, it is *one* word that is involved.

has meanings *a, b*, then *Word2* will have meanings *a', b'*, and *Word3* will also have meanings *a'', b'', and so forth.* One of the best-known examples is the container–contents polysemy.

This can be observed in the following examples:

container	contents
put the *can* in the fridge	**eat the whole *can***
drop the *glass*	**drink the whole *glass***

and so on with CUP, BOTTLE, PACKET, BOX etc, that is, words belonging to the same lexical set.

This *systematic polysemy* is very different from the (obviously motivated but) haphazard polysemy seen in the standard textbook examples. People and tables have **legs** in English and Hungarian (**az asztal lábai** 'the legs of the table'), but not in French, where chairs have feet (**les pieds de la table** 'the legs of the table'); both people and needles have **eyes** in English, but not in Hungarian; a clock has hands in English, but not in Hungarian, where it has **mutatók** 'pointers'. Regular polysemy, by contrast, seems to be universal (but of course not absolute).

A fairly universal regular polysemy type is illustrated by the pairs (i) **open this *book*** vs (ii) **disagree with this *book***. This may be seen as a special case of container–contents polysemy: the former refers to a physical object, the latter to its content. Another typical example is adjectives describing how people feel; these can be used of actions as well. For example, ANGRY may be argued to be polysemous, with one meaning that goes e.g. with ① CROWD, MOB, or RESIDENT, and another that is used e.g. with ② VOICE, FACE, or OUTBURST.

Regular/systematic polysemy in English

1a There's a **squirrel**
1b We don't eat **squirrel**

2a There's a **mink** near the river
2b She wore a **mink** coat

3a He had his hands in his **pockets**
3b He **pocketed** the change and ran off

4a **Shakespeare** wrote plays
4b It's in **Shakespeare** somewhere

5a She doesn't drink **coffee**
5b Three **coffees** please

6a That looks like **silver**
6b It's a **silver** bracelet

7a Do you play the **cello**?
7b The **cellos** came in late

8a I like your **jacket**
8b They've got your **jacket** in the window

9a Have you been to **Rome**?
9b **Rome** denied this

10a She stood by a tall **pine**
10b The desk was made of **pine**

11a I haven't got a **fork**
11b He **forked** the peas into his mouth

12a They all carry **knives**
12b He got **knifed** during a robbery

2.2.2 Polysemy: curse or blessing?

Polysemy is far from being a *defect* of language: it is essential for its efficient functioning. So all-pervasive and indispensable is it that if it did not exist, we would need many more words than there are now. This would mean, for instance, that where now there is just one **raise**, **eat**, **write** and **hurt** (these verbs have been randomly chosen), there would have to be 33 different words for the 33 senses of **raise**; 9 for **eat**; 17 for **write**, and 9 for **hurt**, excluding all kinds of multiword expressions, such as phrasal verbs.[49] This is 68 separate meanings against the 4 polysemous ones available now, for which 68 individual words would be needed. Without polysemy, i.e. being able to attach several senses to a lexeme, vocabularies – mental dictionaries – would be an unbelievable burden on speakers' memory. Polysemy is absolutely needed for language to function in an economical and flexible way.

An important point about homonymy and polysemy is that there is *no real dichotomy* here: either member of a homonymous pair can, and often does, have several senses, thus a given example may illustrate both phenomena. The word **bear**$_1$ was quoted with the gloss 'animal' above, but it also means 'person who sells shares when prices are expected to fall', as found in a learner's dictionary,[50] and has eight nominal senses (two of them with a capital B) in RHWUD (1999). The word **bear**$_2$ was given in the 'carry' sense above, but it can also mean 'tolerate' and 'give birth', among many other things. So homonymy and polysemy – which is widespread anyway – often occur together.

An even more important point that is usually missing from discussions of the "homonymy vs polysemy" issue is the following question: Which of the senses of *word* is meant in these definitions – indeed in all of the above paragraphs? Lexemes, word forms, or grammatical words? Polysemy typically characterizes *all* grammatical words belonging to a lexeme, so it must be the entire lexeme's property: singular **bear** is just as polysemous as plural **bears**, and present-tense **bear** is just as polysemous as past-tense **bore** or the participles **born** and **bearing**. Polysemy, then, appears at the lexeme level – although what we physically observe is the polysemy of word forms and grammatical words or word occurrences, because these *inherit*, as it were, the polysemy from the lexeme. The apparent, "inherited" polysemy of word forms is just *secondary*. Homonymy, by contrast, is a property of word forms, a kind of coincidence: the singular **rose** is a homonym of past-tense **rose**, but that relation is not between two lexemes, just between two random forms of these lexemes.

The Hungarian nominal **várnak** 'to castle', e.g. is homonymous with verbal **várnak** 'they wait', and nominal **várunk** 'our castle' is homonymous with verbal **várunk** 'we wait', but not *all* forms of the lexemes VÁR$_{(n)}$ and VÁR$_{(v)}$ coincide – this means that homonymy is not the lexemes' property. Again, the homonymy of the word form may be inherited by the grammatical word, thus the homonymy of the grammatical word is secondary.

To sum up: because homonymy and polysemy are not applied to the same "word" notion, they cannot, strictly speaking, be seen to oppose and logically complement each other.

[49] RHWUD (1999).
[50] CALD (2008).

Ignorance of this fact is not really harmful – but it goes some way towards explaining why, when homonyms need to be illustrated in a Hungarian grammar class, it is always two *citation forms* (the visible forms of lexemes) that are used, never word forms. **Lép**$_1$ 'step' vs **lép**$_2$ 'spleen'; **nyúl**$_1$ 'rabbit' vs **nyúl**$_2$ 'reach out'; **ég**$_2$ 'sky' vs **ég**$_2$ 'burn' are typical examples, which, while nicely showing the unrelatedness of senses, creates the false impression that it is lexemes that are opposed here. Many of the *forms* of these lexemes, i.e. members of their *paradigms*, do actually display homonymy, but many do not: **lépsz**$_{(v)}$ 'you step' and **lépet**$_{(n)}$ 'spleen-Acc' are not common forms of the lexemes LÉP$_1$ 'step' and LÉP$_2$ 'spleen'; **nyúlt**$_{(v)}$ 'reached out' and **nyulat**$_{(v)}$ 'rabbit-Acc' are not forms shared by NYÚL$_1$ 'reach out' and NYÚL$_2$ 'rabbit'; and **éget**$_{(v)}$ 'burns' and **égen**$_{(n)}$ 'in sky' are not common forms of ÉG$_1$ 'burn' and ÉG$_2$ 'sky'.[51] By contrast, **lépünk** would be a perfectly good example, because it is a shared form of LÉP$_1$ and LÉP$_2$ and it means both 'we step' and 'our spleen'. Also, good examples of forms that could be used never, or very seldom, are because their citation forms differ, e.g. **követ** (Accusative) vs **követ** (Nominative), which are coinciding forms of the lexemes KŐ 'stone' and KÖVET 'envoy'.

Where it is not the citation forms (of two lexemes) that coincide but word forms, i.e. members of their paradigms, it is also customary to talk about *grammatical homonymy*, as in the case of **lépünk** or **követ**. This has also been referred to as syncretism. It also happens that the formal coincidences, i.e. grammatical homonyms, are within one and the same lexeme's paradigm: Hungarian **ennék** 'I would eat' vs **ennék** 'they would eat it', for example, both belong to ESZIK 'eat'.

2.2.3 Polysemy into homonymy

What starts out as polysemy often grows into homonymy. When the relatedness of senses is no longer felt, one polysemous word – a *polyseme* – is "split", i.e. becomes two homonyms. This has happened, e.g. to the lexemes HORN or the Hungarian TOLL.

To a generation of English speakers, HORN used to be polysemous, with two meanings: ① 'hard, pointed, part growing from animal's head' ② 'musical instrument'. Then, when cars came on, it may still have been just one polysemous lexeme, with a third meaning, ③ 'device in vehicle used as signal'.[52] Sense ② was similar enough to ①; then ③ was still similar to ① and ②. Gradually, however, the senses became dissociated – came to be felt unrelated as native speakers were no longer aware of their connection – yielding three homonymous lexemes, HORN$_1$ vs HORN$_2$ vs HORN$_3$.

There are now two Hungarian lexemes, TOLL$_1$ 'feather' and TOLL$_2$ 'pen', which used to be obviously related: pens can still be traced back to the flight feathers of large birds used as writing implements.[53] Fountain pens, and especially other kinds of pen, have long been dissimilar to these quill pens, but once the connection was obvious.

If this situation holds *diachronically*, then it must be the case that *synchronically*, too, there is a huge grey area between polysemy and homonymy, and this area is constantly changing in size. This also indicates that relatedness of meaning is always a matter of degree, since native speakers' (education, linguistic awareness, and thus) intuitions for many words differ. The polysemous MOUTH was mentioned above as having one basic and probably more metaphorical meanings: but just how many different metaphorical ones? The basic sense of MOUTH is surely ① 'opening in face', and ② is probably a different sense in **the mouth of a river**; but there

[51] NYÚL$_1$ and NYÚL$_2$ actually have few shared word forms except for **nyúl** itself, **nyúlnak**, and **nyulunk**$_{(v)}$ = **nyúlunk**$_{(n)}$ when pronounced identically.

[52] Both adapted from CALD (2008).

[53] Latin **penna** means 'feather' (which happens to be an obsolete/poetic word for 'pen' in modern Hungarian).

seem to be more senses. If caves, tunnels, and jars have the *same* kind of object, then we're looking at just three meanings here, the ① face ② river, and ③ cave/tunnel/jar kinds of meaning? But we may be more true to the facts if we split ③ into three different senses – these are all different. But a "cave mouth" and a "tunnel mouth" are more similar to each other than they are to the "jar mouth". And come to think of it: the mouth of a river is *not an opening* at all…

Since this sense identification and *demarcation* problem is hard for the analyst, native speaker consensus can hardly be expected. Are the meanings of **run** in **the buses don't *run*** and ***run* for president** and **his nose is *running*** and ***run* me water** *related?* If you ***play* chess** in the morning and ***play* Hamlet** in the evening, are these related senses? Probably. And when you ***play***[54] **a card**? And when you ***play* the flute**? Do not say that "they must be related, otherwise **run** and **play** would not be used in all these cases": after all the same **bear** form is used in **can't *bear* him** and **the *bear* ran towards us** – and in totally unrelated meanings.

2.2.4 Polysemy vs vagueness

Not all instances of meaning imprecision are cases of polysemy. An expression can be argued to be *vague* in several ways, in which case its sense is imprecise. To be polysemous, it must have at least two separate senses. The lexeme BROTHER is not polysemous just because it can mean both 'younger male sibling' and 'older male sibling'; nor does the existence, in other languages,[55] of special lexemes for these two different notions make BROTHER polysemous: its sense is vague (with respect to age). FRIEND is *vague* with respect to sex, not ambiguous between 'male friend' and 'female friend'. By contrast, GIRLFRIEND displays polysemy: it has two disjoint senses: ① 'female friend' ② 'woman/girl who somebody has a romantic or sexual relationship with'. The lexemes MIDDLE-AGED, RICH or DILIGENT are not polysemous either, only vague in the sense of *ill-defined*.

REPTILE or WEAPON may be argued to be not even vague, just *general*: if someone says they saw a reptile, or that they used a weapon, or that they used a knife, what they saw/used is as clearly delimitable as if they had said they saw/used a snake/gun/penknife, just REPTILE/WEAPON/KNIFE as terms are more general (SNAKE is a *hyponym* of REPTILE, and REPTILE is a *superordinate* of SNAKE; GUN is a hyponym of WEAPON, and WEAPON IS a superordinate of GUN; PENKNIFE is a hyponym of KNIFE, and KNIFE is a superordinate of PENKNIFE).

2.3 Words come in classes. Or do they?

In this, as in almost any linguistic text, *word class* or *syntactic category* or – to use the most traditional term – *part of speech* is taken for granted. These are more or less interchangeable, but this does not mean that there is consensus on what they cover or what their status is.[56] *Part of speech*, often abbreviated to *PoS*, tends to be used recently in computational linguistics, or language technology, contexts.

Word classes seem to most linguistics students one of the quite straightforward notions in grammar, traditional or theoretical. That meaning is not the number one criterion in setting up word classes, or even that it is best left alone, may surprise many students, but the general view is that apart from that, the situation is unproblematic.

---

[54] One definition is 'choose from the ones you are holding and put on the table' (CALD 2008).

[55] Cf. Hungarian **öcs** 'younger brother' vs **báty** 'older brother'.

[56] Or indeed, whether they exist as primary linguistic objects, or they are derivative entities, reducible to more essential ones, i.e. mere "epiphenomena".

A mere glimpse at the contents page of a few simple descriptive grammars of different languages, where parts of speech are elegantly and proudly listed, would convince one that their lists are different; moreover, different grammars of English also seem to diverge on this point. How come?

2.3.1 Descriptive grammars

Let us simplify a bit. English descriptive grammars differentiate about eight word classes. The first four seem to be uniform: *noun, verb, adjective and adverb*. These are usually equated with the *open classes*, which contain *content* or *lexical* words with *lexical meaning*. The second four (or more), the set of *closed class* words, comprised of *grammatical* or *function words*, which just have *grammatical meaning*, shows more variation. These usually include articles (or determiners), conjunctions, pronouns, and prepositions, sometimes interjections. This simplified picture, of course, is mainly seen in abridged grammars for (native speaker) students and pedagogical grammars (for learners).

The largest English descriptive grammar[57] distinguishes not two, but four superclasses:

(a) open classes – noun, full verb, adjective, adverb;
(b) closed classes – preposition, pronoun, determiner, conjunction, modal verb, primary verb[58];
(c) "lesser categories" – numerals and interjections;
(d) "words of unique functions" – the negative particle **not** and the infinitive **to**.

2.3.2 Criteria

It seems evident that there are apparently no principled criteria on which the above classification – and similar taxonomies – has been based. Let us see some of these issues.

- Is the property "open/closed" *more important* than "lexical/grammatical"? Or do "open/closed" and "lexical/grammatical" cover the *same* two territories? How can "lexical" and "grammatical" be *defined*?
- Why do the items in (d) not belong to the closed classes, when they are the *most closed* group?
- Why are prepositions among the closed classes, when a subtype of them, complex prepositions (e.g. **due to**, **with reference to**), which are recognized by this grammar, tend to increase in number? And why do they belong to the *grammatical* classes, when most of them have clearly definable, lexical content? Or can it be that there are *some* prepositions belonging to one, and *some* to the other set? Then two classes of prepositions ought to be set up, not just one. Note that prepositions give a lot of headache to the analyst.
- The verb situation may also remind us of the general issue of *classes* vs *subclasses:* in many classifications (though not in the one above), there is a huge class of verbs uniting such diverse things as lexical verbs, aspectual auxiliaries, and modal auxiliaries.

[57] Quirk et al (1985).
[58] These are **be**, "auxiliary" **have**, and **do**.

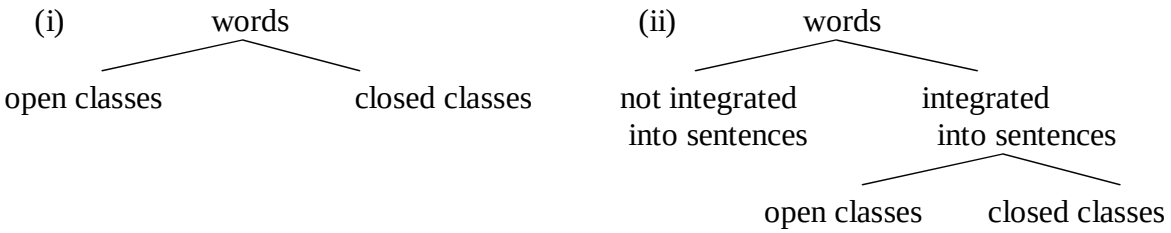
- If these three kinds of "verb" are more *different* than similar, then we should *not lump* them together, but have, for example, "lexicals" and "aspectuals" and "modals".
- If countable nouns are *more different* than similar to uncountable nouns, they should not be classed together either.
- And most generally: what is the basis – and is there *one common* such basis – of classing any items together? This cannot be *meaning*, for reasons well known, and it cannot be *form* in all cases, since many of these have *no morphology*. The only criterion is *distribution*, i.e. *behaviour in larger units*. This means that an exhaustive characterization of all words of a language must be complete before any such list is undertaken.

  "Behaviour in larger units" implies, first of all, that words that do not occur in larger units, i.e. clauses, should be classed apart from the rest that do; the first binary cut in our classification thus ought to be exactly this. The two grammars whose PoS (= part of speech) classifications are summarized below do just that; the first one tacitly, the second expressly.

2.3.3 Two recent approaches: many problems solved, many created

Modern theoretical approaches use strict distributional criteria for setting up word classes. If the notion of word class is based on behaviour in sentences, then only those words will be granted PoS status that are *integrated into sentences*. Most of what are traditionally considered as *interjections*, *e.g.* – **hey! / ouch! / shit! / morning! / bless!** – will either (i) not be PoS-classified at all, or (ii) be in a separate class from the rest, yielding systems like these:

Fig 7



2.3.3.1 Newson et al (2006)

Newson et al (2006), e.g. follows (i) above. Since it professedly places words in crisp categories by using the three features [N], [V], and [θ] (for *thematic*),[59] plus introducing *underspecification* for the [θ] feature, resulting in a total of 12 word classes, this system ignores huge numbers of words that cannot at all be characterized as having these features. Although this is a well-motivated move that gets the theory rid of many unwanted details, it does not make those details non-existent. The system has four *functional* (= non-thematic) and four *thematic* categories, plus 4 types of word *underspecified* for [θ], i.e. neither [+θ] nor [–θ]:

[–θ]=[+F]	underspecified for [F]	[+θ]= [–F]
① *D*eterminers	⑤ measure or group nouns	⑨ *N*ouns
② *I*nflections	⑥ aspectual Auxx; passive Aux	⑩ *V*erbs
③ *Deg*ree words	⑦ "postdeterminers",[60]	⑪ *A*djectives
④ *C*omplementizers	⑧ [–θ/–F] prepositions	⑫ *P*repositions[61]

Rather than some hard-to-define "lexical" property that marks "lexical" (as opposed to "grammatical") words, this system uses potential *predicatehood*, i.e. being able to be a predicate, as a criterion for *thematicity*. Any PoS that does not contain potential predicates is a functional class: this puts prepositions with the thematic, "lexical" group, and thus decides the issue of whether they are "lexical" or "grammatical". A preposition may be a two-argument predicate, most typically with two DP arguments, e.g. the logical formula of the sentence **The picture is on the wall** is *on (picture, wall)*. Unfortunately, as we will see, this does not at all remove the complex problems of prepositions.

2.3.3.1.1 Functional classes

① Determiners include the articles, the traditional demonstratives, certain of the quantifiers, some **wh**-words, and personal pronouns. Personal pronouns, which are either traditionally subsumed under nouns or get a class of their own, are D's here, *as required by the entire theory*. Determiners, i.e. items in ① may occur with items in ⑤ *or* with items in ⑨ but not with both: the phrase **a group** (①+⑤) is possible, **this tourist** or **some tea** (①+⑨) is possible, but ***this group tourists** (①+⑤+⑨) is an impossible sequence. This is because the words in ⑤ not being able to take nominal complements in English,[62] **of**, an "underspecified" preposition – an item of ⑧ – must be inserted between ⑤ and ⑨, yielding the sequence ①+⑤+⑧+⑨.

While the expletive **it** may be claimed to have its place among pronouns (a not too reassuring place, because its distribution is markedly different from any of the other pronouns, including the genuine neuter "personal" pronouns), expletive (existential) **there** has no label *as a word class*. Various other members of the traditional pronoun class also have no place within this classification; they are certainly not to be sought among the determiners.

② Inflections are the functional relatives of verbs. Note that in this approach, while inflections are a *word* class, not all inflections are *words:* I is a syntactic class that contains less-than-word constituents. Recall that half of the finite ones (i.e. the modals) are words, and the nonfinite inflection (**to**) is a word[63]; the rest of the finite I's are affixes.

The status of this I, and especially of infinitival **to** as a [+V] element is even more theory-dependent. Recall that in Quirk et al (1985) this **to** belonged to the "words of unique functions" (along with negative **not**, which has nothing to do with the word **to**, its strange bedfellow) and incidentally has no place in this system of 12 above).

Also missing from this analysis are several oddly behaving auxiliary-like items such as (a) **used to** and **ought "to"**, (b) similarly maverick ones with dual class membership, such as **need** and **dare**, and (c) many of intermediate status, e.g. the semi-auxiliaries (**appear to**, **be about to**, **be going to**, **be *certain* to, be *likely* to**, **be to**, **have to**, **tend to**; **keep -ing**). Some of these resurface under raising adjectives (italicized here), while some just have no place in this PoS scheme.

[59] Or, alternatively [F] (= for Functional).

[60] These are not determiners, and have adjectival features.

[61] Only the classes with an italicized initial letter have convenient abbreviations.

[62] In Hungarian, e.g., the same sequences are grammatical: **egy csoport turista** 'a group *of* tourists'; **egy pohár tej** 'a glass *of* milk'.

[63] Note that in a more refined system, **to** is not an independent word, just a clitic. The modals, by contrast, enjoy greater autonomy, e.g. they do carry stress.

③ Degree words belong to *adverbs* in traditional classifications, along with countless completely different other adverbs. One might think that (all, or most) adverbs have their secure place within Adjectives (⑪) here, but this not so: the most frequent place/time adverbs (**here**, **now**, **then**), which are not formed with **-ly**, have no place in this scheme.

④ Complementizers are one of the smallest classes. All its members have homonyms in another class: (a) complementizer **that**$_1$ is also a determiner (**that**$_2$) in ①; (b) **for**$_1$ is also a preposition (**for**$_2$)[64] in ⑫ ; and (c) **if**$_1$ *would be* a conjunction (**if**$_3$) but conjunctions (connectors) are absent from this system. The odd property displayed by these complementizers and their homonyms is that while

(a) **that**$_1$ originates from **that**$_2$ historically, and

(b) **for**$_1$ is actually called a *prepositional complementizer*, suggesting that it *is* a preposition,

(c) **if**$_1$ and **if**$_2$ show no overlap at all in terms of function. The relationship, then, between these C's and their homonyms, is not uniform.

2.3.3.1.2 Underspecified classes

⑤ The [–θ/–F] noun class of measure or group nouns is a conspicuously large one for non-thematic items, since practically any noun with a suitable meaning (i.e., one that permits it to take part in partitive structures) belongs here. Outside of these partitives, these nouns are thematic in that they can be predicates (although they rarely take arguments, if at all). Within the *partitive structures*, they are not predicates, of course.[65] Although there are huge numbers of nouns that have an individuating function in partitives (some of which are idiomatically chosen, cf. **a loaf of bread**, **a leaf of grass**), it would not make sense to talk about homonymy here claiming that there exists a separate [+θ] N **loaf** and a separate [–θ/–F] **loaf**, and a [θ] N **leaf** along with a [–θ/–F] **leaf**, and so forth for every N that has a double function like these. Such a *homonym analysis* would imply, for example, that while **this is a *leaf*** contains a [θ] **leaf**, the expression **this is a *leaf* of grass** has a [–θ/–F] **leaf**. There is a functional difference between them, but it is not a good idea to capture this in terms of different *word classes.*

⑥ The [–θ/–F] verb class contains perfect HAVE, progressive BE, and passive BE. It is to be noted that these are also *intermediate* not just in terms of *status* but also in terms of their *place:* both the **may have known** and the **to have known** type of "verb groups", e.g., are ②+⑥+⑩.

⑦ The [–θ/–F] items in the "postdeterminer" class have adjectival traits but non-thematic (although some of them can, in formal style, be predicates, e.g. **his faults were many/few**). If only **many**, **few** and **several** belong here, their number is small indeed. Traditionally, postdeterminers include *cardinal and ordinal numerals*; these, however, have no explicitly recognized niche in this system; they are probably best grouped with **many/few**. The traditional "open class quantifiers", customarily also classed with postdeterminers (**a large number of**, **plenty of**, **lots of** etc), also belong here; in this system, they are probably treated as nominal phrases containing a PP. A minor problem is that **many** when used before an article, in a completely different distribution (e.g. **many a day/man**) is not accounted for.

⑧ The underspecified P's have no semantic label, so it is difficult to guess the size of this class. Newson et al (2006) only offers two examples: **of** and **by**.

One of the [–θ/–F] prepositions mentioned – the word **by** – is probably a homonym of the thematic (e.g. spatial) preposition **by**; the item **of**, however, has only a grammatical function (i.e. is only functional). If, however, non-thematicity is all we can state about these prepositions, then *any* complement *P*, i.e. a P headed by a complement PP will be a [–θ/–F] preposition, as the italicized P's in the PP's **keen** [$_{PP}$ ***on*** **music**], **differ** [$_{PP}$ ***in*** **size**], **give the book** [$_{PP}$ ***to*** **her**] etc. It is questionable, though, whether it makes sense to talk about homonymy in all such cases, and claim that there is one [–F] P **on** and one [–F/–θ] **on**, a separate [–F] P **in** and a separate [–F/–θ] **in**, or a [–F] P **to** and a [–F/–θ] **to**, and so on for the many prepositions that double up like these.

It is thus not accidental that underspecified P's have no semantic label: they have no independent meaning, their choice depending on some syntactic head requiring them. Biber et al (2007) calls these *bound* prepositions (**decide *on* the screen**), as opposed to *free* ones (**a fly *on* the screen**).

[64] There is actually a third **for**, the causal conjunction, which has no place either.

[65] The complements of partitive nouns are not arguments, and thus they are not in a thematic relationship with them: e.g. **a box of chocolates**.

2.3.3.1.3 Thematic classes

⑨ Nouns are a unified class, not differentiated by Count/Noncount at word class level.[66]

⑩ What are traditionally verbs are not a unified class here: they are in three classes, grouped according to distribution. Only [+θ] verbs belong here. As noted above, the three kinds of [+V] item also may follow each other *topographically*: **may have/be known** and **to have/be known** e.g. are both ②+⑥+⑩.

⑪Under this framework, adjectives as a class include adverbs[67], and the two are supposed to be just variants in complementary distribution. Because, however, adverbs have always been the largest catch-all class, inevitably many different items in that category have no place in this system either. All non-**ly** adverbs (including the extremely productive **-wise** formation) are like this, as well as connecting elements. Sentence adverbs, or disjuncts, whether truly adverbs or not, are not placed anywhere as a word class.

⑫ As in all other approaches, prepositions are the most recalcitrant class in this system as well, whether we consider the [–θ] or [+θ] or [–θ/–F] type of P:

(a) Of the three complementizers, only **for** is a homonym of a genuine P. The homonym of **that** is a determiner and has nothing to with prepositions. The homonym of **if** is a "conditional conjunction", which has no word class in this arrangement.
Complementizers, then, have a very tenuous connection to prepositions.

(b) there is a problem, as we have seen, with the definition of the [–θ/–F] ones.

(c) The different kinds of P never occur in a string, neither all three, nor any two: there never occurs a ④+⑧+⑫ (or ④+⑧ or ⑧+⑫) sequence. This is not a problem, but it is conspicuous that [+N, –V], [+V, –N] and [+N, +V] strings (e.g. **a dog**; **may feed; many small**) like that are all possible.

The fact that new prepositions do systematically turn up (either through a conversion-like process, like in the case of **regarding, following** and **pending** or through combination, as in the case of **on account of**, **by means of**, **with regard to**) is in harmony with prepositions being an open class. Unfortunately, however, the present approach does not mention them, and it may be surmised that some of them would not be considered a (complex) preposition at all but a PP. Of course, whether a multiword item is entitled to a PoS label is an even more vexing general question.

The problems discussed in 2.3 highlight the fact that for word classes to be set up, hosts of questions need to be decided on, which will all depend on the assumptions about the grammar, i.e. on a particular theory. A reliable list of word classes of a language can be hoped to be provided only on the basis of a whole theory of the grammar of that language. It is expected that any such list will differ.

2.3.3.2 Inserts in Biber et al (2007)

Biber et al (2007) follows (ii) in Fig 7 above. The most important novelty of their PoS arrangement is that it distinguishes three large groups, the third one being new: *inserts*. The three *superclasses* here are *lexical words*; *function words*; *inserts*. This new niche takes care of a huge number of lexical items that clearly have no place otherwise/elsewhere.

Lexical words are nouns, verbs, adjectives and adverbs, in the spirit of most traditional classifications. As expected, function words are both more numerous and more varied, since the lexical superclass only accommodates four classes. Inserts, however, while clearly solving a classification problem, create another: they, too, are far too heterogeneous.

Inserts are a newly recognized category of *marginal* types of word that are not an integral part of sentence structure. At least a part of them may be thought of as an extended class of interjections, such as **cheers**! **bye! hi! hey! erm… uhm… hm… ouch! shit!** However, even the traditional interjections have always been much too heterogeneous, and inserts are even more so: some of them are badly or not at all integrated into sentences, while others are completely *outside* clause

[66] They are also not grouped into Common vs Proper because these do not display genuine distributional differences just maybe divergences of *typical use*.

[67] Or there is a superclass "A" uniting adjectives and adverbs, nicknamed "advectives" in Radford (1988).

structure, being *"sentence words"*[68]. The "unintegrated" nature of inserts – again, of just a minority of inserts – is also manifested in their deviant phonological structure: they may have sounds and sequences not existing in *integrated* words. In that sense, such words – e.g. **ugh** (with its /χ/ sound of **loch**), **tut-tut** (with its two clicks), **yeah** (ending in a lax vowel), **shhhh** (containing no vowel) are not part of the language at all.

While it is a useful idea to separate inserts from the integrated words, apparently inserts are still much too heterogeneous, even to the extent that many are not words but multiword items, or syntactic phrases, e.g. NP's or clauses. It seems obvious, however, that if the class of inserts is set up on a *pragmatic/functional* basis, then its members will be varied from the *formal* point of view – and vice versa.

If, based on their function/distribution, we class the items **hell!** / **shit! / rubbish!** and the like with inserts, then we have created a new class of homonyms, this time between nouns and inserts. It there are too many such pairs, then it is questionable whether it makes sense to talk about a **shit**$_1$ and **shit**$_2$ or **rubbish**$_1$ and **rubbish**$_2$. The expressions **bloody *hell*!** and **holy *shit*!** and the likes of them cause no such problem, but it is questionable whether they deserve *word class* labels since they are phrases. Some inserts are multiword sequences, but not obvious phrases, e.g. **oh dear!** or **bye-bye** or **good-bye**. The items **you know**, **excuse me**, **I mean** are all clausal; one wonders whether clauses also deserve PoS labels.

It would not be fair, however, to hold it against inserts, a *superclass*, their heterogeneity, since as we have seen, even word *classes* have *core/central* and *peripheral* members: categories are not homogeneous. This realization about categorization is to be thanked to *cognitive* approaches to linguistics.

2.3.3.2.1 Multiple membership vs fuzziness

Fuzzy borderlines are to be distinguished from homonymy, where a single form belongs to more than one word class. (In the discussion above, homonymy was mentioned several times.) Since English is impoverished in terms of morphology, this is a very common phenomenon. Sometimes this is captured in terms of conversion, implying a dynamic process of some X becoming Y, but it need not be.

A form such as **right** can be a lexical word (noun, verb, adjective, adverb) or an insert.[69] The following pairs are clearly distinguishable in PoS terms: ***early* arrival** – **arrive *early***; **a *fight*** – **to *fight***; ***narrow* street** – ***narrow* the focus**; so are the five copies of **round**: ***round* of applause** – ***round* face** – ***rounded* the corner** – ***round* the corner** – **turn it *round***.

Traditionally, **before** (along with **after**) belongs to three classes, preposition, adverb, and conjunction. They do have different distributions, which favours the traditional separate treatment; at the same time, *simplicity* of the system is gained is they are lumped together under preposition, the differences being just in terms of complementation. (The former preposition **before** has DP complements, the adverb has no complement, and the conjunction, clausal complements.)

So, with items like **before** it is not even clear whether multiple membership is involved, but fuzziness there is certainly none.

The ideal word classes need to meet two requirements. They should be *large and general* enough to yield useful generalizations, and at the same time *small and fine-tuned* enough to be true to all, or at least most of, the facts. As the properties of a class may vary, there are unclear borderlines between the characteristics of one class and another. The flexibility of language may thus defy any classification system. If distribution, i.e. syntactic behaviour in all its subtleties is the sole basis for word classes – and this is taken seriously – then the number of PoS's will be far greater, and borderline cases will be much more numerous, than suggested by any of the standard crisp and black-and-white classifications. It is to be feared that you can't eat your cake and have it.

[68] Not much used in English linguistics, it is the equivalent of German *Satzwort* and Hungarian *mondatszó*, which are quite general terms.

[69] If there is no separate *adverb* class, and/or *inserts* are not recognized, this will of course be slightly different.

2.4 Above the word

One would think it impossible not to be able to tell a phrase from an affix: their properties are surely very different. To the lay person, the distinction between an "ending" and a word is just as obvious. We have seen, however, that this is by far not so: there are distinct phenomena between the level of "endings" and words. Just as there are dependent words and semiwords between genuine (autonomous) *words* and *affixes*,[70] there also seems to be a grey area between word and phrase.[71]

Affixes — semiwords — dependent words — autonomous words — ??????? — phrases[72]

What kinds of element populate this area, and (if it is not a *continuum* or *gradience* that lies between words and phrases) how many more or less discrete categories may be set up here (*discovered*, remember, not *invented*)? This will be the topic of this section. To put it succinctly: where do words end, and where do phrases begin?

Earlier we suggested that several types of *multiword unit* exist, and offered a preview, without labels attached:

(A) **mousetrap**, **flash drive**, **download, shut down**
(B) **give *somebody a bell***; **take *advantage of something*; walk *down the aisle***
(C) **If it ain't broke, don't fix it**; **Slow and steady wins the day**

to which we now add three:

(D) ***flog* a dead horse; *kick* the bucket; a *feather* in smb's cap**[73]**; green-eyed *monster***
(E) **the whole caboodle**[74]**; to and fro**[75]**; if I had my druthers**[76]
(F) **no man's land**

There is probably consensus on the items in (B) being *phrasal*: they do have a *verbal head* and some structures that are *arguments* or *adjuncts* (italicized). The ones in (D) and (E) are also phrasal; the items in (C) are more than simply phrasal; they are *sentences*. The sequences in (A) are more problematic: they may all be argued to be multiword in one sense or other; they may also be argued to be *compound* words. The items in (D) and (E) are special for other reasons as well.

When is an apparently multiword sequence a word, and when is it a phrase? The sequence of morphemes **hard disk** is a compound, but is **hard drinker** – an Adj + Noun sequence – also one, or a phrase? Several criteria come in handy; some more reliable than others; some contradictory; none foolproof.

---

[70] Recall, though, that the three elements above the level of affixes are termed (kinds of) *word*.
[71] To chart the grey area between affix and word, we didn't have to use data from other languages (except Hungarian). No mention was made, e.g. of what is termed suffixed definite article, arguably an enclitic (as in Norwegian **hus*et*** 'the house' or Romanian **lup*ul*** 'the wolf'.
[72] Although clearly not the task of *morphology*, what kinds of elements are above the phrase is a legitimate question from *lexicology's* point of view.
[73] 'A praiseworthy accomplishment; distinction; honor'; RHWUD (1999).
[74] 'The whole lot, pack, or crowd'; adapted from RHWUD (1999).
[75] 'Alternating from one place to another; back and forth'; RHWUD (1999).
[76] RHWUD (1999) defines **druthers** /ˈdrʌðəz/ as 'one's own way, choice, or preference': **If I had my druthers, I'd dance all night.**

2.4.1 Phrasal verbs

We will say, first of all, that **shut down** is not a compound – it is not one *word* in the first place – because its two parts can be *separated*. It cannot be called a compound *verb* either. The existence of **shut *the system* down** or **shut *it* down** shows that syntactically, this is two words: they can be separated in/by the syntax. The same separation is impossible with **download: *down the file load** or **down it load* are ungrammatical. The sequence **shut down** is called a phrasal verb, which falsely suggests that it is one word; the point, however, in having *"phrasal verb"* is not that it is a special kind of (i.e. multiword) *verb*, but that it is *phrasal*, i.e. is a phrase. It behaves like Hungarian *prefixed verbs* such as **lekapcsol** 'shut down': negation (a syntactic operation) splits them: **nem kapcsol le** 'does not shut down'. Phrasal verbs[77] are items of the lexicon; they are lexical elements, but not words, so *not lexemes* either. Not all lexical items are words – this much we know now.

2.4.2 Compounds

The items **mousetrap** and **flash drive** are compound nouns; the **download** type (cf. **input**, **outrank**, **uphold**, etc) is a special compound verb. We will call **shut down** a *phrasal* verb of the "verb–particle" structure, while **download** a *compound* verb of the "particle–verb" form.

(Whether compounds are written *solid* "as one word"), or hyphenated, or *open* (separately) is irrelevant. There do seem to be more and less widespread varieties for individual words, though, and this makes it difficult to accept the irrelevance of the spelling convention for analysis.

The argument for compound status that is usually used is twofold: (a) *prosodic* and (b) syntactic.

(a) Their *main stress* is on the first member, i.e. they have early stress (*unit stress*), unlike in phrases, which have *phrasal* (i.e. *late*) *stress*. This is a fairly, though not 100% reliable diagnostic of compoundhood.

(b) They are words, which means that they have what is called *lexical integrity*: no syntactic rules can apply to their parts. Their members cannot be separated; they cannot be targets of questions; they cannot normally be anaphorically referred to.

(b-i) You can say e.g. **It's a trap but not for mice**, but not ***It's a trap but not [mouse___]**
– where just the second member of the word has been *omitted/ellipted*.

(b-ii) You can say
I trap [mice] with this gadget — ***[What] do you trap ___ with this gadget*?**
but not
This gadget is a mousetrap — ***[What] is this gadget a [___trap]?***
– where just the first member of the compound has been targeted by the question.

(b-iii) Many speakers do not accept sentences where pronouns refer back (i.e. *anaphorically*) to just a part of a compound (italicized here):
?She's bought a [*mouse*trap] because she's afraid of them
?He was looking for a [*book*rack] but he only found racks for very small ones

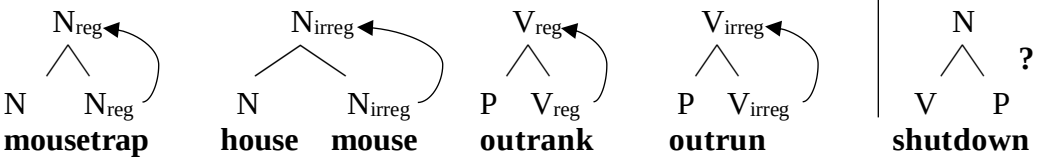
[77] And prepositional verbs (e.g. **look into** 'examine' or **go with** 'choose'), and phrasal-prepositional verbs (e.g. **look down on** 'despise' or **put up with** 'tolerate') are also not single-word verbs, i.e., not words, i.e., not lexemes.
"Phrasal verb" can actually be used inclusively, for (i) this narrow sense of "phrasal verb" *plus* (ii) prepositional verb *plus* (iii) phrasal-prepositional verb.

2.4.2.1 Phrasal verb to noun conversion

Though phrasal verbs are not words syntactically, and consequently not compounds either, there does exist a significant class of words – nouns *converted* from these verbs – which *are* compounds: SHUTDOWN and DOWNLOAD, e.g. are compound noun lexemes. (Note that here, the stressing also changes: **shut** ˈ**down** > ˈ**shutdown** – the latter is the typical compound *stress pattern*.) They are special because they are nouns, although their right-hand member is not one, thus violating the *right-hand-head rule* for English compounds which says that a compound word gets its grammatical features, including its word class and its [±regularity] feature, from the head, the right-hand member. Thus, MOUSETRAP is a (regular) noun since TRAP is one; HOUSE MOUSE is an irregular-plural noun, with its irregularity *percolating* from MOUSE. The compound OUTRANK is a regular verb because RANK is regular, while OUTRUN is irregular because RUN is irregular. See Fig 8.

However, the right-hand member of these converted nouns – SHUTDOWN, BUILDUP, COMEDOWN, COUNTDOWN, DROPOUT, SETUP, HANDOUT etc – is a preposition.[78]

Fig 8



2.4.2.2 Compounds: a right-hand-rule mystery

What exactly gets violated in the SHUTDOWN kind of converted compound is a tricky question, though. Two solutions are possible:

(i) these compounds are left-headed – and their grammatical properties percolate, i.e. *filter* up to the compound word from this left-hand member. (This requires, however, that the V on the left be converted into a N first.)

(ii) they are *headless*, and their word-class status is not "inherited" from inside the compound; the question, then, is where it *does* come from.

It must be stressed that these converted phrasal verbs are not the only instances of left-headed compounds in English (and especially not universally). If **princess royal**, **governor general**, **present perfect** and similar items are analyzed as compounds (rather than phrases – in which case, incidentally, this noun–adjective pattern would also be unusual), then they are left-headed as well.

2.4.2.3 Headless compounds: Shakespeare was no humpback

Syntactic objects must have a head, which may or may not be a *structural head* and a *semantic head* at the same time.[79] By contrast, compounds may or may not have structural heads; if they do, these heads may or may not be semantic heads as well. If they have no head, they are called *exocentric*. Two types will be exemplified.

(i) Noun compounds of the type PICKPOCKET 'person who steals from pockets/bags' or TURNKEY 'jailer', which are truly exocentric: though they are nouns, and their right-hand member is also one (or the pronoun **all**, as in CURE-ALL[80]), but it is not their head *either structurally*

[78] In more traditional frameworks, classified as Adverb or as Adverb(ial) Particle.
[79] Thematic heads are heads in both senses; functional heads are not semantic heads.
[80] And also CATCH-ALL, CARRY-ALL, HOLD-ALL.

or semantically. The plural of **cure-all** is **cure-alls**, but this form cannot come from the plural of **all** because it has none. A PICKPOCKET is not a kind of pocket, a TURNKEY is not a kind of key. (These oppose the *endocentric*, *headed* compounds like **shirt pocket** and **car key**, where the first member is a modifier of the second, and thus the compound itself a *hyponym* of the second member.) These exceptional compounds do not inherit their features from their head, in which respect they are like SHUTDOWN above.

Unlike in SHUTDOWN, however, and quite interestingly, in these compounds the second member is an argument – object – of the first, which is a verb: CEASEFIRE, KILLJOY, SCARECROW, SPOILSPORT and TURNKEY are better-known examples: a CEASEFIRE is about the ceasing of fire, a KILLJOY is someone who "kills" others' joy.

And of course **Shakespeare** belongs here: this name is the combination of **shake** and **spear**, and it meant 'spearman'.[81]

(ii) The second type of truly exocentric compound will be illustrated with HUMPBACK 'person with abnormal curve of the spine'. This is a noun; its right-hand member is also a noun, but this noun is not the head in either sense; this kind of compound has no semantic head modified by the non-head. A humpback is not a kind of back.

The fact that the right-hand member of these compounds is not their structural head is seen in the plural of FLATFOOT. This noun may have an irregular plural, but it can also be **flatfoots**, indicating that the number feature does not percolate from the right-hand member, **feet**.

These compounds are instances of *metonymy* (a kind of *pars pro toto* relationship where *a feature* represents the entire person), the underlying idea being that a BLUEBEARD is a person who *has* a blue beard. Other better-known examples include BLOCKHEAD 'silly person', REDNECK **'**uneducated farm worker', BIGMOUTH **'**a noisy, indiscreet, or boastful person', HIGHBROW 'a person of scholarly and erudite tastes', and EGGHEAD 'an intellectual; highbrow'.

All the items in (A) – (F) in 2.4 above are *lexical items*. Those in (A) are *lexemes*, since they are words (i.e., lexically integral objects, pronounceable on their own, manipulated as units by the syntax). The items in (B) – (F) are not lexemes. The *lexical item* status means, among other things, the psycholinguistically relevant fact that, for one reason or other, these items need to be memorized – remembered, *stored and retrieved* from memory – as wholes. From a semantic point of view, these items are *form–meaning pairings*, similarly to words: it is the *whole* of the verb phrase **walk down the aisle** that means 'get married', and it is the whole nominal phrase **green-eyed monster** that means 'jealousy'. Just by knowing SHUT, you will not know **shut down** (not even that it exists); just because you know everything about the noun BELL, you won't know **give smb a bell** (not even that it exists); knowing BROKE$_{(adj)}$ and FIX$_{(v)}$ will not empower you to say **if it ain't broke, don't fix it**; knowledge of KICK and BUCKET will not enable you to use the *idiom* **kick the bucket**. The expression **no man's land**, which you will recall from earlier, is special because it looks like a normal (genitive) phrase such as **George's car** or **mom's cooking**, but it might remind you of a compound, and may be argued to be just one concept rather than two, as the latter two expressions. **No man's land** (p. 12), then, illustrates the "grey area" between (compound) word and phrase.

[81] Similarly formed surnames are **Shacklock** or **Shakelock**, a 'lock-shaker', i.e. a jailer and; **Shakelance** meaning 'lance-shaker', i.e. 'spearman'.

2.4.3 Idioms

Group (D) contains what are traditionally termed idioms, e.g. **flog a dead horse; kick the bucket**. The most noticeable thing about them – part of their well-known definition – is that their meaning is not the sum total of, i.e. not computable/derivable from, the meanings of their parts. (An expression whose meaning can be computed from the meanings of its constituents and the rules of combination is *compositional*.) All of their component parts, if considered in isolation (i.e. *outside of* these expressions), have some lexical (or grammatical) meaning, but that is irrelevant: it is as though the expression were bracketed after it has been formed out of words, and once it is between brackets, the inside no longer exists for the outside.

It seems that people have a perfectly clear *ideal* of idioms: *prototypical* idioms are those "colourful" ones that are presented and taught to language learners as interesting and useful – and are mostly useless, especially in the quantities they get dumped on the learner. In actual fact, there are many, many more idioms than this over-hyped kind.

A large percentage of *compounds* are *idiomatic*, i.e. non-compositional in terms of meaning. Although they are not classed with idioms, idiomaticity is a feature that compounds have in common with idioms.

Group (E) also lists idioms, but these are special even within that group: they have component parts that do not exist outside of these expressions: **the whole *caboodle***; **to and *fro***; **if I had my *druthers*** are such examples. This may help understand that idioms are form–meaning pairings *as wholes*.

So even if **shut down** and **walk down the aisle** and **flog a dead horse** are not lexemes, they are lexical items. Another angle from which they can be looked at is the mental lexicon, the "word stock" in the native speakers' head. That is when the term *listeme* will come in handy.

2.4.3.1 A few home truths about idioms

Most of what is claimed in this section is true for other languages too, but we will stick to English.[82]

The string **home truth**[83] is an idiom because its meaning cannot be *calculated* (or *deduced*, *guessed*, *predicted*) from the meaning of its *components* (or *constituents*, or *members*). More liberal definitions, however, have cannot "*fully* be calculated" rather than just "be calculated". This means that idiomaticity, as defined by this *semantic opacity* (or *non-transparency*), is a *gradual* phenomenon, i.e. there is a *cline* (or *scale*, *spectrum*, *gradience*) of *idiomaticity*.

For us, an idiom will be a "*multiword* lexical item whose meaning can't be *fully* deduced from that of its constituents". Note that under this definition, a complex *single-word* item whose meaning is not predictable from the meaning of its parts is not an idiom, although under a broader definition it would qualify as one. E.g. the compounds DOORMAN 'person whose job is to stand by the door of a hotel or public building' and FOOTMAN 'male servant whose job

---

[82] These sections on idioms are based on Ayto (2006) and Moon (2006).

[83] CALD (2008) explains it as follows: 'a piece of information which is not pleasant/wanted, but is true'. MED (2008) only lists the plural **home truths** and defines it as 'unpleasant facts/opinions about *you* that someone tells you'. LDOCE has this: **home truth** [countable usually plural] 'a true but unpleasant fact that someone tells you about *yourself*'. RHWUD (1999) has **home truth** defined as 'an indisputable fact or basic truth, esp. one whose accuracy may cause discomfort/embarrassment'.
Note that RHWUD's and CALD's are the broadest definitions, because they do not include the 'about yourself' component. In fact, it is about a *general* truth rather than a truth that may embarrass *you*. Under the MED and the LDOCE definitions, this could never be a section title here, since the (supposedly new) information is not about you but idioms.

includes serving food' are not idioms, even though their meanings are idiomatic; you cannot guess/predict what they mean. Just as idioms show a *scale* of idiomaticity, the idiomaticity of compounds is also gradual, and – unlike that of idioms, which are by definition idiomatic – it ranges from fully idiomatic to non-idiomatic. LADYBIRD, BLUEBOTTLE and TALLBOY[84] are at one extreme, BLACKBIRD, CHATTERBOX and LADYKILLER are in the middle of this continuum, while BEDROOM, RAINCOAT and SANDPIT are at (or near) the non-idiomatic end.

This general definition covers different expressions that, in addition to being used in different *syntactic functions* (from different word classes to sentences), are located at intersecting *gradiences of (a) semantic opacity and (b) grammatical fixity* (or *fixedness*, or *frozenness*). Idioms, then, are situated along these two (not necessarily connected) *spectra*.

(a) Opacity

– At one end of the *semantic opacity spectrum* the idiom is totally *opaque:* many of the most-quoted *figurative* idioms such as **kick the bucket**, **rain cats and dogs**, **eat crow**, **lend smb a hand**, **eat humble pie**, **cut the mustard**, **pig in a poke** belong here. They may be opaque because they contain words with a meaning that is not (generally) used/known outside the idiom. To most speakers, the word **poke** e.g. does not mean 'sack'[85] outside of this idiom. This is a different situation from e.g. **cahoots** (in **be in cahoots with smb over smth**[86] and the examples in 2.5.1 which discusses an even more special case, of words simply not existing outside of certain idioms).

– When all the main elements of an idiom have their standard meaning, it is their combination – the syntactic structure – that is responsible for the opacity. This is why **fish and chips**[87] and **bread and butter**[88] in the sense 'bread spread with butter' are idioms: these two coordinated phrases do not mean simply what the **and** suggests, cf. **I bought fish, and chips and onions**. Note, incidentally, that **bread and butter** is also an opaque idiom when it means 'source of livelihood'. (So there are three strings of the form **bread and butter**: a free phrase; a relatively transparent idiom; and an opaque idiom.)

The closer to the opaque end of the gradience a multiword expression is, the more likely it is to be regarded as a fully-fledged idiom. Many compounds (**pickup** 'an increase/improvement'), as we have seen, and many multiword verbs (**pick up** 'start a sexual relationship with someone you don't know'; **let up** 'of rain: stop or improve') also satisfy the criterion of semantic opacity.

(b) Fixedness

Idioms are often *fixed phrases:* this means that they tolerate no *manipulation* either (i) *lexically* or (ii) *syntactically*, i.e. they tolerate *no replacement* of

(i) any of their members with another item, and *no insertion* of lexical material;
(ii) any of their *structure* with another, *or addition* of other structures.

In fact idioms vary in this respect too – both in terms of (i) and (ii) – and so are best seen along a scale (degrees of) of *fixity/fixedness*.

---

[84] The meanings have not been supplied deliberately.
[85] RHWUD (1999) has this: **poke**[2] 1. *Chiefly Midland U.S. and Scot.* 'a bag or sack, esp. a small one'.
[86] CALD (2008): 'act together with others for an illegal/dishonest purpose'
[87] CALD (2008): 'fish covered with batter […] and then fried and served with pieces of fried potato'
[88] In Hungarian, **vajas kenyér**, with a different syntax. Cf. G. **Butterbrot**.

– At one end of the fixedness spectrum, idioms allow (of widely differing types and compositions[89]) no alteration or insertion: **call it a *week*** is not an acceptable variant of **call it a *day***, and **kick the *bucket*** cannot become **kick the *pail***, and **spill *the* beans** cannot be **spill *your* beans**.

By heart; **for my part**; **on paper**; **by and large**[90] (PP); **raise an eyebrow**; **make head or tail of it**; **stick one's neck out**; **make heavy weather of smth** (VP); **salad days**; **monkey business**; **big deal**; **no end**; **hook, line and sinker** (NP); **dyed-in-the-wool** (AP); **There's the rub**; **If a job's worth doing, it's worth doing well** (sentence) etc are all fixed.

– About the middle of the fixedness scale, some *lexical variation* is possible: e.g. alongside **one's eyes are bigger than one's *stomach*** the same idiom with ***tummy*** and ***belly*** is also around. ***Fit* the bill** alternates with ***fill* the bill**; **drag one's *feet*** with **drag one's *heels*** etc. In extreme cases, there is a whole cluster of different forms around a *metaphor:* **another *nail* in smb's *coffin*** / **a final *nail* in smb's coffin** / **to *nail* down smb's *coffin*** / **hammer the last *nail* into smb's *coffin*** / **drive the first *nail* into smb's *coffin*** / **bolt down smb's**[91] ***coffin* lid**. There is also American–British variation – e.g. **not see the forest**$_{US}$ **/ wood**$_{BR}$ **for the trees** – which is also responsible for some of the lexical variability.
On the other hand, rule-governed modification to structure is not just possible but inevitable. The most obvious variation is *inflection* (tense and person/number)*:* verbal idioms, e.g. typically have different tenses. Idioms (just like individual words) are associated with certain *grammatical structures:* **kick the bucket** has no progressive, while **push up the daisies** only occurs in the progressive.

Transitive verbal idioms, e.g. have a *vacant slot*, or *a variable*, for the (direct, indirect, or prepositional) object or some other functional item to be filled (italicized in the examples): **sweep *smb* off *their* feet; give *smb* a piece of *one's* mind**; **clap eyes *on smb***.

– At the "loosest", i.e. most liberal end of the fixedness gradience are found constructions such as **what is X doing Y**?[92] (where X is the subject, and Y a place expression). As in (a) above, when all the elements in an idiom are *variable slots*, then it is their syntactic structure that allows them to be called idioms, for this overall structure determines their meaning.

Again, the closer to the fixed end of the gradience some multiword expression is, the more likely it is to be regarded as a prototypical idiom. As expected, prototypicality in terms of (a) and (b) get added together, with the *most opaque* and the *most fixed* ones being the clearest exemplars of idiom.

2.4.3.2 Creative variation

In addition to the ordinary rule-governed, or systematic variation in idioms, there is ad hoc *creative (individual, deliberate) variation*. Such creativity occurs when e.g. in some legal setting, someone uses the idiom **leave no stone unturned** in an altered form: **leave no *legal* stone unturned**; this is simple insertion. Even more radical alterations are possible, such as when

[89] These are not just varied structurally, but also different in functional terms: **big deal**, a NP, generally occurs on its own as an interjection; **no end** is used adverbially; **monkey business** may be used in canonical subject/object slots.

[90] This idiom is also special because of its ungrammaticality: different expressions – an adjective (phrase) and a preposition (phrase) – have been coordinated. Cf. also **down and dirty** (P and A coordinated); **in the know** (the word **know** is only a verb here in this idiom).

[91] Allowing for suicide, these all work not just with **smb's** but also with **one's**, i.e. *one's own*, coffin

[92] Translation also shows that this is an idiom. Hungarian has **mit *keres*…** 'what is … *searching*… . German has **was hat … hier zu *suchen*?** 'what business does … have here', literally also 'what is … *searching*?'.

someone says **There'll be no bucket-kicking** [i.e., dying] **here**. In the appropriate circumstances, the form **at the drop of a trilby** [= a kind of hat] may be a variant of **at the drop of a hat**. This is no different from the purposeful violation of any grammatical (morphological, syntactic, semantic) norm.

A typical, albeit hardly definitional, property attributed to idioms is that they are frequent, especially in the colloquial, spoken varieties. One consequence of this – apparently false – premise is that idioms are not just fun but actually ought to be useful in language teaching.

2.4.3.3 Corpus evidence: frequency and variability

In this section frequency and variation will be in focus. Corpus data can be used to demonstrate that many English *figurative* or *colourful* idioms (e.g. **bury the hatchet**, **red herring**), and most proverbs and similes (a large portion of which are idiomatic; e.g. **you can't have your cake and eat it**; **white as sheet**, **thin as a rake**) occur infrequently. Findings do vary, but in huge *corpora* of English texts, very few of these figurative idioms, and no proverbs or similes, are found with high frequencies. By contrast, less "colourful" idioms such as **take place**, **in fact**, **come to think of it** and **give up** are extremely common, forming part of the central vocabulary. There is a visible paradox here: the reason why these colourful, figurative idioms are thought to be so frequent is exactly that they are *prominent:* they "stick out". Their *prominence*, however, is misleading since it comes from *markedness*, which results from *infrequency*.

Corpus data also prove that there is far greater *instability*, or *variability*, of these idioms that is usually thought. This is true for the rule-governed, systematic kind of variability, the normal lexical variation, and the *ad hoc* individual, creative variability as well. Examples have been given in 2.4.3.2.

2.4.3.4 Multiword lexical items vs collocations

Multiword sequences, whether phrasal verbs, prepositional verbs or idioms (or compounds, if they are to be treated multiword) have *semantic cohesion*. *Collocations*, by contrast, are combinations of *independent* words that typically appear together, i.e. co-occur. The adjectives BROAD and WIDE, e.g. are found in different collocations, although they are broadly (not widely!) synonymous.

broad	either	**wide**
accent	**selection**	**angle**
agreement	**spectrum**	**appeal**
daylight	**shoulders**	**area**
grin	**variety**	**distribution**
mind		**experience**
outline		**interests**
smile		**margin**
support		**sidewalk**

Thus, **broad accent** and **wide area** are good, possible, English-sounding (sometimes called "idiomatic") collocations, while ***wide accent** and ***wide mind** are not. Note, however, four things: (a) in some cases both adjectives *collocate* idiomatically (**broad/wide shoulders**); (b) there seem to be worse and less serious violations of what is collocationally possible; (c) because "what goes with what" must be memorized, it may make sense to talk about *listemes* in

(some of) these cases; (d) the collocation may actually be so strong (the cohesion so great) that it even may make sense to talk about *compounds*, e.g. in the case of **wide angle**. The point is that these expressions are normal nominal phrases, adjective–noun collocations, and not multi-word units in most cases. They have no unitary meaning as wholes; their meaning is always the sum total of their parts. If you feel that BROAD and WIDE are not quite synonymous – BROAD is more abstract – then it is not surprising that they also collocate differently.

2.4.4 Collocations

Why are collocations so difficult to get a handle on?

The availability of corpora had brought collocations into the limelight (back at the end of the 1990s). Because corpora allow us to see *patterns* that had been lying low invisibly before, access to corpus data led to some important realizations about "whether it is really the syntax that combines words". There does indeed seem to be an *open choice principle* at work here; it has, however, also become clear that there is an other one, the *idiom principle*.

The sheer frequency of **I love you** (as opposed to e.g. **I like those girls / foods**) and **all the time** (as opposed to e.g. **all the boys / books**) shows that these are *prefabricated* phrases. The words in them are not selected on the fly – are not chosen freely – but are probably retrieved from memory as wholes.

There are two extremes, then, with many positions in between.

IDIOMS	COLLOCATIONS	FREE PHRASES
– frozen/fixed expressions – extreme manifestations of the *Idiom Principle* – mng: non-compositional	– less rigidly fixed – the "twilight zone"	– manipulated by the syntax – extreme manifestations of the *Open Choice Principle* – mng: compositional
be at SIXES and sevens **at ALL** **not at ALL** **at a STONE'S THROW** **put all your EGGS in one basket** **leave no STONE unturned**	**kept my promise** **addled egg** **stark naked** **throw a party** **dead drunk** **pay attention**	**SIX boys / ALL the boys** **buy EGGS** **the PARTY last night** **THROW a STONE** **KEEP a rabbit** **be DRUNK/NAKED**

There is a hierarchy in collocations, so we can distinguish the *base* and the *collocator*; the base determines the other member. The bases have been underlined in the box above.

2.4.5 Lexical bundles

There is another type of expression, also studied by phraseology: these are expressions which are neither idioms, nor collocations, nor free phrases – because they are *not identifiable structural units*, i.e. not phrases of any kind – but which occur frequently (i.e. more frequently than would be justified by the frequency of their members separately). These expressions are *lexical bundles*. Examples include:

the truth is …
the thing is …
… if you wanna know
… (you) know that I mean
… you know what I'm saying – *"know whadd ahm sayn"*

One reason why collocations are so hard to pin down is gradience. Prototype theory, as we have seen, can usefully be applied to linguistics as well. Gradience is true of all linguistic objects/terminology; most terms capture a range of slightly different phenomena and are, therefore, best defined with reference to *prototypical* examples.

A preliminary definition: "a collocation is a combination of two or more words that occur next to each other".

That definition must be amended:
(a) Two, not more words are prototypical.
(b) The words are not always adjacent: in the type **kept my promise** e.g. the base and the collocator are not adjacent.
(c) The phrasing "occur next to each other" is obviously not enough: the words have to occur not by mere chance but need to be frequently used in that combination. That is, we must distinguish *collocation* from just any old *combination* and *co-occurrence*.

If "frequency" is too subjective, corpora help. Table 2 presents data from the BNC[93].

Table 2

single word	freq in the BNC	potential collocation?	freq in the BNC
love	4,074	**I love you**	712 (over 17%)
truth	8,250	**the truth is**	427 (over 5%)
promise	2,305	**kept my promise**	3 (0.1%)
distance	6,829	**keep your distance**	12 (0.2%)

That is, the frequency of **love**, e.g. is 4,074. Of that number, 712 are found in **I love you**, so it is rightly claimed to be a collocation. Both **I love you** and **the truth is** are really frequent, so they *are* collocations.

But: while there are 2,305 instances of **promise**, there are just three examples of the combination **kept my promise**, and while there are almost 7,000 instances of **distance**, there

[93] The British National Corpus stood at 100 million words at the writing of the article this section is based on. Today's English corpora are in the range of 1–2 billion words.

are just 12 hits for **keep your distance**. What is going on here? Are **kept my promise** and **keep your distance** *not* collocations, then?

The four problems with **kept my promise** and **keep your distance** are: ①, ②, ③ and ④.

① We get many more hits than that: we don't just want *word forms* but *lexemes*: we want all occurrences of KEEP. The solution: we must *lemmatize* (bring together all word forms), and then the corpus query finds all occurrences (word forms) of KEEP: **keep, keeps, kept, keeping**. Now we find a lot more examples: **keep my promise** and **keeping my promise** (there probably will be no ***keeps my promise** though).

The results of the improved search that includes (**keep / keeps / keeping / kept ___ distance** and **keep / keeps / keeping / kept ___ promise**) are as follows:

Table 3

single word	freq in BNC	potential collocation?	freq in BNC	improved freq
love	4,074	**I love you**	712 (over 17%)	
truth	8,250	**the truth is**	427 (over 5%)	
promise	2,305	**KEEP * promise**	3 (0.1%)	**85 (3.7%)**
distance	6,829	**KEEP your distance**	12 (0.2%)	**127 (1.9%)**

② Beside the **my** of **my promise**, however, we also need to search for **my / your / his / her / (its) / our / their.** By doing that, we now get various combinations ranging from **I kept my promise** and **she keeps her promise** to **you are keeping your promise** and **they have kept their promise**.[94] (No tabulated data presented tis time).

③ We get even more hits because KEEP and its collocating (object) noun need not be *adjacent*: there are what are termed *discontinuous collocations*. We need to search for **keep/keeps your/her** *etc* **___ distance**, allowing for a word or words to be between the pronoun and the noun. *Wildcards* must be used to register that.

Note that we can have *still* more hits than that because the collocates can be to the left, not just to the right of the base: in **his/their** *(etc)* **____ *promises* were never *kept***, the collocate – **promise** – is the 3rd word to the *left* of the base, **kept**. The satisfying solution is to increase the search *span* to 5 words L/R (i.e. to increase the search domain so as to find five words to the left and five to the right): that will increase the above percentage even more.

④

There is a serious problem with the practical application of recurrence. Consider Table 4:

[94] Both of these (**promise** and **distance**) could actually be in the plural, though with the latter it is less likely.

Table 4

single word	freq in BNC	potential collocation?	freq in BNC
and	2, 689,689	**and as**	8,515

Apparently, the statistical criterion of recurrence does not solve problem of defining collocation. With a frequency of over two and a half mn, the item **and** is more likely to occur much more often before or after *any* other word – but these are accidental occurrences.

There may be two ways out:

(a) Stipulate that collocations must be grammatical units. The problem then will be that some are not: **dogs—bark, horse—neigh** e.g. are subjects and verbs in clauses, not phrases.

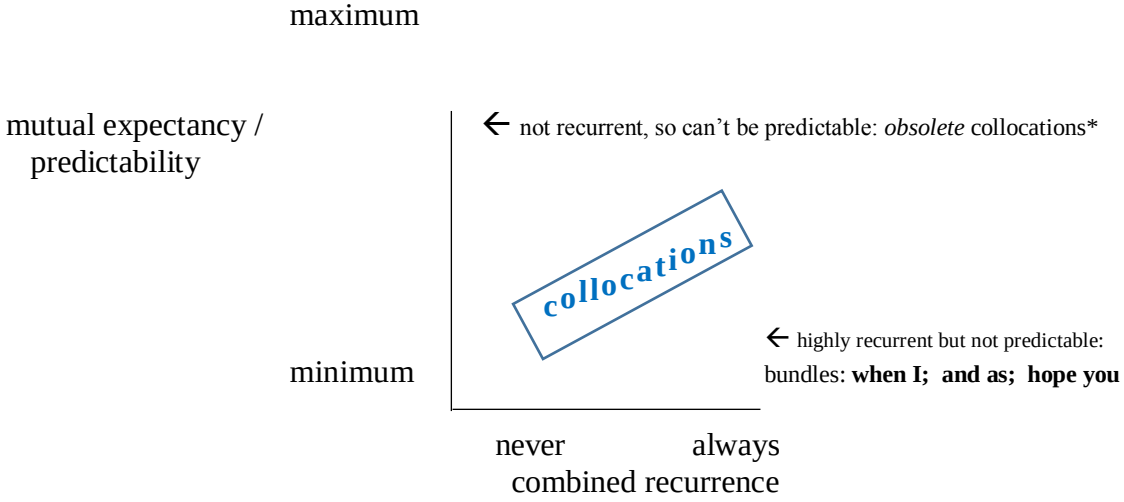
(b) Restrict the notion of collocation to those combinations that show *mutual expectancy*. This will exclude the items in Table 4 above. If two words mutually expect each other, then speakers can *predict* the occurrence of one whenever they encounter the other. The notion of predictability takes the speaker's perspective, while mutual expectancy is seen from the perspective of the language.

Predictability and mutual expectancy capture the *psychological essence* of collocation, i.e. the associative relation between syntagmatically related words.

Summing up what collocations are about: a collocation is (i) a combination of (ii) two or more words that are (iii) adjacent or are within a span of X, which show (iv) mutual expectancy / predictability.

Figure 9 gives one an idea of the awkward status of collocations within phraseology.

Fig. 9



*for a small section of the community

The **collocations** bar shows the extension of the phenomenon of collocation; it does not reach to the origin (the 0): that area is for free phrases.

Combinations with a very high degree of predictability show a tendency towards *idiomatization*.

2.5 Listemes

Listeme is a convenient (though not widely known) term for any morpheme or group/string of morphemes (including words and idioms) that needs to be listed in the speaker's mental lexicon as a unit because something must be memorized about it rather than figured out from its parts. Recall that *arbitrariness* as a *design feature* of language means that there is no necessary/predictable link between a form and its meaning, which means that all *simplex* forms – all morphemes – need to be memorized, i.e. are listemes. But many larger-than-morpheme form–meaning pairings/units, i.e. *complex* forms are also memorized: as we have seen, memorizing **down** is no guarantee for knowing ***shut down*** or ***down*load** or **walk *down* the aisle**, because the meanings of these cannot be figured out from their parts, i.e. their meaning is not compositional. The lexical items **shut down** and **walk down the aisle** and **flog a dead horse**, which are strings of morphemes, are thus listemes.

It is important to stress that an expression need not even be idiomatic, i.e. non-compositional, to be a listeme. Large numbers of compositional phrases, even whole-sentence strings are stored/listed as wholes.

2.5.1 Listemes, lexemes, lexical items

In this largely recapitulatory section, some of the expressions encountered thus far will be compared from the point of view of listemehood, lexemehood, and lexical item-hood.

- The expression MOUSETRAP is a *compound noun*; it is a lexical item; it is a word both *phonologically* and *syntactically;* and a *lexeme*. It is made up of just two components, as the most typical of compounds. Once formed from *two lexemes* – not by speakers on line, but the language system; it is a ready-made item – a pair of virtual brackets is put around its members, metaphorically speaking, which sees to it (even more metaphorically speaking) that neither the syntax nor the semantics has access to what is inside. The contents of the brackets, as we said before, "is a no-go area for the grammar".

  Exactly the same can be said about FLASH DRIVE.

  Notice, incidentally, that the word-internal "syntactic" relation between the components is different in these two compounds: while MOUSETRAP is a trap for mice – it traps mice – a FLASH DRIVE is not a drive for flashes – nor does it drive flashes. Indeed, not knowing the meaning contribution of FLASH to the whole does not hinder one from knowing what the compound as a whole means.[95]

  If /ˈnoʊmænzlænd/ (mentioned in the discussion above) is considered as a compound, the difference in *punctuation* between **no-man's-land** and **no man's land** is just about as irrelevant for its linguistic status as that between a solid and an open compound such as **mousetrap** vs **flash drive**. If it *is* a *compound*, then it is a lexical item, a lexeme, a listeme, and a word. If considered a *phrase*, then it is a lexical item and a listeme, but not a lexeme (and obviously not a word).
- Particle–verb expressions are also genuine *compound words*; DOWNLOAD is a compound *lexeme*, as was claimed above; it is a listeme.
- Unlike the compounds in (A) in 2.4, the expressions in (B)–(F) are various multiword items; they are lexical items, though obviously not words.[96] They are also listemes. They typically contain words in the syntactic/phonological sense, but these

[95] About twenty years ago there was a (by the then standards) large capacity "floppy" disk called **zip disk**, used in **zip drives**. You do not need to know what the **flash** or the **zip** bit in these mean to understand these words.

[96] Several sources use "multiword lexeme"; under our definition, such a term makes no sense.

words *within these multiword units* are not lexemes, not even lexical units, because they are not form–meaning pairings; simply put, they either have no meaning whatsoever, or their meanings are irrelevant.

The expression **give somebody a bell** is different from the rest in (B) – and from **ring a bell** e.g. – in two senses: first, it contains the item *"somebody"*, which is a *variable*. This means that it's evidently not the actual word **somebody**, just its *slot* that is memorized by speakers, which, when retrieved, gets filled by whatever word is appropriate to the situation. The word "something" in **take advantage of something** is a similar variable.

Walk down the aisle is a *lexical item* but obviously not a word in any sense of "word", so not a lexeme; it is a *listeme* though, since it is a unique form paired with a particular meaning. It contains *four words*, of which the third, as we have seen, is "less of a word", i.e. not autonomous but phonologically dependent – a *proclitic*. Being constructed from four *words* does not mean being constructed from four *lexemes*, because those form–meaning pairings are not relevant *in this string*. What is really important about the whole expression is that it is a (kind of) *listeme*, an *idiom*. Note that a speaker's knowledge of the etymology/story of, or the metaphor/image behind, this idiom does not stop it from being an idiom. Fairly *transparent* it may be once you know the meaning, but it is neither *predictable* nor *compositional*.

Flog a dead horse is similar in every respect, but it is even less transparent and less compositional: while you could argue that marriage *may* involve walking down an aisle, what you do while **flogging a dead horse** is not flogging at all. As has already been pointed out, transparency or compositionality are not either-or things but make up a *gradience* of (more) transparent/compositional to non-transparent, i.e. opaque and non-compositional.

- The items in (C) are just two illustrations of varied sentence-length expressions. While most sentences are no doubt new in the sense of being produced online, and thus illustrate what is called *creativity* of language, a great number of them – a lot more than usually thought – are not *ad hoc* generated, but listed: *greetings*, and also *sayings*, *proverbs*, *maxims*, *adages*, *dictums* [note that none of these has a usable definition], any famous quotes, even favourite lines of poetry. You can't really be blamed if you do not know where the exact boundaries between some of these are. The expression **if it ain't broke, don't fix it** would probably be classified as a saying, but **what you see is what you get** – with many, many others – just does not seem to have a label.

The expression **ain't**, a non-standard contracted form, is a "corrupted" version of the word forms **isn't** or **aren't** or **hasn't** or **haven't** (in certain, but not all, of the syntactic uses of these verbs) as well as of **do not** and **does not**. The exact nature – and history – of the contraction process is hard to establish, though.

- The same may be said about the form **don't**, a standard contraction of **do** and **not**. Recall that while **don't** is a contraction, where the **n't** is a clitic, **don't** is also a syntactic word, as it moves as a unit in questions. By contrast, subject + auxiliary contractions contain a *(en)cliticized* auxiliary, but they do not produce syntactic words:

You ain't/don't/can't go → Ain't/don't/can't you go?

while the **I'm** in **I'm here** or the **She's** in **She's seen it** cannot move. This is an important point concerning wordhood on which **not**-cliticization and auxiliary cliticization differ. Both **not**-contractions and auxiliary contractions are words in the phonological sense, but only the former are syntactic words.

- Group (D), as we have seen, contains idioms. We have also seen two important – and often sadly ignored – facts about idiomaticity: that it is (i) *graded*, or *scalar*, rather than an either-or matter; (ii) that it characterizes, to various degrees, all the various types of multi-word units, not just these, figurative, colourful idioms. These expressions are phrasal in terms of structure: verbal and nominal phrases. They can be called multi-word units, multi-word elements (MWE's), or *phraseologisms*; within that category, they are idioms.[97]

The expressions in (D) are lexical items but not lexemes; they do, however, have a syntactic *head*[98], which does assume the function of a lexeme when it gets into the syntax and gets inflected. In those idioms where this is at all possible – e.g. in **flog a dead horse** – the head will be inflected, i.e. have a paradigm, i.e. be a lexeme: **flogged/flogs a dead horse**.

- Group (E), as we have seen, contains special idioms, whose components may have no existence outside of these multi-word units, and may have no meaning whatever: **the whole *caboodle***; **to and *fro***; **if I had my *druthers***. Other like examples include **days of *yore***;[99] ***ulterior* motives/reason/motive**;[100] **in fine/good *fettle***[101]; **in the *offing***;[102] ***wend* one's way**.[103] It could, nevertheless, be argued that where this unique word in these idioms may be replaced with a synonym (as in the definitions in Footnotes 100, 101, 103), it does have a meaning: **ulterior**, **fettle** and **wend** may be such items. This unique word, you will recall, is like a "cranberry morpheme", one with no meaning, in complex words such as the **cran-** in CRANBERRY or the **rasp-** in RASPBERRY, or the **Wednes-** in WEDNESDAY.

Thus, the **cran-** is to CRANBERRY at the morphological level as **offing** is to **in the offing** at the level of the lexicon. The word **offing**, then, may be nicknamed a "cranberry word",[104] and **in the offing**, a "cranberry idiom".

2.5.2 Phrasemes

Multi-word units are referred to by many names: phrasemes, phraseologisms, phraseological expressions, set expressions, set phrases, multi-word elements, multi-word sequences, multi-word expressions. The problem, however, is that it is nowhere explained what exactly they mean.

Below is a chart with a possible classification of phrasemes, followed by examples of phrasemes. See how easy/hard it is to classify them using the guidelines of the chart.

---

[97] Note that in English, neither *expression* nor *phrase* means the same thing as the Hungarian "kifejezés", which basically translates English "idiom". *Expression* has a wider application: it is used for any *string*, i.e. *chunk* of language, while *phrase* is a syntactic term, thus opposing *clause* and *word*.

[98] Verbs in VP's, but the nouns in the NP's, not the DP's; here, nouns are the relevant heads lexically.

[99] **Of yore** = of a long time ago.

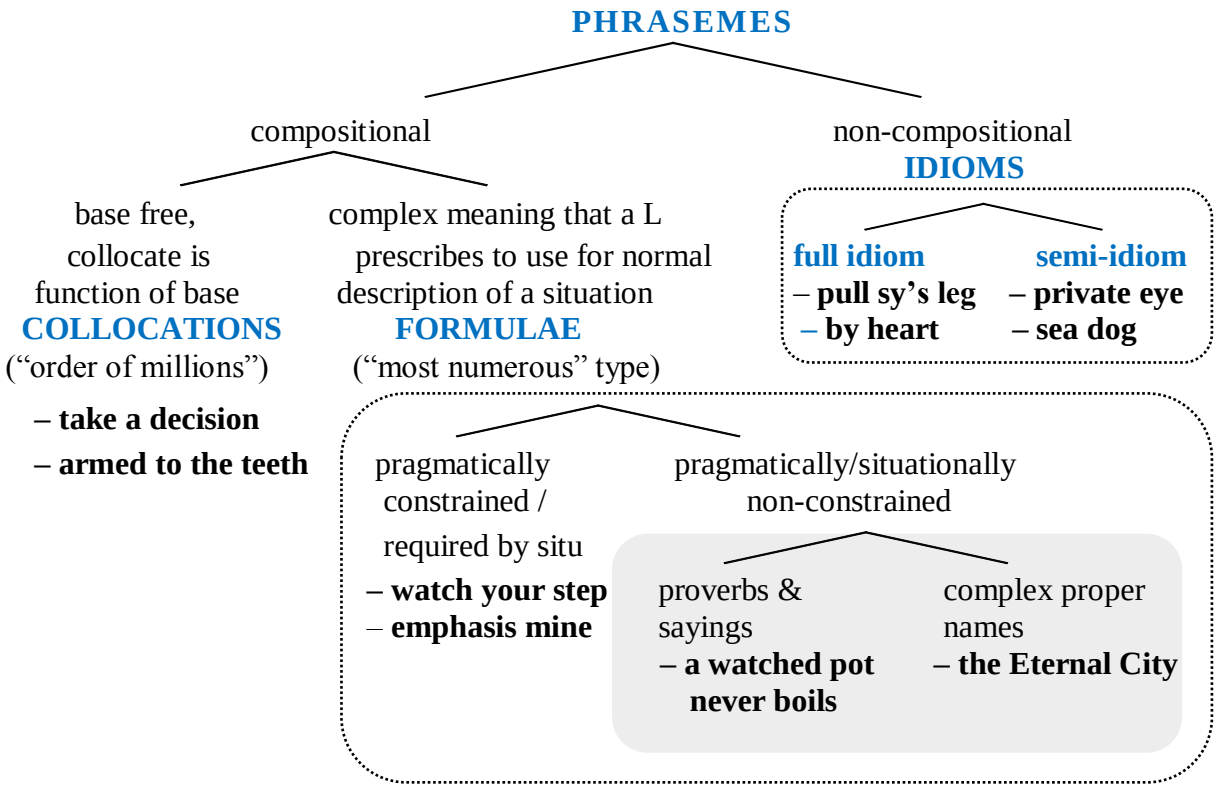
[100] **An ulterior reason** = a secret purpose/reason for doing something.

[101] **In fine/good fettle** = in good condition.

[102] **In the offing =** likely to happen soon.

[103] **Wend one's way** = turn/direct one's way.

[104] Unfortunately, *cranberry word* is also used for words that *contain* a cranberry morpheme; WEDNESDAY in that sense, is a cranberry word, containing a *cranberry morpheme*, **Wednes-**.



2.5.2.1 Assorted phrasemes

1. kick the bucket
   – no replacement w/ (quasi)synonym
2. pull smb's leg
3. wet paint
   – not 'caution, painted'
   (no asterisk needed!);
   – R. ostorozhno, okrasheno;
   – H. frissen mázolva 'freshly painted'
4. in other words
   – R. inache govor'a
5. to make a long story short
   – R. koroche govor'a
6. take a shower
7. come to one's senses
8. put smth on the map
9. bull session
10. play chicken
11. bluestocking
    – R. sin'ij chulok
12. put smb through his paces
13. go ballistic
14. have a cow
15. by heart
16. bone of contention
    – R. jabloko razdora 'the apple of discord'
17. private eye
18. start a family
19. sound asleep
20. make a decision
    – Br. take a d.
21. make an apology
22. black coffee
23. laugh in one's sleeve
24. (if) you've seen one, you've seen all
25. happy birthday to you!
    – cf. "older" H. Isten éltessen!
26. the Red Planet
27. we all make mistakes
28. what is your name?
    – H. hogy hívnak?
    – R. kak teb'a zovut?
29. best before
    – R. srok godnosti
    – Fr. À consommer avant…
    – G. Mindestens haltbar bis…
30. break a leg!
31. no parking
32. roger!
33. ————
    – Pol. smacznego!
    – H. jó étvégyat!
    – G. Guten Appetit!
34. hold the line!
35. watch your step!
36. all you can eat

2.5.3 Lexemes but not listemes

The derived words LISTEMEHOOD and LEXEMEHOOD, which have been used in the text of the discussion above, are lexical items, words, and lexemes at the same time; however, for most speakers of English they are probably *not listemes*. Rather, they are understood by resorting to the same morphological rules that have produced them. If you, a reader, encountered LISTEMEHOOD and LEXEMEHOOD for the first time in the context of this passage, then you must also have "reverse-engineered" the morphological rule that put LISTEME or LEXEME and the *derivational suffix* **-hood** together.

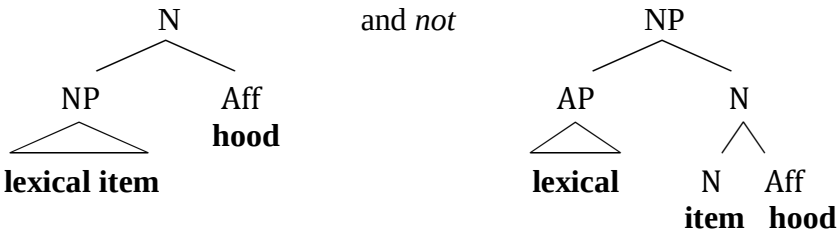
The expression LEXICAL ITEMHOOD (or LEXICAL-ITEM-HOOD) is also a lexical item, a word, and a lexeme at the same time. It is also not a listeme, similarly to these other two **-hood** words (or even less so: it is an even more *ad hoc* formation). It also shows that a phrase can exceptionally be suffixed: the **-hood** attaches not to ITEM but to the entire string **lexical item**. Note that the variant with two hyphens actually suggests this analysis: that /ˌleksɪkl ˈaɪtəmhʊd/ is *one word*.[105]

The structure, then (irrespective of the written form) is this:

word (N)		(Noun) phrase
[N [NP **lexical item**][Aff **hood**]]	and *not*	[NP [AdjP **lexical**] [N [**item**][Aff **hood**]]]

in tree diagram form:

Fig 9



The alternative – if we do not wish to allow phrases to be affixed, i.e. allow there to be phrases within words – would be to say that LEXICAL ITEM is (already) a compound word, of the adjective–noun structure. Then **lexical item-hood** is a normal *affixed compound*, like **mousetrap-s** or **download-ed**, i.e.

[N [N [N **mouse**] [N **trap**]] [Aff **s**]] [V [V [P **down**][V **load**]] [Aff **ed**]]

tree diagrammed:

Fig 10

[105] Without the two hyphens, the interpretation of **lexical itemhood** might be that this is a *lexical kind of itemhood* – a wrong analysis; **lexicalitemhood** is a more than unlikely variant.

Annotated select bibliography of lexicology *and* lexicography

Mostly books and monographs; only exceptionally does it include articles, which are even harder to access.

Aarts, B & McMahon, A eds (2006) *The Handbook of English Linguistics*. Blackwell.
☆ Includes several writings relevant for lexicology and lexicography; all of them are listed below. The handbook contains studies of "all aspects of Present-Day English… from a variety of different angles, both descriptive and theoretical, but with a methodological outlook firmly based on the working practices developed in modern contemporary linguistics".

Aarts, B & Haegeman, L (2006) "English word-classes and phrases". In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics*. Blackwell. 117–145.
☆ Good place to start. Has lots of syntax, not just word classes. They're syntactic categories, after all.

Adams, V (1973) *An introduction to modern English word-formation*. Longman.
☆ Detailed account of just that: English word formation, now rather obsolete.

Adams, V (2013) *Complex words in English*. Routledge.
☆ Originally planned as the 2nd edition of Adams (1973), it is an inclusive, and mainly technical, account of word formation in English.

Aitchison, J (1994) *Words in the mind. An introduction to the mental lexicon*. 2nd ed. Blackwell.
☆About how humans learn, store, understand, and retrieve words. Part I/1 nicely summarizes the differences between the mental lexicon and print dictionaries.

Akmajian, A, Demers, R A, Farmer, A K, Harnish, R M (2010) *Linguistics. An introduction to language and communication*. 6th edition. MIT.
☆ Accessible broad overview of linguistics, with illustrations from English. No separate chapter on dictionaries but the lexicon is adequately treated in both Chs 2 & 6. Relevant parts: Ch 2: Morphology, Ch 6: Semantics.

Allan, K (2001) *Natural language semantics*. Blackwell Publishers Ltd.
☆ Relevant parts: 1: Some fundamental concepts for semantics 2: Words and worlds and reference 3: The lexicon and the encyclopedia 4: Morphology and listemes

Aronoff, M & Rees-Miller, J eds (2003) *The handbook of linguistics*. Blackwell Publishing.
☆ Serious but accessible handbook on all aspects of language, chapters are by known authors. Relevant parts: Ch 9 Morphology by Spencer, A Ch 10 The lexicon by Cruse, D A.

Atkins, S & Rundell, M (2008) *The Oxford guide to lexicography*. OUP.
☆ A serious textbook on the making of dictionaries. Has accompanying reader: Fontenelle (2008).

Ayto, J (2006) "Idioms". In: Keith Brown ed. *Encyclopedia of Language and Linguistics*. Elsevier.
☆ An encyclopedia-entry-sized summary of idiom types and properties.

Baerman, M, Brown, D & Corbett, G G (2005) *The syntax–morphology interface. A Study of Syncretism*. CUP.
☆ An advanced treatment of syncretism.

Bauer, L (1998) *Vocabulary*. Routledge.
☆ Step-by-step, hands-on introduction to vocabulary, with lots of in-text exercises.

Bauer, L (2003) *Introducing linguistic morphology*. 2nd ed. Edinburgh University Press.
☆A two level introduction to morphology – an entry-level 53-page Part One: Fundamentals, followed by a more technical 240-page Part Two: Elaboration.

Bauer, L (2004) *Morphological productivity*. CUP.
☆A serious treatment of productivity as it affects morphological systems.

Bauer, L (2006) "Compounds and minor word-formation types". In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics*. Blackwell. 483–506.
☆ See Aarts, B & McMahon, A eds (2006).

Behrens, S J & Parker, J A eds (2010) *Language in the Real World. An introduction to linguistics*. Routledge.
☆ Ch 4 What dictionaries reveal about language and dictionary makers
is an accessible introduction to the workings and types of dictionaries.

Biber et al (2007) *Longman Grammar of spoken and written English*. Longman.
☆ Massive descriptive grammar, corpus-driven, new *and* reliable.
Section A 2.1–2.5 is staple reading for word classes; Section E 13 treats lexical bundles, idiomatic phrases, and binomials, among others.

Blevins J P (2006) "English Inflection and Derivation". In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics.* Blackwell. 507–536.
☆ See Aarts, B & McMahon, A eds (2006).

Booij, G (2007) *The grammar of words. An introduction to morphology*. 2nd ed. OUPÍ
☆ Serious treatment of morphology. Examples mainly from English and Dutch.

Brinton, L J & Brinton, D M (2010) *The linguistic structure of modern English*. John Benjamins.
☆A general introduction to the description of English. Relevant parts: Chapter 4: Internal structure of words and processes of word formation; Chapter 6: lexical semantics.

Brown, K ed (2006) *Encyclopedia of Language and Linguistics*. Elsevier.
☆ A monumental work on all kinds of linguistic subjects.

Carstairs-McCarthy (2002) *An Introduction to English Morphology: Words and Their Structure*. Edinburgh University Press.
☆A reliable, general introductory text.

Coates, R (1999) *Word structure*. Routledge.
☆ Step-by-step, hands-on introduction to morphology, with lots of in-text exercises.

Coleman, J (2006) "Lexicography". In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics.* Blackwell. 581–600.
☆ See Aarts, B & McMahon, A eds (2006).

Cowie, A P ed (2009) *The Oxford history of English lexicography*. Vol I: *General-purpose dictionaries*. Vol II: *Specialized dictionaries*. Clarendon Press Oxford.

Cruse, D A (2000) *Meaning in language. An introduction to semantics and pragmatics*. OUP.
☆ A textbook on (especially lexical) semantics and pragmatics. Most relevant part: Part 2 Words and their meanings.

Cruse, D A (1986) *Lexical semantics.* CUP.
☆ An advanced book about the meaning of words; takes a descriptive approach.

Crystal, D (1967) *English word classes.* Lingua 17, 24–56.
☆ Very insightful approach to the definition/groupings of the word classes of English. A classic.

Crystal, D (2008) *A dictionary of linguistics and phonetics*. 6th ed. Blackwell Publishing.
☆ Reliable source if you need one like this.

Crystal, D (ed) (1997) *The Cambridge encyclopedia of language*. 2nd ed. CUP.
☆ A wonderful volume, offers a wealth of – textual & pictorial – information on all aspects of language & English. Relevant parts: 17: Semantics 18: Dictionaries

Crystal, D (ed) (2003) *The Cambridge encyclopedia of the English language*. 2nd ed. CUP.
☆ Very enjoyable but no-nonsense book about language for the general reader.
Relevant parts: 8: The nature of the lexicon. 11: The structure of the lexicon. 12: Lexical dimensions 14: The structure of words 15: Word classes.

de Schryver, G-M (2003) "Lexicographers' dreams in the electronic-dictionary age" In: *International Journal of Lexicography*, Vol 16 No2. OUP.
☆ Anticipates much of the fate of post-print dictionaries.

É Kiss, K, Kiefer, F & Siptár, P (1998) *Új magyar nyelvtan* [New Hungarian grammar]. Osiris. [In Hungarian]
☆ Has syntax/morphology/phonology part. The morphology part, a survey of (mainly Hungarian) morphological issues, is relevant. Uses four notions of word, including "morphological word"; *phonological word* is defined differently from the text above.

Fasold, R W & Connor-Linton, J eds (2006) An introduction to language and linguistics CUP.
☆ A comprehensive textbook. Relevant parts: Ch 2 Words and their parts Ch 4 Meaning.

Fontenelle, T (2008). *Practical lexicography. A reader*. OUP.
☆ A collection of works, all of them important contributions to lexicology *and* lexicography not easily available outside of this volume. Accompanies Atkins, S & Rundell, M (2008).

Fromkin, V, Rodman, R & Hyams, N (2003) *An introduction to language*. 7th ed. Wadsworth.
☆ An accessible introduction to language, one of the long-time staples intro lingx books worldwide. Chs 3 and 5 are especially down-to-earth. Relevant parts: Part 2. Ch 3: Morphology. The words of language. Ch 5: The meanings of language.

Fromkin, V, Rodman, R & Hyams, N (2011) *An introduction to language*. 9th ed. Wadsworth. Cengage Learning.
☆ New edition of Fromkin, V, Rodman, R & Hyams, N (2003). Relevant parts: Ch 1 Morphology Ch 3 The meaning of language

Geeraerts, D (2010) *Theories of Lexical Semantics*. OUP.
☆ Serious treatment of possible approaches to lexical semantics.

Hanks, Patrick (2013) *Lexical analysis. Norms and exploitations*. The TIT Press.
☆ Readable up-to-date treatment of some important issues of the lexicon.

Hartmann, R R K & James, G (1998) *Dictionary of Lexicography*. Routledge.
☆ Not the biggest imaginable, but a reliable source for definitions of notions in lexicography.

Haspelmath, M (2002) *Understanding morphology*. Arnold.
☆ Advanced book on linguistic morphology, "demonstrating the diversity of morphological patterns in human language and elucidating broad issues that are the foundation upon which morphological theories are built" [from the Preface of the 2nd ed; see below].

Haspelmath, M & Sims A D (2010) *Understanding morphology*. 2nd ed. Hodder Education.
☆ 2nd edition of Haspelmath, M & Sims A D (2010). The material has been substantially restructured and some topics have been expanded. The goal was to bring foundational issues to the forefront [from the Preface].

Hudson, R (1995) *Word meaning*. Routledge.
☆ Easy, practical introduction to words, lexicology, and lexical semantics.

Hurford, J R & Heasley, B (1983) *Semantics. A coursebook*. CUP.
☆ Very reader-friendly coursebook with lots of in-text exercises.

Jackson, H (1982) *Analyzing English. An introduction to descriptive linguistics*. 2nd ed. Pergamon Press.
☆ Especially relevant: Part Three: Words.

Jackson, H (1988) *Words and their meaning*. Addison Wesley Longman Limited.
☆ Accessible introductory textbook devoted to (Engish) lexicology and lexicography, now rather dated but still reliable. The

Jackson, H (2002) *Lexicography*. Routledge.
☆ Accessible book devoted to (basically: English) lexicography.

Jackson, H & Amvela, E Z (2007) *Words, meaning and vocabulary*. Continuum.
☆ Simple, not-too-technical overview of English lexicology.

Jeffries, L (1998) *Meaning in English. An introduction to language study*. Palgrave.
☆ Accessible introduction to language. Relevant part: 3 Words and meaning.

Jeffries, L (2006) *Discovering language. The structure of modern English*. Palgrave Macmillan.
☆ Ch 3 (Words) is a basic introduction to morphology and word classes.

Katamba, F (2005) *English words. Structure, history, usage*. Routledge. 2nd ed.
☆ Accessible introduction to many aspects of English words.

Kearse, K (2006) "Lexical semantics". In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics.* Blackwell. 557–580.
☆ See Aarts, B & McMahon, A eds (2006).

Kenesei, I (2007) "Semiwords and affixoids. The territory between word and affix". In: *Acta Linguistica Hungarica* 54: 263-293.
☆ Treats the shady but very much existing area between ("full") word and (genuine) affix.

Kennedy, G (1998) *An Introduction to Corpus Linguistics*. Longman.
☆ An introductory book on corpus linguistics, includes chapters on corpus design, corpus analysis, and the applications of such analysis.

Kiefer, F ed (2000) *Strukturális magyar nyelvtan 3. Morfológia*. Akadémiai Kiadó.
☆ Huge volume (technical in places) overviewing (not just Hungarian) morphological issues. In Hungarian.

Kiefer, F ed (2006) *Magyar nyelv*. Akadémiai. [In Hungarian]
☆ Relevant parts: 3: Alaktan [Morphology] by Kiefer, F 4: Szófajok [Word classes] by Kenesei, I 7: Szemantika [Semantics] by Kiefer, F & Gyuris, B

Kiefer, F ed (2008) *Strukturális magyar nyelvtan 4. A szótár szerkezete*. Akadémiai Kiadó.
☆ Huge volume (technical in places) overviewing (mainly Hungarian) lexical issues. In Hungarian.

Kilgarriff, A (1997) "Putting frequencies in the dictionary". In: *International Journal of Lexicography*, Vol. 10 No. 2 OUP.
☆ Frequency information about words and whether it is useful in dictionaries.

Kirkness, A (2004) "Lexicography" In: Davies, A & Elder, C ed *A handbook of applied linguistics*. Blackwell Publishing.
☆ Accessible overview of lexicographic issues.

Kreidler, Ch W (2002) *Introducing English semantics*. Taylor & Francis.
☆ Introduction to the principles of linguistic semantics at university level, as the blurb says.

Kroeger, P R (2005) *Analyzing grammar. An introduction*. CUP.
Relevant parts: 2: Analyzing word structure, 13: Derivational morphology, 14 Valence-changing morphology, 15 Allomorphy, 16 Non-linear morphology, 17 Clitics.
☆ Mostly on morphology (lexicology/lexicography not in self-contained chapters); technical at places.

Kuiper, K & Allan W S (1996) *An introduction to English language. Sound, word and sentence*. Macmillan Press Ltd.
☆ Relevant parts: Ch 5 The form and function of words Ch 6 Word meanings and vocabularies.

Landau, S L (2001) *Dictionaries. The art and craft of lexicography*. 2nd ed. CUP.
☆ A very well-written book on lexicography (more American than British), on how dictionaries are researched and produced. Examines & explains all features of dictionaries; illustrations from various works.

Lieber, R (2009) *Introducing morphology*. CUP.
☆ Comprehensive treatment of morphology, intended for undergraduate students with no more background than an introductory course in linguistics.

Lieber, R (2004) *Morphology and lexical semantics*. CUP.
☆ "Explores the meanings of morphemes and how they combine to form the meanings of complex words, including derived words…, compounds… and words formed by conversion" [from the blurb of the book].

Lipka, L (1992) *An Outline of English Lexicology. Lexical Structure, Word Semantics, and Word-Formation.* 2nd ed. Max Niemeyer Verlag Tübingen.

Malmkjær, K ed (2002) *The linguistics encyclopedia.* 2nd ed. Routledge.
☆ Medium-sized encyclopedia of linguistics.

Martsa, S (2007) *English morphology. An introduction*. Pécsi Tudományegyetem / Nemzeti Tankönyvkiadó.
☆ Comprehensive treatment of English morphology for Hungarian students.

McArthur, T (1992) *The Oxford Companion to the English language*. OUP New York.
☆ Huge encyclopedia covering all aspects of (the English) language. Fairly traditional, rich on detail.

McEnery, T & Gabrielatos, C (2006) "English Corpus Linguistics" In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics.* Blackwell. 33–71.
☆ See Aarts, B & McMahon, A eds (2006).

McGregor, W (2009) *Linguistics. An introduction.* Continuum.
☆ An accessible introductory text. Relevant parts: Ch 3 Structure of words: morphology Ch 4 Lexicon Ch 6 Meaning.

Mel'čuk, I (2006) *Aspects of the theory of morphology*. (D Beck editor). Trends in Linguistics. Studies and Monographs 146. Mouton de Gruyter.
☆ Serious volume, theory-laden, technical in places.

Merrison, A J, Bloomer, A, Griffiths, P & Hall C J (2014) *Introducing language in use*. 2nd ed. Routledge.
☆ New coursebook for introductory courses, use-based, with many examples. Relevant parts: Ch 5 Words Ch 6 Semantics.

Meyer, Ch F (2004) *English corpus linguistics. An introduction*. CUP.
☆ Accessible overview of corpus linguistics, with English in focus.

Meyer , Ch F (2009) *Introducing English linguistics*. CUP.
☆ Broad overview of the subject of English linguistics, relatively recent and reliable. Ch 6 is a readable treatment of semantics, morphology and dictionaries.

Minkova, D & Stockwell, R (2006) "English Words". In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics.* Blackwell. 461–482.
☆ With an emphasis on history, it treats the vocabulary of English, mainly in terms of size, type, and token frequency.

Moon, R (2005) "Multi-word Items" In: *The Oxford Handbook of the Word* ed. by John R Taylor.

Moon, R (2006) "Corpus approaches to idiom" In: Keith Brown ed. *Encyclopedia of Language and Linguistics*. Elsevier.
☆ A useful roundup of idioms.

Newson, M et al. (2006) *Basic English Syntax with Exercises*. 2006. Bölcsész Konzorcium, ELTE.
English Syntax at the BA level and upwards; especially relevant is the chapter on word categories.

O'Grady, W, Dobrovolsky, M & Katamba, F. *Contemporary linguistics* (1996) 3rd ed. Longman.
☆ Comprehensive introduction to linguistics, looks at not only how language is structured but also how it is used functionally & socially. Ch 6: Interfaces: morphology and phonology; morphology and syntax Ch 7: Semantics: the analysis of meaning.

Pavey, E L (2010) *The structure of language*. *An introduction to grammatical analysis*. CUP.
☆ Accessible introduction. Relevant parts: Ch 2 The structure of words Ch 4 The structure of meaning.

Pethő, G T (2004) *Poliszémia és kognitív nyelvészet. Rendszeres főnévi poliszémiatípusok a magyarban*. [Polysemy and cognitive linguistics. Systematic polysemy types in Hungarian]. PhD dissertation. ELTE.
☆ Advanced discussion of polysemy types; especially relevant part is II. A főnevek poliszémiája a magyarban [Polysemy of nouns in Hungarian].

Pinker, S (1994) *The language instinct*. Penguin Books.
☆ A must-read for any student of language. Available in other and more recent editions. Most relevant part: 5 Words, words, words.

Pinker, S (1999) *Words and rules. The Ingredients of Language*. Basic Books.
☆ "Illuminates the nature of language & mind by choosing a single phenomenon and examining it from every angle imaginable": the phenomenon of regular and irregular verbs [from the blurb of the book]. The chapter titles are witty – typical Pinker – but tell little about the topics.

Plag. I (2003) *Word-formation in English*. CUP.
☆ Serious textbook about what the title says.

Plag, I (2006) "Productivity". In: Aarts, B & McMahon, A eds *The Handbook of English Linguistics.* Blackwell. 537–556.
☆ See Aarts, B & McMahon, A eds (2006).

Quirk, R et al (1985) *A comprehensive grammar of the English language*. Longman.
☆A huge descriptive grammar of English from the dawn of the corpus era. Includes Appendix I: Word formation, a 70-page treatment of affixation, conversion, compounding and "miscellaneous modes".

Radford, A (1988) Transformational grammar. A first course. CUP.
☆Time-honoured introduction to TG. Some of Part 2 Structure, 4 Noun phrases, 5 Other phrases, and 7 The lexicon may be relevant.

Radford, A, Atkinson, M, Britain, D, Clahsen & H, Spencer, A (2009) *Linguistics. An introduction.* 2nd ed. CUP.
☆ Part II: Words is a comprehensive treatment of all matters lexical, from morphology and lexical semantics through processing and the mental lexicon to lexical variation and disorders.

Readings in morphology. (1994) JATE BTK Angol Tanszék. JATEPress 1994.
☆ Basic readings from the 70s and 80s, some classics.

Richter, B ed (2006) *First steps in theoretical and applied linguistics*. Bölcsész Konzorcium.
☆ A basic step-by-step and topic-by-topic introduction to linguistics for BA students of English. Relevant parts: Ch 5: Words, meanings and their relationships, Ch 9: Keywords in context: corpus linguistics, Ch 10: What is in a dictionary? Lexicography.

Riemer, N (2010) *Introducing semantics*. CUP.
☆ Up-to-date introductory textbook to semantics for undergraduates.

Rizo-Rodríguez, A. (2008) *Review of five English learners' dictionaries on CD-ROM*.
☆ Review of the features of the five learners' dictionaries available at the time of writing.

Rundell, M (2008) "The corpus revolution revisited". In: *English Today* 93, Vol 24, No 1.
☆ An update on the rise of electronic language corpora and their impact on dictionaries. Looks back on a 1992 article by Rundell, M & Stock, P *The corpus revolution* in English Today.

Saeed, J I (2003) *Semantics*. 2nd ed. Blackwell Publishing.
☆ Serious book on (all types of, not just lexical) semantics. Relevant parts: Ch 2 Meaning, thought and reality Ch 3 Word meaning.

Sinclair, J (1991) *Corpus, Concordance, Collocation*. OUP.
☆ A classic on corpus linguistics by the creator of the Bank of English and one-time editor of Collins–Cobuild learners' dictionaries.

Sinclair, J (2003) *Reading concordances. An introduction*. Pearson Education Limited.
☆ Practically oriented description of concordances and their use in lexicology and lexicography.

Sinclair, J (2004) *Trust the text. Language, corpus and discourse*. Routledge. Edited with Ronald Carter.
☆ Develops Sinclair's idea that "the analysis of… naturally occurring texts… and, in particular, computer processing of texts have revealed quite unsuspected patterns".

Singleton, D (2000) *Language and the lexicon. An introduction*. Arnold.
☆ A comprehensive and accessible guide to how/where the lexicon interfaces, or how it fits in, with other aspects of language.

Spencer, A & Zwicky, A M (1998) *Handbook of morphology*. Blackwell.
☆ Serious volume of writings by well-known experts, covering the whole of morphology.

Štekauer, P & Lieber, R eds (2005) *Handbook of word-formation*. Springer.
☆ Serious treatment of all aspects of word formation in English.

Sterkenburg, P ed (2003) *A Practical Guide to Lexicography*. John Benjamins Publishing Company.
☆ Written by various experts, it is a rather badly edited book; read it if you are not too particular about typos (among other things). Its saving grace is that there are few of its type around.

Stockwell, R & Minkova, D (2001) *English words. History and structure*. CUP.
☆ Mostly about the learned vocabulary of English, has a strong historical bias, is technical accordingly.

Todd, L (1987) *An introduction to linguistics*. Longman York Press.
☆ Relevant parts: Ch 4: Morphology, Ch 5: Lexicology, Ch 7 Semantics.

Yule, G (2006) *The study of language. An introduction*. 3rd ed. CUP.
☆ Accessible introduction to language. Relevant parts: Words and word formation processes & Morphology.

Zgusta, L (1971) *Manual of lexicography*. Janua Linguarum. Series maior 39. Prague: Academia / The Hague, Paris: Mouton 1971. (in cooperation with V. Cerny)
☆Ancient but comprehensive, still considered by some a standard in the field, it is what the title says.

Dictionaries referenced

CALD (2008) *Cambridge Advanced Learner's Dictionary*. 3rd (CD-ROM) edition.

LDOCE *Longman Dictionary of Contemporary English*. CD-ROM. Updated 4th ed. Longman.

MED (2008) *Macmillan English Dictionary for Advanced Learners*. CD-ROM. 2nd ed. Macmillan Publishers.

RHWUD (1999): *Random House Webster's Unabridged Dictionary*. CD-ROM version 3.0. Random House Inc.

TESz (1967–1984) *A magyar nyelv történeti-etimológiai szótára*. Vols I-II-III. Akadémiai Kiadó.