What Is a Word?

Since the beginnings of grammatical study in Europe, the concept of a **word** has been considered to be of central importance. But there are several ways of defining words, and these are not equivalent. We therefore need to examine the several definitions, and to understand the differences among them. Otherwise, we will have no chance of providing sensible answers to such questions as the following: How many words are there in English? Are English *dog* and *dogs* the same word or different words?

It is usual to distinguish four definitions of 'word'. We will look at each of these first from the point of view of English. Then we will examine some data from other languages. Finally, we will look at some special cases.

1. Orthographic words. An **orthographic word** is a written sequence which has a white space at each end but no white space in the middle.

If you look at the paragraph above, you will find it easy to pick out the orthographic words: *Since*, *the*, *beginnings*, *of*, and so on. Very obviously, orthographic words exist only in written texts, and they have no existence in speech.

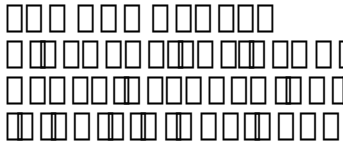
Consider the sequence *ice cream*. As we will see later, this is in some important respects a single English word, but nevertheless it is two orthographic words, not one. If we wanted to, we could write it as *ice-cream*, in which case it would be one orthographic word, but this spelling is not usual.

Such fluctuation is by no means rare in English, however. We might choose to write any of *land owners*, *land-owners* or *landowners*. All three forms have the same meaning, but the first is two orthographic words, while the second and third are one orthographic word each. We encounter both *pocket knife* and *pocketknife*, *common sense* and *commonsense*, *toy shop* and *toyshop*, and hundreds of others of the same kind. The rules of English orthography simply do not specify which compounds should be written with a white space and which not, and individual preferences vary considerably.

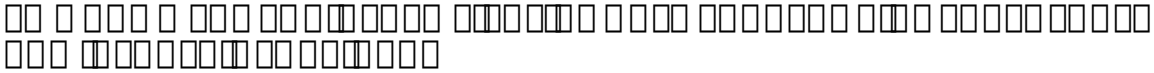
Of course, this freedom to choose where to put white spaces is far from absolute. In most cases, the rules of our orthography dictate where the white spaces should go, and failure to conform produces manifest illiteracy.

However, orthographic words are generally of very little linguistic interest. They are important in learning to write a language in an educated manner, and they may be of interest in a linguistic study of writing systems. But, for most linguistic purposes, orthographic words are irrelevant.

What about other languages? The concept of orthographic words can be readily extended to any writing system which uses white spaces more or less as we use them in English. But not all writing systems do this. Most ancient languages, including Phoenician, Greek and Latin, were very often written with no white spaces or other dividers. Here is an example from Greek.



If this passage were divided into words in the way that later became conventional in Greek, it would look like this:



You can see that the concept of orthographic words is of no use with such a writing system. The idea of separating words was slow to take hold among the ancients, but eventually some peoples began to write their languages with dots or strokes to separate the words. Here is an example from Oscan, an ancient language of Italy (the original text was written right-to-left, but we'll ignore that here):

STATUS•PUS•SET•HURTIN•KERRIIIN•VEZKEI•STATIF•...

Here the dots are doing the same job as the white spaces in English, and so we can usefully talk about orthographic words in this text.

2. Phonological words. A **phonological word** is a piece of speech which behaves as a unit of pronunciation according to criteria which vary from language to language.

In English, the most useful criterion is this one: a phonological word contains only one main stress. Consider the following sentence, as you would pronounce it in relaxed colloquial speech:

The rest of the books'll have to go here.

There are five main stresses here, falling on the words *rest*, *books*, *have*, *go* and *here*. This sentence therefore contains five phonological words. One obvious way of breaking up the utterance into phonological words is as follows:

[*the rest*] [*of the books'll*] [*have to*] [*go*] [*here*]

(You might argue for [*'ll have*] [*to go*], but we won't pursue that issue here.) You can see that not all of these units correspond to units that we might want to recognize for other purposes: for example, *of the books'll* is certainly not a unit of grammatical structure, but we nevertheless pronounce it as a single phonological word.

In other languages, phonological words may be identified by other criteria. Turkish, for example, has a phenomenon called **vowel harmony**, by which a single phonological word is permitted to contain only certain sequences of vowels. As a consequence, each Turkish suffix exhibits multiple forms differing in vowel quality. For example, the suffix meaning 'in' appears as either *-de* or *-da*, according to the requirements of harmony: *ev* 'house' forms *evde* 'in the house', but *oda* 'room' forms *odada* 'in the room'. And the suffix meaning 'my' has four forms: *ev* 'house' gives *evim* 'my house'; *at* 'horse' gives *atım* 'my horse'; *göz* 'eye' gives *gözüm* 'my eye'; and *top* 'gun' gives *topum* 'my gun'.

Phonological words are important in the study of pronunciation, but they are irrelevant to the study of grammar.

3. Lexical items. A **lexical item** (or **lexeme**) is an abstract unit of the lexicon (vocabulary) of a language, with a more or less readily identifiable meaning or function.

A lexical item is a word in the sense in which a dictionary contains words. It is also the sense of 'word' in cases like "How many words are present in the vocabulary of an ordinary English-speaker?" and "I learned twenty new words of French today."

A lexical item is an abstract unit, and it must be represented in speech or writing by one of the possibly several forms it can assume for grammatical purposes. For example, if we want to mention canine animals, we must use either the singular form *dog* or the plural form *dogs*. But these two grammatical forms both represent the same single abstract unit, the same lexical item. We can conveniently represent that lexical item as DOG. Then *dog* and *dogs* are the two possible forms of the lexical item DOG.

In the same way, we recognize a lexical item TAKE, which can be represented by any of the five grammatical forms *take*, *takes*, *took*, *taken* and *taking*.

A dictionary provides entries for lexical items. So, for example, we do not expect to find separate entries in the dictionary for *dog* and *dogs*: we expect to find only one entry for the lexical item DOG. Likewise, we expect to find only one entry for the lexical item TAKE, and not five entries for the five forms of that lexical item.

(A small complication. Some dictionaries have a policy of providing entries for irregular forms like *took*, but an entry for one of these is purely a cross-reference to the main entry. So, the entry for *took* will say merely "see *take*".)

In English, the lexical item BE, uniquely, has eight different forms, but no other English lexical item has more than five, and most have fewer than this – sometimes only one. However, in grammatically more elaborate languages, such as French, Russian, Latin and Arabic, a lexical item may have several dozen different forms. And there exist languages in which certain lexical items (usually verbs) can have several hundred forms.

3.1. Citation forms. We need to talk about lexical items far more often than about other kinds of words. But we have a problem. A lexical item is an abstract unit, and it must always appear in one or another of its various forms. Therefore, we are obliged to choose one of those forms to represent the lexical item when we want to talk about it. The form we choose is called the "citation form" of the lexical item. The **citation form** (or **dictionary form**) of a lexical item is the particular grammatical form of it which we use in naming it, talking about it, and entering it in a dictionary.

In English, we are exceptionally fortunate. Almost every English lexical item has one form which carries no grammatical marking at all. This form, the **base form**, is the natural choice for the citation form.

For a noun like DOG, we choose the singular *dog* as the citation form, and not the plural *dogs*. For the verb TAKE, we choose the infinitive *take*, and not an inflected form like *taking* or *took*. For the adjective BIG, we choose the positive form *big*, and not an inflected form like *bigger*. For the preposition WITH, there is nothing to talk about: this lexical item has only the single form *with*, and this is therefore its citation form.

A further problem may arise with lexical items which are defective. A **defective** lexical item is one that lacks some of the forms normally exhibited by a lexical item of its class. For example, the nouns FURNITURE and HAPPINESS are defective, since they have no plural forms, while the nouns OATS and POLICE have no singular forms. And the verbs MUST and BEWARE both lack a number of the forms exhibited by most verbs. If one of the missing forms happens to be the one we would ordinarily choose as the citation form, then we are forced to make a different choice. For example, since OATS has no singular, we must choose the plural form *oats* as the citation form.

In other languages, we are not always so fortunate as to find one form of a lexical item which carries no grammatical marking at all. For example, in the Romance languages, such as French, Spanish and Italian, a verb has no form which carries no marking at all: every verb form in these languages bears a grammatical marking. We are therefore obliged to choose as the citation form one of these grammatically-marked forms. In all these languages, the conventional choice is the infinitive, which, unlike the English infinitive, bears a grammatical ending. So, for example, the verb meaning 'sing' is cited as French *chanter*, Spanish *cantar* and Italian *cantare*.

The position is much the same in Latin. We might therefore choose the infinitive as the citation form of a verb, and linguists occasionally do this, but, among classicists, the conventional citation form is the first-person singular present indicative active form. So, the verb for 'sing' is cited, not as its infinitive *canere*, but as *cano* 'I sing', and this is the form under which the verb is entered in a Latin dictionary.

Much the same occurs in Arabic, but an Arabic verb has no infinitive. The conventional citation form of an Arabic verb is its masculine third-person singular past-tense form. So, for example, the verb meaning 'write' is cited as *kataba* 'he wrote', and this is the form under which the verb is entered in a dictionary.

4. Grammatical word-forms. A **grammatical word-form** (or **GWF**, or **grammatical form**) is one of the several forms that may be assumed by a lexical item for grammatical purposes.

So, for example, *dog* and *dogs* are grammatical forms of the lexical item DOG, and *take*, *takes*, *took*, *taken* and *taking* are all grammatical forms of the lexical item TAKE. Sometimes a lexical item exhibits only one grammatical form. For example, the lexical items OATS, POLICE, BEWARE and WITH have only one grammatical form apiece: *oats*, *police*, *beware* and *with*. Nevertheless, it may still be useful to distinguish the lexical item OATS from its sole grammatical form *oats*, and likewise for the others.

We have already seen that an English lexical item usually has one grammatical form which carries no grammatical marking at all, its **base form**. All the other grammatical forms of a lexical item, the ones that carry grammatical markings, are the **inflected forms** of that lexical item. So, for example, DOG has the base form *dog* and the inflected form *dogs*, while TAKE has the base form *take* and the inflected forms *takes*, *took*, *taken* and *taking*. But POLICE, BEWARE and WITH have no inflected forms at all, while OATS has *only* the inflected form *oats*. (A complication: since the form *police* is plural, it must bear an abstract inflection for plurality. Some linguists might therefore prefer to say that *police* is an inflected form, even though it carries no overt marking.)

In some analyses, the number of grammatical forms exhibited by a lexical item may be larger than the number of overtly distinct forms. Here is a simple example. Some English verbs exhibit different forms for the past tense and the past participle. The verb TAKE is one of these: *She took the exam*, but *She has taken the exam*. Many verbs, however, present identical forms forms for the past tense and the past participle. An example is FINISH: *She finished the exam*, but *She has finished the exam*. Since the distinction between the past tense and the past participle is essential for English verbs in general, most linguists prefer to say that past-tense *finished* and past-participle *finished* are two *different* grammatical forms of the verb FINISH.

Here is a slightly less obvious example. For any English verb, the present-tense forms (usually apart from the third-singular form) are identical to the infinitive. We can see this with TAKE: *You should take an umbrella* (infinitive *take*), but *I always take an umbrella* (present-tense *take*). Even though there is *no* English verb which exhibits different forms for the infinitive and the present tense, many linguists (not all) would still prefer to say that we are looking at two different grammatical forms here, because of the difference in grammatical behaviour.

We finish with a third example, even more subtle. The *-ing* form of an English verb has at least three different grammatical functions. We will illustrate these with *taking*. First, *taking* can be a participle: *Taking her umbrella, she slipped out of the house.* Second, it can be part of a progressive (continuous) verb-form: *Susie is taking her umbrella.* Third, it can be a gerund: *Taking an umbrella would be a good idea*. Some linguists (again, not all) would prefer to say that all three of these functions represent *different* grammatical forms of TAKE, even though the three forms are always identical for every verb.

These decisions are not matters of absolute truth: they are merely matters of analytical taste and convenience, and linguists vary in their preferences.

By the way, there is one more item of the form *taking*. This is the verbal noun, and it appears in the film title *The Taking of Pelham 1-2-3*. But this *taking* is not a grammatical form of the verb TAKE, and in fact it's not a verb-form at all: it's a noun, and it represents a distinct lexical item TAKING. This noun TAKING is related to the verb TAKE by a derivational process, much as the noun ARRIVAL is related to the verb ARRIVE. See the next section.

[A warning. Quite a few writers, including some who should know better, confuse verbal nouns in *-ing* with gerunds, and apply the label "gerunds" to all of them indiscriminately. This is not a practice you should imitate. The gerund *taking* is a verb-form (for example, it can take an object, as in the example above), and it is an inflected grammatical form of the verb TAKE. However, the verbal noun *taking* is not a verb-form at all, but a noun: it behaves like a noun, and it has no verbal properties at all.]

5. Inflection and derivation. We have seen that a lexical item can appear in several grammatical forms, some of which carry inflections (overt grammatical markings). It is essential to distinguish the process of inflection from the quite different process called derivation.

Inflection is the variation in form of a lexical item for grammatical purposes.

Derivation is the construction of a new lexical item from another lexical item, usually by the addition of an affix (a prefix or a suffix).

Take the lexical item CAT, which has two grammatical forms: *cat* and *cats*. But what about the word *catty*, as in *a catty remark*? How should we classify this?

This is not another inflected form of CAT. Instead, it is a different lexical item, CATTY. How do we know this? For one thing, CAT is a noun, and all its grammatical forms are nouns, while CATTY is an adjective.

The relation between CAT and CATTY is one of derivation: we say that CATTY is derived from CAT by the addition of a suffix.

Since CATTY is a different lexical item from CAT, we expect to find CATTY entered in the dictionary, and our expectations are fulfilled: look at any dictionary of English.

In the same way, DOGLIKE is a different lexical item from DOG, and it too deserves its own entry in the dictionary. A few more examples of derivation by suffixation are HAPPINESS from HAPPY, ARRIVAL from ARRIVE, SINGER from SING, SLOWLY from SLOW, KINGDOM from KING, MOTHERHOOD from MOTHER, NUDISM from NUDE and REDDISH from RED. Some examples involving prefixation are REWRITE from WRITE, UNHAPPY from HAPPY, NON-SMOKER from SMOKER and PRE-WAR from WAR.

Some examples of derivation are irregular and messy, such as DESTRUCTION from DESTROY, CALCULABLE from CALCULATE, COMPREHENSIBLE from COMPREHEND and MEDICAL from MEDICINE.

In principle, we expect to find a separate entry for every lexical item obtained by derivation. In practice, however, dictionaries often cut corners. When a derived lexical item has a regular form and a transparent meaning, as with UNHAPPY, HAPPINESS, REWRITE and SLOWLY, a dictionary may elect to save space by not providing an ordinary entry for the word. A large dictionary may simply list the word without definition, either at the end of the entry for its source word (look for *slowly* under *slow*), or at the bottom of the page (look at the words in *un-* in a big dictionary), while a small dictionary may omit the word altogether. This is done merely to save money, and not to deny that all these words are genuine lexical items which really deserve their own entries.

English exhibits an interesting type of derivation called **zero-derivation**, or **conversion**, by which a lexical item is simply shifted from one word-class to another, without adding any material. For example, the adjective *brown*, as in *brown shoes*, has been converted into a verb, as in *brown the meat*, and the noun *shoe*, as in *new shoes*, has been converted into a verb, as in *shoe the horses*. English allows this with some freedom, and examples can be elaborate. For example, the noun *smoke*, as in *full of smoke*, has been converted into a verb, as in *Susie smokes*, and this verb has been converted into a different noun, as in *Susie was having a smoke*.

In all these cases, we must regard the zero-derived word as belonging to a different lexical item from the source, because it belongs to a different word-class. So, we must recognize a lexical item BROWN_1, which is an adjective, and another lexical item BROWN_2, which is a verb, and similarly for the other cases.

In the examples we have just been looking at, it is obvious which items have been derived from which other items. But the direction of derivation is not always obvious. Take *kiss*. We have a verb KISS, as in *Susie kissed Mike*, and a noun KISS, as in *Susie gave Mike a kiss*. Presumably one of these is derived from the other. But which one is it? This is not obvious, and we may have to make an arbitrary decision.

6. Multi-part and discontinuous words. English and some other languages exhibit certain items which present awkward problems from the point of view of wordhood. These are the items which appear to be words by some criteria, but which exist in two or three pieces, sometimes separated by other material.

The most obvious English examples are the phrasal verbs. A **phrasal verb** consists of a simple verb plus a particle, or rarely two particles. Examples are *make up*, *take off*, *turn on*, *ring up* and *put up with*. A phrasal verb represents two (or three) orthographic and phonological words, but it must usually be considered a single lexical item, since its meaning is often unpredictable from the meanings of its components. Consider the several items of the form *make up*: *She made up a story*; *She made up her face*; *She made up with her boyfriend*; *She made up the numbers*.

Many phrasal verbs can occur discontinuously: *She took off her dress* or *She took her dress off*; *She turned on the light* or *She turned the light on*; and so on.

It appears, then, that we must recognize lexical items which occur in pieces, sometimes in pieces which are separated by other material.

English has another class of awkward verbs, the prepositional verbs. A **prepositional verb** consists of a simple verb (possibly plus an adverb) plus a preposition. Examples are *call on*, *look at*, *take to*, and *look down on*: *We called on Susie*; *Look at the rainbow!*; *I didn't take to our new neighbours*; *Susie looks down on New Agers*. Prepositional verbs differ from phrasal verbs in ways that are explained in reference grammars of English.

In these examples, sequences like *on Susie* and *on New Agers* appear to be syntactic units, prepositional phrases. Nevertheless, it appears that we must recognize *call on* ('visit') and *look down on* ('regard with contempt') as single lexical items, because of their meanings. We are thus faced with conflicting analyses: the item *on* is syntactically part of a prepositional phrase, but lexically part of a multi-part verb.

Examples like these present formidable problems of analysis, and they stretch our conception of words to the breaking point.

Other languages present further examples of multi-part and discontinuous words. Consider Mandarin Chinese. In Chinese, '(the) can' is *guòn*, but 'into the can' is *dào guòn lǐ*, where 'into' is expressed by the discontinuous item *dào...lǐ*.

7. Content words and grammatical words. Lexical items are commonly divided unto content words and grammatical words.

A **content word** (or **full word**, in the Chinese terminology) is a lexical item which has semantic content – that is, which has a readily identifiable meaning.

A **grammatical word** (or **empty word**) has little or no identifiable meaning, but has one or more grammatical functions.

A content word can be defined, and it can often be translated into another language with some ease. English content words include *house*, *dog*, *give*, *big* and *carefully*. An English dictionary will provide reasonable definitions of these items. And we can say that *house* is equivalent to Welsh *tŷ*, French *maison*, Spanish *casa*, Turkish *ev*, Swahili *nyumba* and Japanese *ie*. (But note that, because of lexical differences among languages, not all content words can be translated quite so straightforwardly.)

A grammatical word cannot be defined, and looking for an equivalent in another language is often pointless. English grammatical words include *of*, *the*, *and*, *have*, *who* and *if*. A dictionary cannot provide a definition of one of these, but can only give an account of its grammatical functions: asking for the "meaning" of the word *of* is a waste of time. The English phrase *a bottle of wine* contains the two grammatical words *a* and *of*; its Basque translation *boteila bat ardo* is literally 'bottle a wine', with no equivalent of *of*; and its Welsh equivalent *potel gwin* is literally 'bottle wine', with no equivalent of *a* or *of*.

Some words appear to be borderline in this classification, with clear grammatical functions but some degree of identifiable semantic content, such as *in*, *with*, *we*, *this* and *for*. The best policy is to classify these items as grammatical words, since their grammatical functions are usually much more important than any traces of meaning they may possess.

8. Clitics. A **clitic** is an item which represents a lexical item and a grammatical word-form but which does not make a phonological word by itself.

Since a clitic cannot form a phonological word by itself, it must be phonologically bound to a more substantial item, its **host**, with which it forms a phonological word, possibly together with other clitics.

Among the English clitics are the articles *a(n)* and *the*, which are bound to a following host, as in *a book* and *the man*, and the auxiliary *'ll*, which is bound to a preceding host, as in *John'll do it*.

French possesses a large number of clitics. In the French sentence *Il te le donnera* 'He'll give it to you', the pronouns *il* 'he', *te* 'you' and *le* 'it' are all clitics bound to the following verb.

Turkish has a clitic *de ~ da* 'also, too', which is bound to a preceding item and must agree with that item in vowel harmony: *siz de* 'you too', but *onlar da* 'they too'.

A clitic which precedes its host, like the English articles and the French pronouns, is a **proclitic**. A clitic which follows its host, like English *'ll* and the Turkish item, is an **enclitic**.

Some linguists would suggest that clitics are items whose claim to wordhood is somewhat marginal. From a historical point of view, a clitic is perhaps to be seen as an item which has lost its status as an autonomous word, but which has not yet been reduced to the status of an affix (a prefix or a suffix).

9. Short forms. English exhibits a number of short forms of several kinds, and these short forms merit some discussion from the point of view of wordhood. Among non-linguists, there is a habit of labelling all or most of these short forms "abbreviations", but this practice is misconceived. We will distinguish true abbreviations from the other cases.

9.1. Abbreviations. An **abbreviation** is a short way of writing a word or a phrase, using only letters of the alphabet and possibly full stops. An abbreviation is strictly a written form. It has no pronunciation of its own, and it can only be pronounced by pronouncing the full form which it abbreviates – or, in some cases, by spelling it out letter by letter.

Familiar English abbreviations include *Prof.* for *Professor*, *Sgt* for *Sergeant*, *Dr* for *Doctor* (only as titles accompanying names), *BC* for *before Christ*, *mph* for *miles per hour*, *kg* for *kilograms* and *C* for *degrees Celsius*. Very many of our common abbreviations are derived from Latin: *a.m.* for *in the morning*, *lb.* for *pounds*, *e.g.* for *for example*, *i.e.* for *in other words*, and *etc.* for *and other things*. The exceptional forms *Mrs* and *Ms* are treated as abbreviations, even though no longer forms exist.

An abbreviation consisting of only one or two letters may sometimes be spelled out letter by letter. This is commonplace with *a.m.* ("ay-em") and *BC* ("bee-see"). In other cases, this practice is usual only in reading out a written text, as with *e.g.* ("ee-jee") and *i.e.* ("eye ee"), and saying "ee-jee" or "eye-ee" in speech may be perceived as eccentric.

An abbreviation usually qualifies as an orthographic word, but it is not a word in any other sense. With a few marginal and idiosyncratic exceptions, an abbreviation is not part of the spoken language at all.

9.2. Logograms. A **logogram** is a written character which is not a letter of the alphabet but which conventionally represents a word, or rarely a sequence of words. Familiar logograms include the digits, like '5' for *five*, monetary symbols like '£' for *pounds*, and arithmetical symbols like '+' for *plus*, '=' for *equals* and '%' for *per cent*. Another is '❤' for *love*, as in 'I ❤ NY', but this one is confined to certain informal contexts.

A logogram is a representation of a lexical item or of a grammatical word-form. It fits the definition of an orthographic word given above, but some linguists might prefer to restrict that definition to cases consisting of letters of the alphabet, in which case a logogram would not be an orthographic word.

9.3. Contractions. A **contraction** is a conventional brief way of pronouncing a sequence of two (or rarely three) words which often occur together.

A contraction always has a distinct written form. Typical contractions are *I'm* for *I am*, *it's* for *it is* or *it has*, *she'll* for *she will*, *couldn't* for *could not*, *won't* for *will not*, *hadn't* for *had not*, and (more informally) *she'd've* for *she would have*.

As you can see, some contractions have unexpected forms, such as *won't* for *will not*. And some contractions which were once common have become obsolete or regional, such as *'tisn't* for *it is not*.

A contraction is always a single orthographic word and a single phonological word. But it represents two (or three) lexical items, and two (or three) grammatical word-forms.

9.4. Acronyms and initialisms. In contemporary English, we have become very fond of coining new words in the following way: we begin with a phrase consisting of several words; we extract the initial letters of the most important words in that phrase; and we put the resulting sequence of letters together to form our new word – which almost always has the same meaning as the original phrase.

Here are a few examples:

BBC	*British Broadcasting Corporation*
FBI	*Federal Bureau of Investigation*

NATO or *Nato*	*North Atlantic Treaty Organization*
AIDS or *Aids*	*acquired immuno-deficiency syndrome*

We may distinguish two kinds of outcome here, and there exist national differences in terminology. Sometimes the resulting form can only be pronounced by spelling it out letter by letter, as with *BBC* and *FBI*. In the British tradition, a formation of this kind is an **initialism**. In other cases the resulting form can be pronounced like an ordinary word, as with *NATO* and *AIDS*. In the British tradition, a formation of this kind is an **acronym**.

However, in the American tradition, the label "acronyms" is applied to all such formations without distinction, and the term "initialism" is not used. British and American books will therefore differ in their terminology.

There is a further trans-Atlantic difference. In all varieties of English, initialisms like *BBC* and *FBI* are written entirely in capital letters. The Americans also write most acronyms (in the British sense) entirely in capitals, and so they write *NATO* and *AIDS*. But the British commonly write these things with only an initial capital, and the usual British forms are *Nato* and *Aids*.

But not all acronyms are written with capitals. Some acronyms have become perfectly ordinary lexical items, and they behave accordingly. For example, the phrases *self contained underwater breathing apparatus* and *light amplification by the stimulated emission of radiation* have given rise to the acronyms *scuba* and *laser*. Both of these are ordinary English lexical items, entirely unremarkable apart from their origins.

The formation of *laser* illustrates the canonical pattern of formation, in which small grammatical words are ignored in constructing the acronym. But many acronyms are deliberately constructed in a non-canonical manner, in order to obtain a result which can be easily pronounced. A good example is *radar*, from *radio detection and ranging*.

An initialism or an acronym is a lexical item. An acronym does not differ from any other lexical item, except perhaps in its unusual written form. An initialism has both an unusual written form and an unusual pronunciation, but otherwise it is an ordinary lexical item.

Initialisms are sometimes confused with abbreviations, but they are not abbreviations. First, an initialism always has its own pronunciation, distinct from the pronunciation of the longer form which it represents. Abbreviations do not usually have their own pronunciations. Second, an initialism, being a lexical item, can appear in a structural position in a sentence in which a lexical item is appropriate. For example, we can say or write *the BBC's decision*, in which the initialism *BBC* bears the possesive suffix *-'s*. No abbreviation can behave like this. We may occasionally encounter a written – not spoken – form such as *the sgt's weapon*, but this is not ordinary English orthography: it is an example of **shorthand**, and shorthand is beyond the scope of our discussion.

9.5. Clipped forms. A **clipped form** is an item which is obtained by extracting a piece from a longer word or phrase. The process of extraction is **clipping**.

In all but very rare cases, a clipped form has the same meaning as the longer form from which it is obtained. In contemporary English, we are inordinately fond of clipped forms. We have *gym* from *gymnasium*, *porn* from *pornography*, *flu* from *influenza*, *fridge* from *refrigerator*, *phone* from *telephone*, and *gator* from *alligator*. As these

examples show, any convenient part of a longer word may be clipped. Even discontinuous pieces can be clipped, as with *sci-fi* for *science fiction*, *sitcom* for *situation comedy*, *biopic* for *biographical picture*, and British *maths* for *mathematics* (compare American *math*).

A clipped form is not an abbreviation. It is a genuine lexical item, just like any other lexical item, and it is unusual in no way apart from its origin.

A clipped form accepts the grammatical inflections which are typical of its word class. For example, nouns obtained by clipping can pluralize: *gyms*, *fridges*, *phones*, *gators*. Verbs obtained by clipping can take ordinary verbal inflections. For example, the noun *disrespect* has given rise to a noun *diss*, which in turn has yielded a verb *diss* 'treat with disrespect', and this verb behaves like any other verb: *He was dissing me.*

A clipped form can enter into compounds like any other lexical item: *gym shoes*, *porn star*, *fridge magnet*, *phone book*. In some cases, the longer form would be abnormal or impossible in the same position: who would say ??*gymnasium shoes* or ??*pornography star*?

Clipped forms are entered in dictionaries like other lexical items, and they are legal in Scrabble™.

Sometimes a clipped form displaces its original longer form. For example, the clipped forms *piano* and *bus* have completely supplanted their sources, *pianoforte* and *omnibus*, and *mob* has displaced its (Latin!) source *mobile vulgus* 'the fickle crowd'. And the words *bra* and *cello* are so close to replacing their source words *brassière* and *violoncello* that many people do not even know the longer forms.

Note the following pair of examples. As observed above, English has an abbreviation *Prof.* for *Professor*, as in the written form *Prof. Chomsky*. But we also have a lexical item *prof*, obtained from *professor* by clipping, as in *The Physics Department plans to hire two new profs.* The abbreviation *Prof.* and the lexical item *prof* are not the same item at all, and they should not be confused.

10. Do all languages have words? Our intuitions suggest that a word is a unit which is much smaller than a sentence, and that a sentence typically consists of a sequence of words. But this account does not hold straightforwardly for all languages. In some languages, it can be difficult to draw a distinction between sentences and words. Such languages are known as **polysynthetic** languages.

Here is a typical sentence from Yup'ik, an Eskimo language of Alaska:

Kaipiallrulliniuk. 'The two of them were apparently really hungry.'

The written form given here represents the ordinary pronunciation of this sentence. But, at a somewhat more abstract level, we can analyses this sentence into a sequence of structural units, or **morphemes** (if you haven't yet learned about morphemes, you will be meeting them soon):

kaig- 'be hungry'
-piar- 'really'
-llru- PAST
-llini- 'apparently'

-*u*- INDICATIVE
-*k* ‘they two’

The sentence consists of a verb stem *kaig*- ‘be hungry’ followed by a string of suffixes. In effect, then, the whole sentence is merely a grammatical form of this verb. As you can see, the forms of some of these morphemes are altered when they are joined in sequence.

In a sentence of this kind, there is nothing that we can call a word, except for the sentence itself. And such sentences can be formidably long:

Ayaqaqucuaryuumiitqapiallruyugnarquq-qaa
‘I guess she probably didn’t really want to go for those short little trips, did she?’

We might wonder whether Yup’ik has any words at all, apart from its sentences. But not all Yup’ik sentences are like these examples. Some sentences consist of shorter units which we can reasonably call words:

Maurluqa ayunek pitlallruuq waten amllerivkenaki qillertaqluki enemun agartaqluki.
‘My grandmother used to pick Labrador tea leaves, just a few like this, and tie them
together and hang them inside the house.’

The words here may be glossed as follows:

maurluqa ‘my grandmother’
ayunek ‘Labrador tea leaves’
pitlallruuq ‘used to pick’
waten ‘like this’
amllerivkenaki ‘just a few’
qillertaqluki ‘and tie them together’
enemun ‘inside the house’
agartaqluki ‘and hang them’

So, Yup’ik does have words after all, though you can see that even the shortest Yup’ik words appear to be more complex than most English words.