3 ENGLISH MORPHOLOGY

Information on English Morphology is available with lemma lexicons and wordform lexicons. If you are interested in inflectional morphology, then you should use a wordforms lexicon, and if you are interested in derivational and compositional morphology, you should use a lemma lexicon.

3.1 MORPHOLOGY OF ENGLISH LEMMAS

The morphological analyses given for lemmas in the CELEX databases always use the *headword* form of the lemma, because this form (unlike Dutch) is usually the shortest in any inflectional paradigm, without any visible inflectional endings. However, when discussing English morphology, *stem* is the normal term used to describe this form, and so in this section *stem* is used instead of headword, just to fit in with common practice.

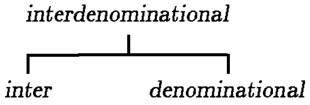
Before finding out details about each of the columns available, you should look at the sections below which try to give some explanation of the methods used to obtain the analyses given in the database. You will then know what CELEX means by terms such as *immediate segmentation, hierarchical segmentation, compound, derivation*, and *derivational compound*. After all that, you'll understand more clearly what each of the various columns has to offer.

3.1.1 HOW TO SEGMENT A STEM

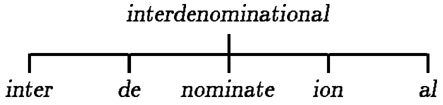
The first and most fundamental type of segmentation is *immediate segmentation*. This simply involves splitting a stem into its largest constituent parts. If you continue to carry out immediate segmentation until there is nothing left to segment, you arrive at the stem's *complete segmentation*. Depending on your requirements, you can look at a complete segmentation in two forms. The first is the *flat* form, which shows every morpheme that makes up the stem. The second is the *hierarchical* form, which, as well as pointing out the individual morphemes in a stem, also shows all the analyses which have to be made to identify those morphemes. The flat segmentation gives the conclusion reached, while the hierarchical segmentation shows the working.

To illustrate the three types of segmentation, take as an example the word *interdenominational.*

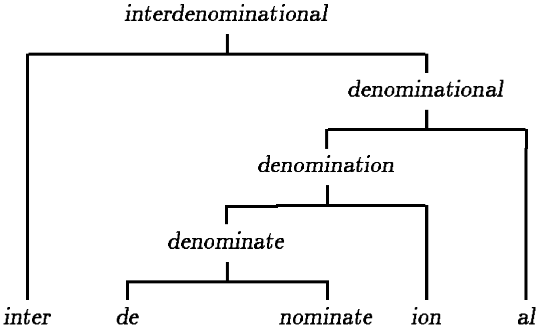
The first type of analysis 'Immediate segmentation' gives the affix *inter* plus the stem *denominational*:



The second type of analysis 'complete segmentation (flat)' shows you what you get if you keep applying immediate segmentation, namely the constituent morphemes of *interdenominational*: the affix *inter* plus the affix *de* plus the stem *nominate* plus the affix *ion* plus the affix *al.*



The third type 'complete segmentation (hierarchical)' shows you the full analysis of the word, including each individual immediate segmentation carried out. It gives you enough information to produce a hierarchical tree diagram like this one:



For most stems in the database, representations of each of these three types of segmentation are available. Sometimes there is more than one representation, because certain stems can have more than one immediate segmentation. To explain this fully, the next section describes the basic analyses that result from immediate segmentation.

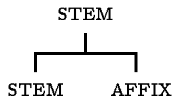
3.1.2 TYPES OF ANALYSES

When you attempt to split a stem into its biggest component parts, the result is always some combination of *stems* or *flections* and *affixes*. A flection—such as the *freezing* in *freezing point*—is treated the same as a stem, so that whenever an analysis involves a stem, you know that the stem could also be a flection. The most straightforward analysis of all is a stem which consists of only one (free) morpheme: it is *monomorphemic*, and clearly can't be split up. Every other stem, however, consists of one smaller stem or affix plus at least one affix or one other stem, and can be termed either a *Derivation*, a *Compound*, a *Derivational Compound*, or a *Neo-classical Compound*. It is important to understand the differences between these four terms, since they are at the heart of the morphological information CELEX provides. So, in the subsections below, each is defined in terms of stems and affixes. Examples are given, and simple 'tree' diagrams illustrate the appropriate immediate analyses.

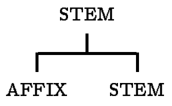
3.1.2.1 THE DERIVATION

A DERIVATION involves affixation, whereby affixes can be added to an existing stem or flection to form a new stem. The immediate analysis always takes one of four possible forms:

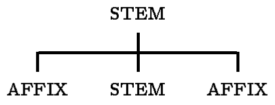
(i) a binary split into a stem or flection plus an affix (the word *careful* for example: *care* + *ful*).



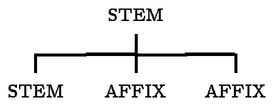
(ii) a binary split into an affix plus a stem or flection. For example, the word *barometer* is analysed as *baro* + *meter*.



(iii) a triform split into an affix, a stem or flection, and an affix (the word *extracurricular* for example: *extra* + *curriculum* + *ar*).



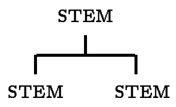
(iv) a triform split into a stem or flection, an affix and another affix. Such words can be derivations of inflected forms – the word *falteringly*, for example, is analysed as *falter* + *ing* + *ly*, which is a stem plus an inflectional affix plus a derivational affix. Alternatively, they can be lexicalised forms of inflected derivations like *countrified*, analysed as *country* + *ify* + *ed*, which is a stem plus a derivational affix plus an inflectional affix. This sort of analysis is only appropriate when the stem and the affix which immediately follows it don't together form a lemma, because otherwise—as with the word *inflationary*—the immediate analysis would be like type (i) above, a stem plus an affix (*inflation* + *ary*).



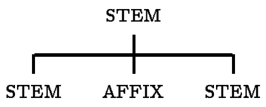
3.1.2.2 THE COMPOUND

A compound is the joining of two stems or flections into one new stem. The immediate analysis always takes one of two forms:

(i) a binary split into two stems (the word *nameplate* for example: *name* + *plate*).



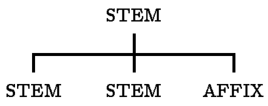
(ii) a triform split into a stem or flection, an affix (simply a 'link' morpheme), and a stem or flection (the word *bandsman* for example: *band* + *s* + *man*).



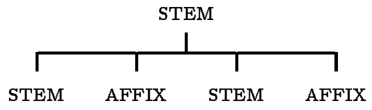
Words which consist of more than two stems aren't analysed as compounds, since they normally have the structure of a phrase or sentence. So headwords like *nevertheless*, *Australian Rules football* and *be-all-and-end-all* don't get a morphological analysis.

3.1.2.3 THE DERIVATIONAL COMPOUND

A derivational compound is a compound which can only be formed in combination with a derivational affix (as opposed to a simple link morpheme). The immediate analysis normally takes the form of a triform split into a stem or flection, another stem or flection, and an affix (the word *icebreaker* for example: *ice* + *break* + *er*).

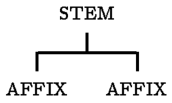


A couple of words can be analysed as a quaternary split into a stem, an affix, a stem, and an affix (the word *brinksmanship*, for example, is analysed as *brink* + *s* + *man* + *ship* and *whippersnapper* is analysed as *whip* + *er* + *snap* + *er*). However this is a very rare form of analysis.



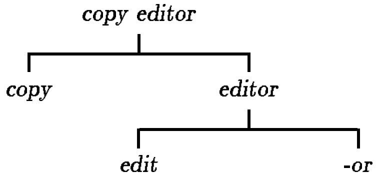
3.1.2.4 THE NEO-CLASSICAL COMPOUND

A NEO-CLASSICAL COMPOUND is a word which appears to be made up of two affixes, neither of which can occur as a word in its own right, like *aerodrome* (*aero* + *drome*) or *neurology* (*neuro* + *ology*). The affixes which combine to form this type of compound are generally known as *combining forms*.

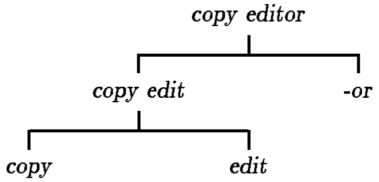


3.1.2.5 THE NOUN-VERB-AFFIX COMPOUND

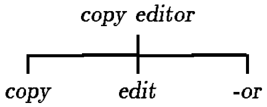
Problems sometimes arise with the analysis of words which look like derivational compounds. The general definition of a derivational compound is normally sufficient, but when the second stem is a verbal form, things become more complicated. A stem which comprises a noun plus a verb plus an affix can normally be considered a derivational compound, but some people may want to treat it as an ordinary compound or derivation. The distinction is important, since it can affect not only the appearance of a single immediate segmentation branch, but also the appearance of a complete hierarchical tree. The stem *copy editor* is such a 'problem' compound. If you consider it to be an ordinary compound (the stem *copy* plus the stem *editor)*, its complete hierarchical tree looks like this:



If you consider it to be an ordinary derivation (the stem *copy-edit* plus the affix *-or)*, its complete hierarchical tree looks like this:



But if you consider it to be a derivational compound, the first immediate segmentation gives you the stem *copy* plus the stem *edit* plus the affix *-or*, which gives the full hierarchical tree a different appearance:



3.1.3 HOW TO ASSIGN AN ANALYSIS

When you're faced with a headword that needs to be analysed, how do you work out the correct analysis? How did the people at CELEX who carried out the morphological analysis by hand arrive at the answers contained in the database? In particular, in the case of noun-plus-verb-plus-affix words, how did they decide which of the analysis types discussed in the previous section were appropriate?

To illustrate the principles used in analysing the information, there are two diagrams, given as Tables 8 and 9 below. The first illustrates the general strategy adopted for each headword, and the second deals with the special problems that arise with noun-plus-verb-plus-affix words. In both diagrams abbreviations are used: `S` means *stem*, and `A` means *affix*, making it easy to refer back to the sections above which define derivations, compounds, and derivational compounds in terms of stems and affixes. When an analysis is *acceptable*, it means that the component parts identified are current stems or affixes, and that the word can be defined as a derivation, a compound, or a derivational compound according to the definitions given in sections 3.1.2.1–3.1.2.3 above. An acceptable stem is one which appears in the *Collins English Dictionary* without being marked as 'obsolete' or 'archaic'.

Following the first diagram, analysis starts with an attempt to see if the word under scrutiny is just the same as an already existing word with a different word class. The word *railroad*, for example, can be used as a verb, and it is said to come from the corresponding noun *railroad*. This phenomenon is called *conversion* or *zero derivation*, since there is no difference in the form of the two words even though they have a different word class. Conversion is explained in full under section 3.1.4 'Status and Language codes'. If conversion has occurred, the analysis need go no further: in ***MorphStatus*** the word gets the code `Z`, and ***NVAffComp*** and its subordinate columns ***Der, Comp,*** and ***DerComp*** are all set to `N`.

If the word is not a conversion, then the next step is to check whether it fits with the definition of a *derivation* given in section 3.1.2.1 above. For example, the word *calculator* is analysed as the stem *calculate* with the suffix *-or*, *encircle* is analysed as the prefix *en-* with the stem *circle*, and *unflappable* as the prefix *un-* plus the stem *flap* plus the suffix *-able*. In all three cases, the word is classified as a *derivation*: in ***MorphStatus*** the word gets the code `C` to indicate that it is *complex*, and ***NVAffComp*** and its subordinate columns ***Der, Comp,*** and ***DerComp*** are all set to `N`.

If the word turns out not to be a derivation, then the next stage is to see if it fits with the definition of a *compound* given

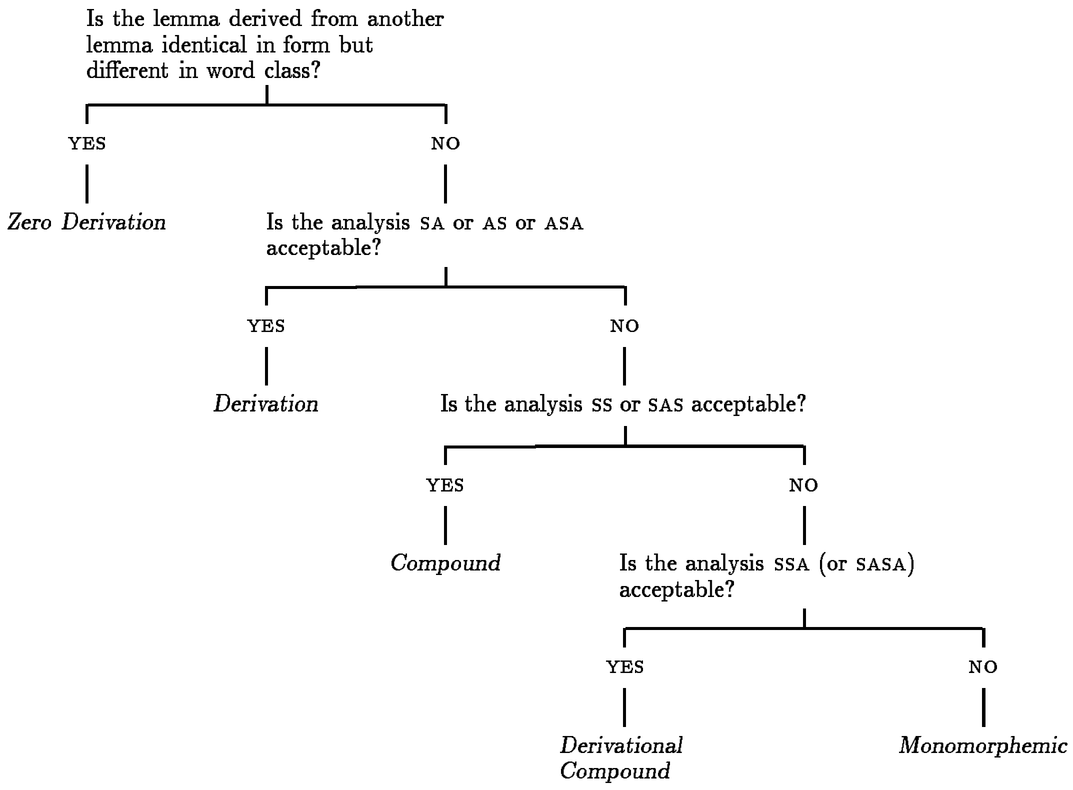

Is the lemma derived from another lemma identical in form but different in word class?
YES
NO
Zero Derivation
Is the analysis SA or AS or ASA acceptable?
YES
NO
Derivation
Is the analysis SS or SAS acceptable?
YES
NO
Compound
Is the analysis SSA (or SASA) acceptable?
YES
NO
Derivational Compound
Monomorphemic

Table 8: How to carry out morphological analysis

in section 3.1.2.2 above. For example, the noun *keyboard* is analysed as the stem *key* plus the stem *board*, and *groundsman* as the stem *ground* plus the infix *-s-* plus the stem *man*. In both cases, the word is classified as a *compound*: under the ***MorphStatus*** column it gets the code `C` to indicate that it is *complex*, and ***NVAffComp*** and its subordinate columns ***Der, Comp,*** and ***DerComp*** are all set to `N`.

If the word still hasn't been classified, then the last stage is to see whether the word is a *derivational compound*, as defined in section 3.1.2.3 above. For example, the adjective *barefaced* is analysed as the stem *bare* plus the stem *face* plus the affix -ed. The word is therefore classified as a *derivational compound*: under ***MorphStatus*** it gets the code `C` to indicate that it is *complex*, and ***NVAffComp*** and its subordinate columns ***Der, Comp,*** and ***DerComp*** are all set to N.

It is possible that the word might not fit into any of the above categories: this makes the word *monomorphemic*, and the 'analysis' is simply the word itself – *chair*, for example, or *llama*. Under ***MorphStatus*** the code M is given, and ***NVAffComp*** and its subordinate columns ***Der, Comp,*** and ***DerComp*** are all set to N. In other cases where no analysis can be carried out, the code under ***MorphStatus*** indicates why. You can read about these codes in section 3.1.4 'Status and language codes'.

3.1.3.1 THE NOUN-VERB-AFFIX COMPOUND

The general scheme explained above is enough to arrive at an analysis in most cases. However, difficulties in applying the system arise when you start considering so-called *noun-verb-affix compounds* – those words which contain a verbal element, which aren't conversions, and which could be analysed as a nominal stem plus a verbal stem plus an affix. Examples of such words are *stockholder* and *copy-editor*. This type of compound is characterised by a Y in the ***NVAffComp*** column. As the diagram below shows, just because they *could* be analysed in such a way, it doesn't mean they necessarily *should* be. *Stockholder* is both a compound and a derivational compound, and *copy-editor* is a derivation, a compound and a derivational compound. The approach outlined below is designed to keep as many morphologists as possible happy with the information available in the database: it's possible to choose for yourself whether to restrict your lexicon to just one type of analysis, or to permit them all, according to your own requirements.

The first step is to see whether the ***NVAffComp*** word you are dealing with can be classified as a derivation, in accordance with the definition in section 3.1.2.1 above. Take the word *dive-bomber* as an example – it can be analysed as the stem *dive-bomb* plus the affix *-er*. The verb *dive-bomb* is accepted as legitimate because it occurs as such in the *Collins English Dictionary* (CED). So this word does meet the definition of a derivation, and thus gets the code Y in the ***Der*** column. Another example is the word *bricklayer*: since a verb *bricklay* doesn't exist (according to the CED) it can't be analysed as a stem plus an affix, and so it gets the code N in the ***Der*** column.

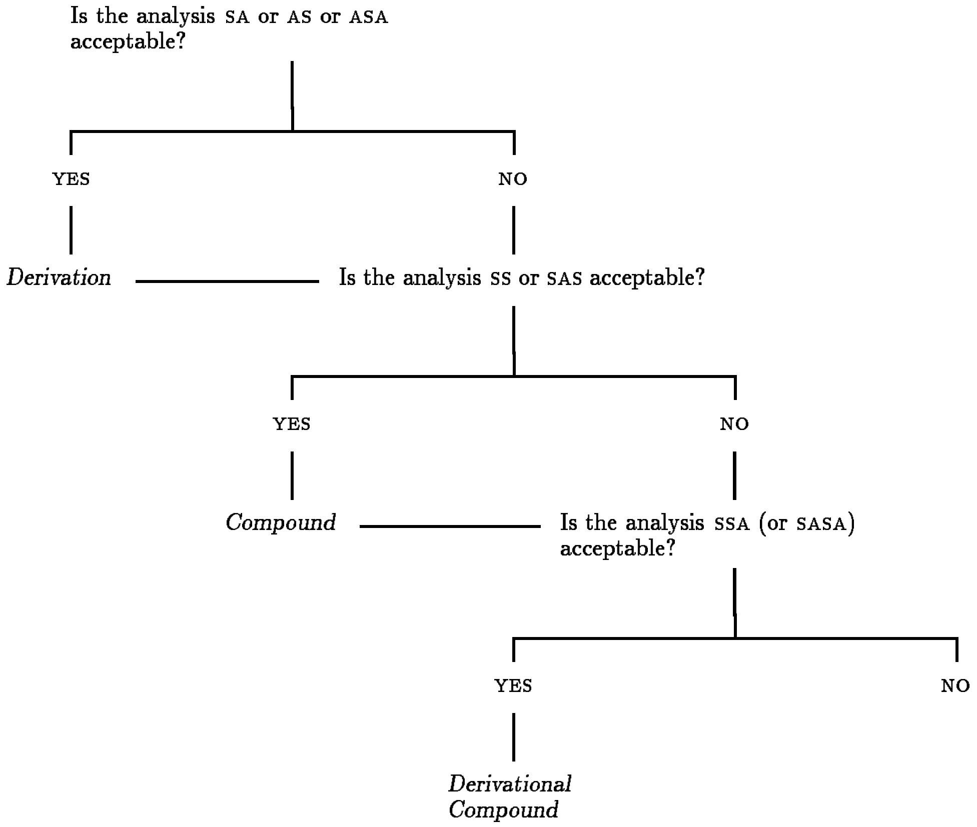

Is the analysis SA or AS or ASA acceptable?
YES
NO
Derivation
Is the analysis SS or SAS acceptable?
YES
NO
Compound
Is the analysis SSA (or SASA) acceptable?
YES
NO
Derivational Compound

Table 9: Dealing with noun-verb-affix compound analyses

The next stage is to see whether the ***NVAffComp*** word in question can be classified as a compound, as defined in section 3.1.2.2 above. Again, the word *dive-bomber* meets the definition: it can be analysed as the stem *dive* plus the stem *bomber* and it is a particular sort of bomber. It can therefore be called a compound, and it gets the code `Y` in the ***Comp*** column. Applying the same rules to *bricklayer* produces the opposite result: a *bricklayer* isn't really a particular sort of *layer*, since (according to the CED) *layer* doesn't mean 'someone who lays things'. So, *bricklayer* gets the code `N` in the ***Comp*** column to show that it isn't a compound.

The last stage is to decide whether the word is a derivational compound, as defined in section 3.1.2.3 above. This time *dive-bomber* does not qualify. Even though it can have the structure stem (*dive*) plus stem *bomb* plus affix (-er), the noun *dive* cannot be the object of the verb *bomb* – you can't talk about 'bombing a dive'. Since *dive* isn't any sort of obligatory complement to the verb *bomb*, *dive-bomber* is not a derivational compound, and therefore gets the code N in the ***DerComp*** column. Applying the rules to *bricklayer* again produces the opposite result. It can have the structure stem (*brick*) plus stem (*lay*) plus affix (*-er*), and *brick* is the object of the verb *lay*; it is possible to talk about 'laying bricks'. So since this time *bricks* is some sort of complement to the verb *lay*, *bricklayer* gets the code Y in ***DerComp*** column to show that it is a derivational compound.

To illustrate this process further, more examples are given in the table below.

Word	Classifications			
	NVAffComp	***Der***	***Comp***	***DerComp***
typesetter	Y	Y	N	Y
dive-bomber	Y	Y	Y	N
copy-editor	Y	Y	Y	Y
stockholder	Y	N	Y	Y
churchgoer	Y	N	N	Y
cub reporter	Y	N	Y	N

Table 10: Example noun-verb-affix compound analyses

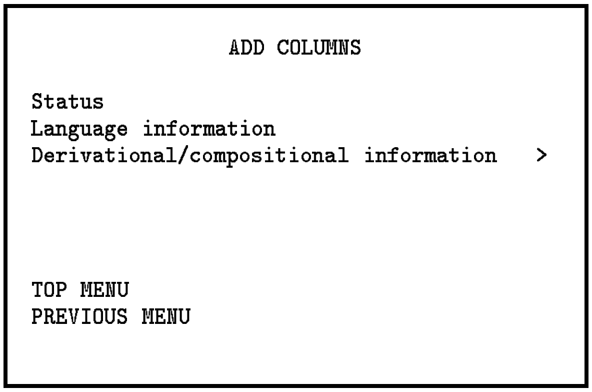
It shows six examples and the codes each one gets in ***NVAffComp***, ***Der***, ***Comp***, and ***DerComp***, as a quick way of showing how the words are classified in the ***NVAffComp*** analysis scheme. In the database, however, each separate *analysis* gets a separate row, and if you looked up analyses for these six words you would get something like this:

Stem	MorphNum	NVAffComp	Der	Comp	DerComp	Def	Imm
typesetter	1	Y	Y	N	N	Y	typeset+er
typesetter	2	Y	N	N	Y	Y	type+set+er
dive-bomber	1	Y	Y	N	N	Y	dive-bomb+er
dive-bomber	2	Y	N	Y	N	Y	dive+bomber
copy-editor	1	Y	Y	N	N	Y	copy-edit+or
copy-editor	2	Y	N	Y	N	Y	copy+editor
copy-editor	3	Y	N	N	Y	Y	copy+edit+or
stockholder	1	Y	N	Y	N	Y	stock+holder
stockholder	2	Y	N	N	Y	Y	stock+hold+er
churchgoer	1	Y	N	N	Y	Y	church+go+er
cub reporter	1	Y	N	Y	N	Y	cub+reporter

If you've followed in full the explanation of how CELEX carried out its morphological analysis of English words, most of this example lexicon should be clear. The columns it contains, along with other columns are described and defined in the sections that follow. Using the columns available you can control the number of analyses you see for each stem, as well as the type of analyses, by means of restrictions on the 'number' and 'status' columns which are defined below. You can decide for yourself whether your lexicon should contain just one 'default' analysis per stem, or whether it should contain more than one analysis per stem. In cases where a stem can be analysed as a derivation, a compound or a derivational compound, you can choose to include whichever type you prefer, leaving out the other type. In short, you have the freedom to build lexicons which contain morphological information in the form you most prefer.

3.1.4 STATUS AND LANGUAGE CODES

The first `ADD COLUMNS` menu you see after you select the 'Morphology' option is this one:



```
                ADD COLUMNS

Status
Language information
Derivational/compositional information   >




TOP MENU
PREVIOUS MENU
```

Before dealing with the various derivational/compositional information columns, which form the bulk of the available morphological information, the first two columns are dealt with here.

The first column simply tells you by means of a single code whether each stem is morphologically simple, morphologically complex, or a conversion, or why it is as yet unanalysed. The table below shows the codes that are used, and it is followed by a description of each of the eight codes. Just before the description concludes with the column definition, there is a diagram which illustrates the strategies CELEX used to determine a status code for each stem.

Status	Code	Example
Morphological analysis available:		
Morphologically complex	C	*sandbank*
Monomorphemic	M	*camel*
Conversion (Zero Derivation)	Z	*abandon*
Contracted form	F	*I've*
Morphological analysis unavailable:		
Morphology irrelevant	I	*meow*
Morphology obscure	O	*dedicate*
Morphology may include a 'root'	R	*imprimatur*
Morphology undetermined	U	*hinterland*

Table 11: Derivational morphology status codes

If a stem contains at least one stem plus at least one other stem or affix, then it is said to be morphologically complex. Details of how the stem can be analysed are given in the derivational/compositional segmentation columns described in the section below. Thus if a stem has the morphological status code C for 'complex', you know that information about its derivational and/or compositional morphology are available in the database.

If a stem is monomorphemic, then it contains only one morpheme, and no further analysis is required. The morphological status code M means 'monomorphemic', and you know that a simple one-stem analysis is given as the derivational and/or compositional morphology for each stem with this code.

If a stem appears to be derived from another stem which is identical in form but different in word class, it gets the code Z for 'zero derivation' or conversion. The noun *delinquent*, for example, can be said to derive from the adjective *delinquent*. Normally derivations from one word class to another are clearly marked by means of an affix – *sheepish* is an adjective derived from the noun *sheep*, for example. But conversions, on the other hand, are not so marked: it's as if an affix containing nothing had been added to the original stem.

Naturally enough, when conversion occurs, it's not immediately obvious which stem is the original and which is the derivative. In analysing these words, CELEX adopted a strategy for determining the *direction* of the conversion (that is, if a verb has been converted into a noun, the derivation is *in the direction* of the noun). Table 12 indicates the normal direction of conversion. Conversion in the opposite direction is also possible provided that it is specified in the *Shorter Oxford English Dictionary* (SOED).

Default direction	Example
VERB—NOUN	*paint*
ADJECTIVE—NOUN	*parallel*
ADJECTIVE—ADVERB	*pretty*
ADJECTIVE—VERB	*pale*
PREPOSITION—ADVERB	*past*

Table 12: Direction of conversions

The status code F indicates that the 'analysis' given is in fact a contraction. The single contraction *'d* can represent *had, would* or *did*, and the complex contraction *he's* represents *he* *is* or *he has*. Each contracted form gets its own row in the database and the status code `F`.

In the case of monomorphemic stems, complex stems, conversion stems, and contractions of stems, morphological analyses are provided in the various segmentation columns. However there remains a large number of stems which have no analysis, and in such cases, codes indicating the reasons for the lack of analysis are given in this column, and these codes and reasons are explained below.

First of all, sometimes even attempting morphological analysis is not appropriate for a particular stem. Usually this is true when the stem is an exclamation or an interjection of some sort (*gosh*, *prithee* or *meow*, for example), or when it is a proper noun – *Spooner* and *Germany* aren't analysed. In addition, those few words which seem to have taken on the structure of a short sentence (or at least consist of three or more stems), like *nowadays* or *whodunit*, don't get an analysis. So, whenever a stem has the code `I` for 'irrelevant', you know that a morphological analysis isn't considered necessary, and that its entries in the segmentation columns described below are therefore empty.

Some stems are recognizable recent loanwords which have achieved some sort of currency in English – words like *virtuoso* or *pretzel* or *mazurka*. Since providing analyses for such stems would, in many cases, mean delving into the morphology of languages not covered by CELEX, they simply receive the code `U` for 'undetermined'. The languages loanwords originate from are shown in the next column, ***Lang***.

On other occasions, an analysis seems possible, but cannot be fully explained. The stem *tabby*, for instance, appears to consist of the productive suffix *-y* plus what might be another stem *tab*. However *tab* bears no immediate relation to the adjective *tabby*, so that *tabby* gets the code `O` to indicate that the morphological analysis is 'obscure'.

In most cases morphological analysis is carried out on a *synchronic* basis: the stems or affixes which make up a word must occur in modern, current English, regardless of the historical origins they might have. On many occasions, however, an etymological root could explain the morphology of a stem which would otherwise be unanalysable. The stem *patrimony*, for example, appears to be made up of a Latin prefix *patri-* and what may be a Latin suffix *-mony*. Stems like this, which could be analysed on the basis of the historical root of its constituent parts, are given the code `R` for 'root'.

If you want to understand this coding system more fully, then you can examine Table 13 , which is a diagram that sets out a scheme for arriving at appropriate morphological status codes. Starting with a stem whose morphology is relevant (that is, it doesn't belong with those stems that have the code `I`), you can work out the code it should have by following the diagram through. This strategy is the one actually used by CELEX to determine the correct codes.

This column can be used to eliminate from your lexicon stems for which there are no morphological analyses, allowing you to concentrate on those which do. Simply add a restriction which states that you only want stems which are morphologically complex: `MorphStatus = C`.

The column which contains these morphological status codes has the following FLEX name and description:

MorphStatus
(MorphStatusLemma) `Morphological status`

The second column contains codes that identify the particular geographical origins of lemmas, including the foreign languages from which some of the lemmas have been borrowed. The headword *conquistador* thus gets the code `S` to indicate that it is of Spanish origin, and the verb *waffle* gets the code `B` because it's reckoned to be a peculiarly British usage. Whenever a lemma doesn't have a code, it isn't a recent borrowing from another language, and it doesn't have associations with one regional variety of English. Table 14 below sets out all the codes used and their meanings.

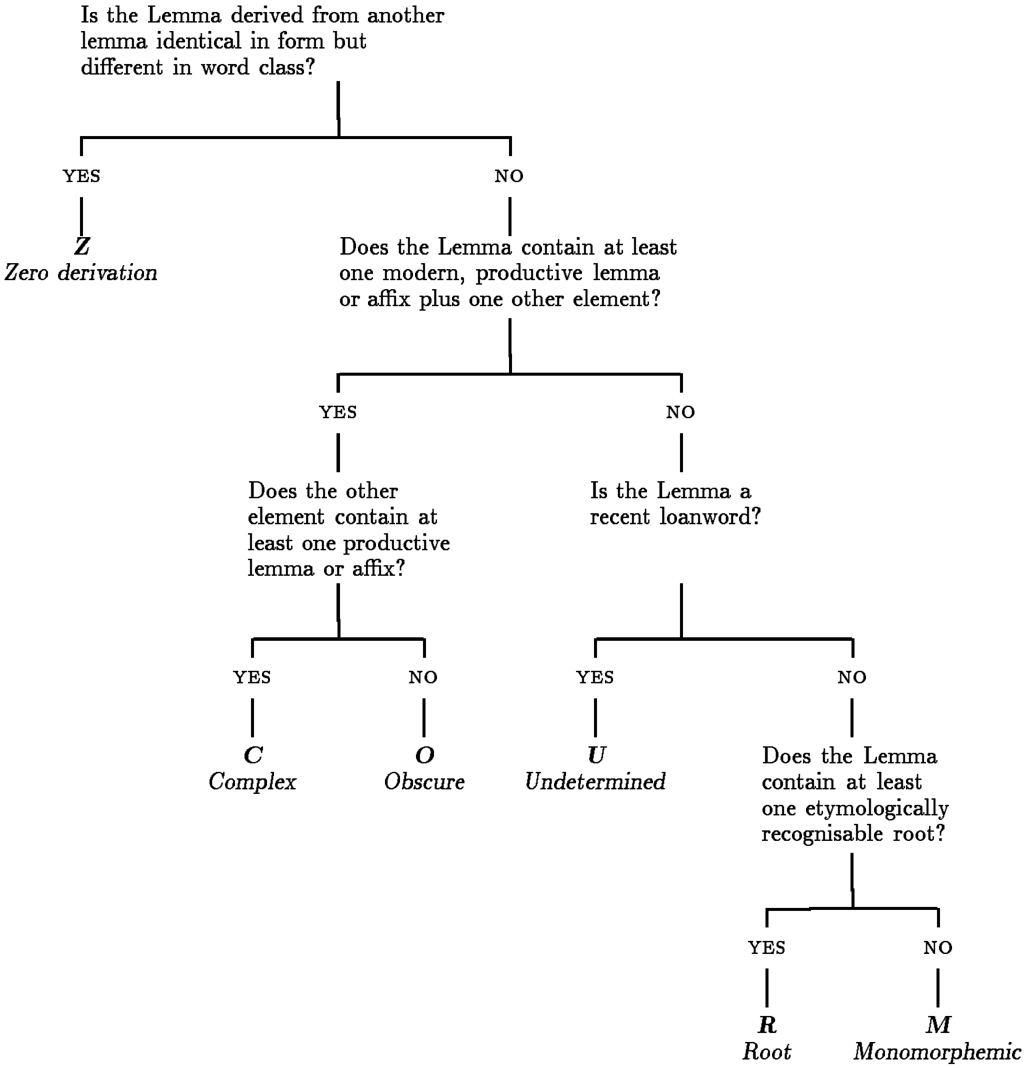

Is the Lemma derived from another lemma identical in form but different in word class?
YES
NO
Z
Zero derivation
Does the Lemma contain at least one modern, productive lemma or affix plus one other element?
YES
NO
Does the other element contain at least one productive lemma or affix?
Is the Lemma a recent loanword?
YES
NO
YES
NO
C
Complex
O
Obscure
U
Undetermined
Does the Lemma contain at least one etymologically recognisable root?
YES
NO
R
Root
M
Monomorphemic

Table 13: How to assign **MorphStatus** *codes*

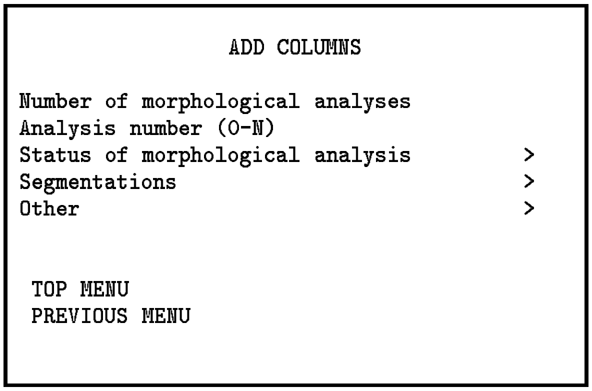
Language Code	Meaning	Example
A	American English	*billfold*
F	French	*patisserie*
B	British English	*divvy*
D	German	*sauerkraut*
G	Greek	*eureka*
I	Italian	*cicerone*
L	Latin	*emeritus*
S	Spanish	*siesta*

Table 14: Language codes for English headwords

The FLEX name and description of the column which contains these codes is as follows:

Lang (LangLemma) Language information

3.2 DERIVATIONAL/COMPOSITIONAL INFORMATION



```
                    ADD COLUMNS

Number of morphological analyses
Analysis number (0-N)
Status of morphological analysis            >
Segmentations                               >
Other                                       >


 TOP MENU
 PREVIOUS MENU
```

These options give you information about the derivational and compositional morphology of *stems*, including how many analyses are available for each stem, a unique number for each analysis, an indication of the way in which each analysis has been made, and a marker for the 'default' analyses for each stem.

The first option is a column which simply indicates how many analyses have been made for each stem. For example, *backfire* has one analysis, *flashbulb* has two, and *treasurer* three. The number of analyses for each stem also equals the number of rows that stem can have with distinct analyses, since each morphological analysis is assigned to its own individual row.

You can use this column to construct restrictions for your lexicon. A simple example would be one that includes in your lexicon only those stems which have more than one analysis. This would take the form `MorphCnt > 1`. The FLEX name and description of this column are as follows:

MorphCnt
(MorphCntLemma)
`Number of morphological analyses`

The second option is a column which identifies each analysis of a particular stem. Each different morphological analysis of a stem is assigned to a different row, and this column gives the number of the row. Thus the adjective lemma *flashbulb* has two rows: one has the ***MorphNum*** 1, the other has the ***MorphNum*** 2 or a stem. description of this column are as follows:

MorphNum
(MorphNumLemma)
`Morphological analysis ID`

3.2.1 ANALYSIS TYPE CODES

Under the 'status of morphological analysis' option there are five 'yes/no'-type columns which, when you use them to construct restrictions, can help you extract the analyses you want from the many stem segmentations available.

Each distinct morphological analysis of each stem has a number, and is given (in several different forms) on its own row in the database. These columns give simple information about each analysis, and are particularly useful whenever a stem is a noun-verb-affix compound. (A noun-verb-affix compound, as discussed in section 3.1.2.5, can correctly be analysed as a derivation, a compound, or a derivational compound.) The five columns in question are called ***NVAffComp, Der, Comp, DerComp***, and ***Def***.

Whenever ***NVAffComp*** contains a `Y`, you know that 'yes, this row contains a stem which is considered a noun-verb-affix compound, and which therefore might be analysed in three different ways'. And naturally whenever it contains an `N`, you know that the row contains a stem which is *not* considered a noun-verb-affix compound. The FLEX name and description of this column are as follows:

NVAffComp
(NVAffCompLemma) `Noun-verb-affix compound`

Whenever ***Der*** contains a `Y`, you know that 'Yes, this row contains a noun-verb-affix compound which is analysed as a derivation'. And whenever it contains an `N`, you know that the row contains a noun-verb-affix compound which is *not* analysed as a derivation or a stem which is not of the noun-verb-affix compound type. The FLEX name and description of this column are as follows:

Der
(DerLemma) `Derivation analysis`

Whenever ***Comp*** contains a `Y`, you know that 'yes, this row contains a noun-verb-affix compound which is analysed as a compound'. And again, `N` means that the row contains a noun-verb-affix compound which *isn't* analysed as a compound or a stem which is not of the noun-verb-affix compound type. The FLEX name and description of this column are as follows:

Comp
(CompLemma) `Compound analysis`

Likewise whenever ***DerComp*** contains a `Y`, you know that 'yes, this row contains a noun-verb-affix compound which is analysed as a derivational compound'. And naturally, `N` means that the noun-verb-affix compound *isn't* analysed as a derivational compound or that it is a stem which is not of the noun-verb-affix compound type. The FLEX name and description of this column are as follows:

DerComp
(DerCompLemma) `Derivational compound analysis`

If a stem has more than one analysis, it's sometimes helpful to be able to identify one which is the best or most useful, or at least to discard unwanted alternatives. Whenever ***Def*** contains a `Y`, you know that ‘yes, this row contains a default analysis’, and when it contains an `N`, you know that the row contains another, non-default analysis.

Since there are three types of analyses which can be assigned to a complex stem, there might also be up to three default analyses for one word: a default derivation analysis, a default compound analysis, and a default derivational compound analysis. (Of course, not many words are eligible for three default analyses.) While morphological analysis was being carried out, rules were formulated to determine which analyses should take precedence over the others, and these rules are explained in Table 15 below.

The left-hand column gives the problem which those doing the analysis came up against, the central column shows which of the two possible analyses should be the default (or ‘take precedence’), and the right-hand column illustrates the principle with an example. The first part of the table shows the preferential order for derivations, and the second part shows the preferential order for compounds. A part for derivational compounds isn’t necessary since they are only analysed in one way.

If, despite the range of analyses available, you only want just *one* default analysis, then you can get it by making a restriction on ***MorphNum***: `MorphNum = 1`. The first analysis for a lemma is always a default analysis. Analyses which are derivations take precedence over compounds, and likewise compounds take precedence over derivational compounds.

Using this column in conjunction with the three preceeding columns, you can construct restrictions which select or omit the analyses you specify. The FLEX name and description of this column are as follows:

Def `Default analysis`
(DefLemma)

To illustrate how you can use these columns, imagine that you have chosen ***Imm*** and ***ImmClass*** as the form of morphological analysis you want to see. ***Imm*** shows the analysis, and ***ImmClass*** shows the word class of the analysed parts (these columns, and the other columns containing the same analyses in different forms, are described in the sections

Option	Solution	Example
Preferential order for the analysis of derivations:		
stem + affix or affix + stem	stem + affix takes precedence over affix + stem	*disavowal* is analysed first as *disavow* + *-al*, then as *dis-* + *avowal*.
verb ending in *-ate, -ete, -ote* or *-ute* + the affix *-ion* or verb not ending in *-ate, -ete, -ote* or *-ute* + an affix like *-tion*	The verb with the higher frequency in the COBUILD type list takes precedence.	*annunciation* is analysed first as the verb *announce* plus the affix *-iation*, and then as the verb *annunciate* plus the affix *-ion*.
adjective + suffix *-ly* or adjective + suffix *-ally*	The adjective with the higher frequency in the COBUILD type list takes precedence.	*problematically* is analysed first as *problematic* + *-ally*, and second as *problematical* plus *-ly*.
verb + suffix or noun + suffix	Verb takes precedence when the suffix is *-able, -er, -or* or *-ure*; noun takes precedence when the suffix is *-ery, -ism, -ist, -ous, -some* or *-y*.	*comfortable* is analysed first as the verb *comfort* plus the suffix *-able*, and second as the noun *comfort* and the suffix *-able*. *chatty* is analysed first as the noun *chat* plus the suffix *-y*, and second as the verb *chat* plus the suffix *-y*.
verb + suffix *-age* or noun + suffix *-age*	When the word denotes action or an instance of a phenomenon, then the verb takes precedence; when the word denotes a measure or collection of something, then the noun takes precedence.	*leakage* is analysed first as the verb *leak* + the suffix *-age*, and second as the noun *leak* + the affix *-age*.
prefix *a-* + noun or prefix *a-* + verb	Verb takes precedence.	*aglow* is analysed first as the prefix *a-* plus the verb *glow*, and second as prefix *a-* plus the noun *glow*.
Preferential order for the analysis of compounds:		
verb + noun or noun + noun	Verb + noun takes precedence.	*checkpoint* is analysed first as the noun *check* plus the noun *point*, and second as the verb *check* plus the noun *point*.
noun + noun or noun + verb	Noun + noun takes precedence.	*windfall* is analysed first as the noun *wind* plus the noun *fall*, and second as the noun *wind* plus the verb *fall*.

Table 15: How to order multiple analyses of compounds and derivations

following this one). Then say that you are interested in two stems *dive-bomber*, which has three different analyses, and *typesetter*, which has two. Both words are noun-verb-affix type words which may be derivations or compounds or derivational compounds, and this accounts for four of the analyses given. However, for the compound analysis of *dive-bomber*, the stem *dive* is analysed as a verb but can also be thought of as a noun, which gives an extra analysis.

First you can decide whether you want just one default analysis for each stem, or whether you want to see all the available analyses.

If you want to see all possible segmentations, then you don't need to add extra restrictions. As the ***MorphCnt*** column indicates, there are three analyses given for *dive-bomber* and two for *typesetter*, so this is what the unrestricted example lexicon looks like:

Stem	MorphNum	NVAffComp	Der	Comp	DerComp	Def	Imm	ImmClass
dive-bomber	1	Y	Y	N	N	Y	dive-bomb+er	Vx
dive-bomber	2	Y	N	Y	N	Y	dive+bomber	VN
dive-bomber	3	Y	N	Y	N	N	dive+bomber	NN
typesetter	1	Y	Y	N	N	Y	typeset+er	Vx
typesetter	2	Y	N	N	Y	Y	type+set+er	NVx

Derivations take precedence over compounds, so for both words the first row, with analysis number 1, contains the derivation and gets Y under ***Der***. And since each word has only one possible derivation analysis, both are also default analyses, and therefore get Y under ***Def*** too. The N under ***Comp*** and ***DerComp*** confirm that they are not compounds or derivational compounds.

Compounds take precedence over derivational compounds, so for *dive-bomber* the next two rows contain the two compound analyses, with analysis numbers (***MorphNum***) 2 and 3. Both get the code Y under ***Comp***. Since verb + noun compounds take precedence over noun + noun compounds, the verb + noun analysis is a default analysis: it gets 2 as its ***MorphNum***, and Y under ***Def***. The noun + noun analysis gets 3 as its ***MorphNum***, and N under ***Def***. The N codes under ***Der*** and ***DerComp*** confirm that neither of these analyses is a derivation or a derivational compound.

The last row in the lexicon gives the derivational compound analysis of *typesetter*, with `Y` under ***DerComp***. Since it is the only possible derivational compound analysis, it is also a default analysis, and therefore gets `Y` under ***Def*** too. The `N` under ***Der*** and ***Comp*** confirm that it is not a derivation or a compound.

However, rather than including all four forms in your lexicon, you might want to ignore the derivation and derivational compound analyses, and just see the compound analyses. To do this for all the stems in the database, you should add an 'expression' restriction to your lexicon which states that `Comp = Y`. In the example lexicon, this one restriction produces the following result:

Stem	MorphNum	NVAffComp	Der	Comp	DerComp	Def	Imm	ImmClass
dive-bomber	2	Y	N	Y	N	Y	dive+bomber	VN
dive-bomber	3	Y	N	Y	N	N	dive+bomber	NN

In the same way, if you want to examine derivational compound analyses, and leave out all the other analyses, you should add an 'expression' restriction to your lexicon which states that `DerComp = Y`. In the example lexicon, this restriction produces the following result:

Stem	MorphNum	NVAffComp	Der	Comp	DerComp	Def	Imm	ImmClass
typesetter	2	Y	N	N	Y	Y	type+set+er	NVx

Rather than seeing a number of analyses, you might prefer to look at just one straightforward default analysis, no matter how many alternatives are given in subsequent rows. Again, you can quickly construct restrictions to make this possible. The quickest way is to use the ***MorphNum*** column, which gives a number to each analysis of each stem. You can say `MorphNum = 1`, which means that only the very first analysis of each stem appears in your lexicon.

Sometimes there may be more than one default analysis. If you want to see just the default analysis of each compound, you should use these two restrictions: `Def = Y` and `Comp = Y`. In the example lexicon, this means that the non-preferred noun + noun analysis is left out:

Stem	MorphNum	NVAffComp	Der	Comp	DerComp	Def	Imm	ImmClass
dive-bomber	2	Y	N	Y	N	Y	dive+bomber	VN

These explanations may appear complicated, but by reading them, you can get to know the important restrictions that you can use to extract the types of analyses you really want.

3.2.2 IMMEDIATE SEGMENTATION

Immediate segmentation is the least detailed form of analysis offered here. It doesn't give you a full analysis, right down to all the smallest elements a stem contains; rather it is a simple, one-level breakdown of a stem into its next biggest elements. So, while complete segmentation is equivalent to a full analytical tree, immediate analysis can be thought of as a close look at a particular level.

There are ten columns which present the immediate segmentation of stems to you. The first gives the orthography of the analysed elements. The next three give more general codings, so that using the FLEX options `SHOW` and `QUERY`, you can look for stems which have a particular form – a preposition plus a noun, say, or a stem plus a stem plus an affix, and so on. The remaining six deal with particular features which sometimes occur in morphological analysis: stem allomorphy, affix substitution, opacity, derivational transformation, infixation and reversion.

In the first column, you get the orthography of the first-level elements themselves, each separated by a + sign. Diacritical markers are not included. Thus the stem *nameplate* is shown as `name+plate`, in accordance with the various rules discussed in section 3.1.1. Note that each element is given in the form of a headword or an affix, even when the original word doesn't use that particular form. Thus the stem *liturgical* is analysed as *liturgy+ical*, where *liturg* is rewritten in the normal form of the stem *liturgy*. The FLEX name and description of this column are as follows:

Imm ***(ImmLemma)***	`Immediate segmentation`

The second column is like the first, except that where the first column gives you the orthography of each element, this column gives you the word class of each element, leaving out any + signs. Single letter labels are used to represent the syntactic class of each element – which is unlike many of the syntactic codes used in other parts of the database. The use of a single character means that there is no possibility of a code becoming ambiguous, since each character is unique. Table 16 shows you the labels used in this column:

Word Class	Label
Noun	N
Adjective	A
Numeral	Q
Verb	V
Article	D
Pronoun	O
Adverb	B
Preposition	P
Conjunction	C
Interjection	I
Single contraction	S
Complex contraction	T
Affix	x

Table 16: Word class labels (immediate segmentation)

Using these codes, *nameplate* is given the code `NN`, to indicate that it is made up of two nouns (a compound), and *emigration* has the code `Vx` to indicate that it is made up of a verb and an affix (a derivation). The FLEX name and description of the column that gives you these codes are as follows:

ImmClass (ImmClassLemma) `Immediate segmentation, word class labels`

The third column provides more detailed information about the syntactic categorization of verbal stems. The basic codes used are exactly the same as the `ImmClass` column, except that instead of the `V` code to represent a verb, any one of a number of codes is given. Table 17 shows you these codes, along with their meaning.

Verbal sub-category	Label
Intransitive	1
Transitive	2
Intransitive & transitive	3
Unmarked for transitivity	0

Table 17: One-character verbal subclass labels

In this column, the word *emigration* has the code `1x`. It is exactly the same as the code in the previous column, except that the `V` is replaced by the number `1`, indicating in more detail what sort of verb it is.

The FLEX name and description of this column are as follows:

ImmSubCat
(ImmSubCatLemma)

```
Immediate segmentation, subcat labels
```

The fourth immediate segmentation column simply tells you whether the elements identified are stems or affixes. Upper case `S` indicates a stem, upper case `A` indicates an affix, and upper case `F` indicates a flectional form of a stem. Thus *emigration* is represented as `SA`, and *bagpipes* as `SF`. The FLEX name and description of this column are as follows:

ImmSA
(ImmSALemma)

```
Immediate segmentation, stem/affix labels
```

The fifth immediate segmentation column concerns stem allomorphy. Within a word, a stem sometimes takes a form different from the one used when it is written down as a word in its own right. When morphological analysis is noted down, any resulting stems are given their normal stem form, because it's easiest to understand. An example is the word *abundant*, which comprises the stem *abound* and the affix *ant*. Note the difference between what appears in the original word (*abund*) and its regular stem form (*abound*): each has the same meaning; the only difference between them is their spelling. This is an example of *derivational* stem allomorphy, since a new word has been derived by linking a different form of stem to an affix.

Another sort of stem allomorphy sometimes occurs with conversions – that is, words which change their word class without the addition of an affix (*sleep* is both a noun and a verb, for example). When conversion occurs, and the form of the stem seems to have altered, the process can be termed *conversion with allomorphy*. The verb *halve* is an instance of conversion with allomorphy, since it is a conversion from the noun form *half*. Thus *half* and *halve* are considered *allomorphs*: two different forms or representations of the same stem. There are three types of conversion with allomorphy. The first is the voicing of the final consonant with the addition of a final *-e*: thus the verb *thieve* is considered to be a conversion of the noun *thief*. The second is the same process in reverse – the removal of a final *-e*, and the devoicing of the last consonant: thus the noun *belief* is a conversion of the verb *believe*. The third is the change in spelling from final *s* to *c*: the noun *practice* can thus be thought of as a zero-derivation from the verb *practise*.

The next type of stem allomorphy is *flectional allomorphy*, a relatively rare type. When the irregular past tense of a verb is used as an adjective, both are said to derive from the infinitive form, so that the adjective *drunken* comes from the verb *drink*. The same is true for past participle forms: the adjective *born* thus derives from the verb *bear*.

There are two other categories which are dealt with under stem allomorphy even though they're not really instances of stem allomorphy – *clippings* and *blends*. Clippings are shortened forms of words which do not change word class. For example, *phone* is a simple clipping of *telephone*. Sometimes a clipping consists of more than one morpheme – *vibes* is a clipping of *vibraphone* which contains the stem *vibraphone* and the affix *-s*, and *hanky* is a diminutive form consisting of the stem *handkerchief* and the affix *-y*.

A blend is a word which is made up of two stems, at least one of which may be shortened. The word *smog* is made up of the stems *smoke* and *fog*, and *paratrooper* consists of the stems *parachute* and *trooper*. Note that the definition of a blend only allows for stems, not affixes.

The table below summarizes the five types of allomorphy and shows the codes used to identify them in the ***ImmAllo*** column.

Stem Allomorphy	Code	Example
Blend	B	*breathalyse*
Clipping	C	*phone*
Derivational	D	*clarify*
Flectional	F	*born*
Conversion	Z	*belief*

Table 18: Stem allomorphy codes

The FLEX name and description of this column are as follows:

ImmAllo
(ImmAlloLemma)

`Stem allomorphy, top level`

The sixth immediate segmentation column marks stems with a morphological analysis involving *affix substitution*. This is the process whereby an affix replaces part of a stem when that stem and the affix join to form another stem. For example, *active* is analysed as the stem *action* and the affix *-ive*; the affix *-ion* has disappeared, and the new affix *-ive* has taken the place of the old one. So, this column gives `Y` for yes if the immediate analysis of the stem involves affix substitution, or `N` for no if it does not. The FLEX column name and description of this column are as follows:

ImmSubst
(ImmSubstLemma)

`Affix substitution, top level`

The seventh column identifies those words whose analysis is *opaque* – that is, words made up of morphemes which are recognisable, but where the meaning of the head element isn't reflected in the meaning of the full word. An example of this is *accordion*: it appears to be made up of the verbal stem *accord* (the head element) and the affix *-ion*. Since the semantic link between *accord* and *accordion* is far from obvious, the analysis is marked as being opaque, and it gets a `Y` in this column. Words whose analyses are morphologically and semantically clear get the code `N`. The FLEX name and description of this column are as follows:

ImmOpac
(ImmOpacLemma)

`Opacity, top level`

The eighth immediate segmentation column gives simple expressions to illustrate any orthographic alterations the analysis of a word involves. A morpheme boundary is marked by a `#`, and letters removed from either side of a morpheme are prefixed by a `-`, and letters which are added are prefixed by a `+`. Letters which do not change are considered part of a morpheme, and not shown. A simple example is `#` – this is the pattern for the word *unable*, since it consists of the affix *un-* and the stem *able*, and neither morpheme alters. On the other hand, *undersized* is given as `#-e#`, since nothing happens to the first morpheme *under-*, the final *e* of the second morpheme *size* is removed, and nothing happens to the last morpheme *-ed*. The FLEX name and description of the columns that contain these expressions are as follows:

TransDer
(TransDerLemma) `Derivational transformation, top level`

The ninth column indicates which stems have an immediate analysis involving derivation by means of an infix. Usually, derivational affixes are added to the beginning or end of a stem, but in some cases the affix is inserted into a multi-word, as in derivations from verb-and-particle combinations like *hanger-on* from *hang on* and *looker-on* from *look on*. Stems marked for this type of infixation get the code `Y` in this column, all other analyses get the code `N`.

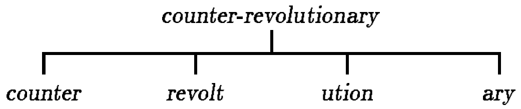
ImmInfix
(ImmInfixLemma) `Infixation, top level`

The last immediate segmentation column deals with stems analysed as conversions from multi-words which have undergone reversion of their parts in the process of conversion. For instance, the noun *downpour* is considered a conversion of the verb *pour down*, and the adjective *off-putting* is derived from the verb *put off* via its flection *putting off*. Whenever a stem is analysed in this way, this column yields a `Y` code. In all other cases, an `N` code is given.

ImmRevers
(ImmReversLemma) `Reversion, top level`

3.2.3 COMPLETE SEGMENTATION (FLAT)

Complete segmentation is ‘complete’ in the sense that it identifies all the morphemes a stem contains. This is in contrast to immediate segmentation, which only picks out the next two (sometimes three) morphological elements. The complete segmentation discussed in this section is also *flat*, which means that you can see what the constituent morphemes are without knowing the details of the full morphological analysis which has been carried out. When you draw a morphological ‘tree diagram’, this information gives the outermost branches only; you cannot analyse any further, and you cannot see the intermediate levels. So, when you want to see the complete, flat, segmentation of *counter-revolutionary* for example, you get this sort of information:



There are three columns with complete segmentation (flat) information. The first contains the morphemes themselves. The second contains the word class of each morpheme, and the third simply states whether each morpheme is a stem or an affix. The last two columns are useful when you're looking for a stem with a particular combination of morphemes: using the FLEX `SHOW` and `QUERY` options, you can hunt out stems which are made up of a noun plus an affix plus a noun, say, or all the stems which contain at least three other stems.

The first column gives you each stem split into its morphemes by + signs. Thus the stem *counter-revolutionary* is written in the following way:

```
counter+revolt+ution+ary
```

No diacritics are included. The FLEX name and description of this column are as follows:

Flat
(FlatLemma) `Flat segmentation`

The second column uses single-letter codes to represent the word class of each morpheme.

Word Class	Label
Noun	N
Adjective	A
Numeral	Q
Verb	V
Article	D
Pronoun	O
Adverb	B
Preposition	P
Conjunction	C
Interjection	I
Single contraction	S
Complex contraction	T
Affix	x

Table 19: Word class labels (flat segmentation)

Using these codes, the stem *counter-revolutionary* is given as `xVxx`. The FLEX name and description of the column are as follows:

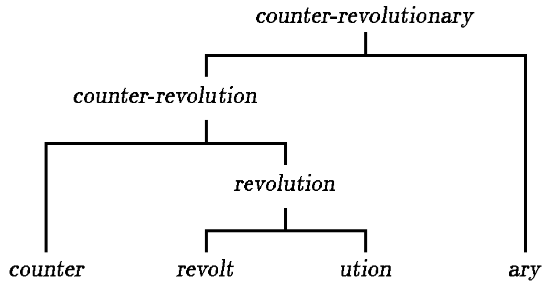
FlatClass (FlatClassLemma) `Flat segmentation, word class labels`

The last column simply indicates whether each morpheme is a stem, a flection or an affix. Upper case `S` means Stem, upper case `F` means Flection, and upper case `A` means Affix. The full code for *counter-revolutionary* is thus `ASAA`. The FLEX name and description of this column are as follows:

FlatSA (FlatSALemma) `Flat segmentation, stem/affix labels`

3.2.4 COMPLETE SEGMENTATION (HIERARCHICAL)

Complete, hierarchical segmentation gives the most detailed analysis available for each stem. It is called *hierarchical* because it can cover several different levels: it is arrived at after immediate analysis has been carried out on every stem that can be identified within a larger stem. With this information, you can draw a complete morphological ‘tree diagram’, from the root to the outermost branches, with every intermediate branch fully represented. So, for the stem *counter-revolutionary*, you can get the following morphological analysis:



There are six columns which give information about the full segmentations of stems. Three of them give the hierarchical segmentations themselves. The simplest of these tells you what the constituent morphemes of the stem are, indicating with algebra-like brackets the structure of the 'tree'. Also available are similar bracket notations which supply a word class label alongside each element on each level, or the word class without the spelling of the element itself. The remaining two columns indicate whether stem allomorphy or affix substitution has occurred anywhere in the full hierarchical analysis.

The first column provides all the information you need to draw a tree diagram like the one above – that is, the constituent morphemes of a stem each delimited by a comma and enclosed in brackets which indicate its complete morphological structure. The stem *counter-revolutionary* thus looks like this:

```
(((counter),((revolt),(ution))),(ary))
```

Each identifiable stem or affix is enclosed by a pair of brackets, beginning with the brackets round the full original stem. Then there are brackets round the stem *counter-revolution*, and subsequently round the stem *revolution*. Finally there are brackets round each of the four morphemes.

The FLEX name and description of the column which contains morphological analyses in this form are as follows:

Struc `Structured segmentation`
(StrucLemma)

The next two columns use extra labels to indicate the word class of each segment. They are given between square brackets to the right of each closing round bracket, so that every segment on every level within the original stem has a word class code. The word class codes used are as follows:

Word Class	Label
Noun	N
Adjective	A
Numeral	Q
Verb	V
Article	D
Pronoun	O
Adverb	B
Preposition	P
Conjunction	C
Interjection	I
Single contraction	S
Complex contraction	T

Table 20: Word class labels (complete segmentation)

The codes used for affixes are combinations of these word class labels. The stem *counter-revolutionary* can be represented as follows:

```
(((counter)[N|.N],((revolt)[V],(ution)[N|V.])[N])[N],(ary)[A|N.])[N])
```

This example illustrates the special form affix codes take. There are two elements in each affix code which are separated by a vertical bar `|`. In front of the vertical bar is a single code which is the word class of the stem which the affix in question helps to form. After the vertical bar comes a combination of single letter codes which indicate the word class of each element within the stem formed, and the position of the affix itself is given by a dot.

In the *counter-revolutionary* example above, the code given alongside the affix *counter* is `[N|.N]`. The `N` before the bar means that the affix *counter* helps to form a stem which is a noun (*counter-revolution*). The `.N` after the bar means that the segmentation of the noun *counter-revolution* is affix plus noun. These detailed codes can help you to identify the way affixes are used, and to get lists of stems which contain affixes used in particular contexts: the fact that the second part of the *counter* code is `.N` helps you to see at once that this affix helps to form a derivation in conjunction with a noun.

Sometimes a pair of affixes can only be used together, as in the word *aerodrome* – the word *aero* does not exist and the word *drome* does not exist. In such cases, `x` marks the other affix, and denotes that the affixes must occur in combination with each other: so-called *combining forms*. The code for the *aero-* of *aerodrome* is thus `[N|.x]`, and the code for the *-drome* is `[N|x.]`.

So, this column is particularly useful for two things. First, you get the word class of each stem in the segmentation alongside the orthographic representations of individual morphemes. Second, you get detailed information about each affix each stem contains. The FLEX name and description of this column are as follows:

StrucLab
(StrucLabLemma)

`Structured segmentation, word class labels`

The next column shows the hierarchical structure of each stem by means of round brackets and commas, and the full word class labels between square brackets, just as with the previous column. The only difference is that in this column the orthographic representation of the constituent stems and affixes is missed out altogether. Thus the stem *counter-revolutionary* gets the following representation:

```
((()[N|.N],(()[V],()[N|V.])[N])[N],()[A|N.])[N]
```

This column again helps you to search for stems which have a particular morphological structure and particular combinations of syntactic elements. The FLEX name and description of this column are as follows:

StrucBrackLab
(StrucBrackLabLemma)

`Structured segmentation, word class labels only`

The fourth hierarchical segmentation column deals with stem allomorphy. Within words, stems sometimes take a form different from their generally accepted stem form. When a morphological analysis is noted down, the resulting stems are given their normal stem orthography. An example is the word *inedible*, which comprises the affix *in-*, the stem *eat* and the affix *-ible*: note the difference between *ed* and *eat*, where the one element is spelt two different ways. This is stem allomorphy. If stem allomorphy occurs at any point in a stem's complete hierarchical segmentation, a code is given in this column to show what sort of stem allomorphy occurs. The table below shows the codes, and you can read more about what each code means in section 3.2.2 above – they are the same codes used in the ***ImmAllo*** column.

Stem Allomorphy	Code	Example
Blend	B	*breathalyse*
Clipping	C	*phone*
Derivational	D	*clarify*
Flectional	F	*born*
Conversion	Z	*belief*

Table 21: Stem allomorphy codes

The FLEX name and description for this column are as follows:

StrucAllo
(StrucAlloLemma)

`Stem allomorphy, any level`

The fifth hierarchical segmentation column marks stems with a morphological analysis involving *affix substitution*. This is the process whereby an affix replaces part of a stem when that stem and the affix join to form another stem. For example, *melodic* is analysed as the stem *melody* plus the affix *-ic*; the affix *-y* has disappeared, and the new affix *-ic* has taken the place of the old one. So, this column gives `Y` for yes if the complete analysis of the stem involves affix substitution, or `N` for no if it does not. The FLEX name and description of this column are as follows:

StrucSubst
(StrucSubstLemma)

`Affix substitution, any level`

The sixth and last hierarchical segmentation column identifies those words whose analysis is completely or partly *opaque* – that is, words made up of morphemes which are recognisable, but where the meaning of the head element isn't reflected in the meaning of the full word. An example of this is *ladykiller*: it appears to be made up of the noun stem *lady* and the noun stem *killer* (which can subsequently be analysed as *kill* plus *-er*). Since the meaning of the head element *killer* doesn't relate directly to the meaning of the full word, the analysis is marked as being opaque, and it gets a `Y` in this column. Words whose analyses are morphologically and semantically clear get the code `N`. The FLEX name and description of this column are as follows:

StrucOpac `Opacity, any level`
(StrucOpacLemma)

3.3 OTHER CODES

The remaining three columns give counts of various sorts: the number of *components* (i.e. stems and affixes) in the immediate analysis of each stem, the number of *morphemes* a stem contains after complete segmentation, and the number of *levels* involved in the complete hierarchical analysis of each stem.

The first of these columns is the simple count of the number of components each stem contains. The normal figure is two; words are generally split into two parts each time one level of morphological analysis takes place. Sometimes three components can be identified: derivational compounds are usually analysed as a stem plus a stem plus an affix, as are normal compounds which are joined with a special 'link morpheme' (*-a-*, *-o-*, or *-s-*). And of course, monomorphemic words only contain one component. Any stems which cannot receive an adequate morphological analysis (for the reasons given in section 3.1.4) get the number 0.

Some examples: in the stem *counter-revolutionary*, the number of components is two (the stem *counter-revolution* and the affix *-ary*), and for *law-breaker* it is three (the stem *law*, the stem *break*, and the affix *-er*.

The FLEX name and description of this column are as follows:

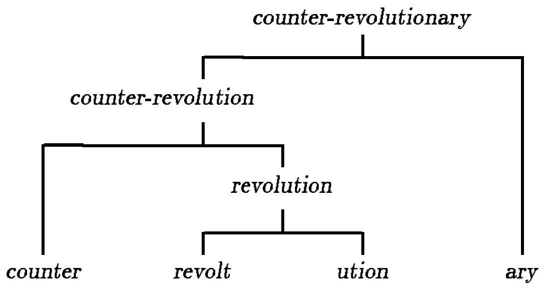
CompCnt `Number of morphological components`
(CompCntLemma)

The second column gives you the number of morphemes in each stem. For words without a morphological analysis, the number given is zero. The number of morphemes in the stem *counter-revolutionary* for example is four, while for *lawbreaker* it is three.

The FLEX name and description of this column are as follows:

MorCnt `Number of morphemes`
(MorCntLemma)

The last of the three columns gives a count of the number of levels in the complete hierarchical segmentation described above, which is best illustrated by another look at the tree diagram illustrating the analysis of *counter-revolutionary*:



Including the stem at the top, the diagram covers four lines: this is the *number of levels* the stem has. It is the number of times you can carry on doing immediate analysis when you analyse a particular stem in full. Do not confuse it with the number of all the immediate analyses required to arrive at the complete hierarchical segmentation; any one *level* of analysis may include more than one immediate segmentation. Monomorphemic stems always get the number 1, while stems without analysis (for reasons explained in section 3.1.4) get the number 0.

The FLEX column name and description of this column are as follows:

LevelCnt `Number of morphological levels`
(LevelCntLemma)

3.4 MORPHOLOGY OF ENGLISH WORDFORMS

There are two types of morphology information available for the wordforms given in the CELEX database: first, information about the lemma which underlies each family of wordforms, and second, a simple identification of the inflectional features which are specific to each wordform, either in the form of thirteen 'yes/no' feature columns or one column with feature identification codes.

Dictionaries present their lexical information under bold-type headwords, which are used instead of listing every individual inflected form separately. Such a form is often called the *canonical form*, since it represents a full canon of inflections. Thus the word *eat* is understood as referring not only to the form *eat* itself, but also the forms *eats, eating, ate*, and *eaten*. To print full details about every inflected form separately would result in a lot of needless repetition and enormous books which no one could lift from the bookshelf. However, for many applications, lemma information has to be listed for each individual wordform, and in a CELEX lexicon of type wordform, you can do just that when you include certain 'morphological' columns. This is done by providing a link between the wordform information and the lemma information. When you choose the option `Lemma information` from the `ADD COLUMNS` menu, you are in fact being allowed into the lemma information by the back door. You can now look up information specific to a particular wordform in your lexicon, and at the same time see general information which is common to all the other forms in the same inflectional paradigm. One particularly useful type of lemma information you can use in your wordform lexicon is the syntactic information, which can give the word class of any wordform you are looking at. There is also an important distinction which you may be able to draw upon with the frequency information. The wordform lexicon gives you a COBUILD frequency figure specific to each wordform, while the lemma information available lets you see the sum frequency for all the inflectional forms in the same paradigm, a figure referred to as the *lemma frequency*.

All the lemma information has already been defined elsewhere in this linguistic guide, so there is no point in repeating it here. All that needs to be pointed out is that the column names used in a real lemma lexicon differ from those used in the lemma information option in the morphology of wordforms. When a FLEX column name and description are defined in the course of lemma lexicon text, the column name given in brackets is the name of the column when it is used as part of a wordforms lexicon. Usually this name is identical to the lemma lexicon name, except that the word *lemma* is added to the end.

ExampleName
(ExampleNameLemma)

The column names used for lemma information in a Wordforms lexicon are given in brackets, as this Example Name shows.

All the other details and definitions remain the same in both cases. So, when you're looking for the columns of lemma information provided with a wordforms lexicon under morphology, just go back to the original lemma information: it's all there.

3.4.1 INFLECTIONAL FEATURES

There are thirteen special columns available only with a lexicon of type wordforms. Each one corresponds to a particular inflectional attribute which a wordform can have. There can only be one of two codes in each column: Y for 'yes, this wordform has this attribute', or N for 'no, this wordform does not have this attribute'. These columns are therefore useful for constructing restrictions on your lexicons, restrictions which need not be 'on view': it's unlikely that you will want to look at the contents of these columns with the SHOW option. (If, on the other hand, you want to have a label which lets you see at a glance all the inflectional features each wordform has, then you should use the 'type of flection' codes described in the next section.)

An example. To make a lexicon which gives you all first person, present tense verb forms in the database, you have to include at least three columns in the wordforms lexicon you create, namely a column which gives the orthographic representations you prefer, along with ***Pres*** and ***Sin1*** (which are amongst the thirteen columns described below). You must then construct two restrictions for your lexicon, one stating that ***Pres*** must be equal to Y, and another stating that ***Sin1*** must be equal to Y. You can then format your lexicon to make sure that ***Pres*** and ***Sin1*** are not 'on view': that way, when you SHOW or EXPORT your lexicon, you just get the list of words you require without two lists of Y's. To this basic lexicon you can of course add any other columns you require, either the orthographic and frequency information specific to each wordform, or the general lemma information—particularly syntax—which is available through the 'Morphology of English wordforms' options.

The first inflectional features column indicates whether a wordform is a singular form of any sort. This means past and present tense verb forms such as *hibernated* or *babbles*, or nouns such as *sagacity*. The FLEX name and description of this column are as follows:

Sing `Inflectional feature: singular`

The second column indicates whether a wordform is a plural inflection of any sort. This means past and present tense verb forms such as *hibernate* or *submerged*, or nouns such as *jocularities*. The FLEX column name and description of this column are as follows:

Plu `Inflectional feature: plural`

The third column marks all the wordforms which are positive forms – that is, not comparative or superlative forms like *better* and *best*, but plain adjectival forms like *good* or *often*. Thus adjectives like *goofy* and *idiomatic* or adverbs like *seldom* and *idiomatically* get the code Y, while all other forms get the code N. The FLEX name and description of this column are as follows:

Pos `Inflectional feature: positive`

The fourth column marks all the wordforms which are comparative forms, almost always adjectives. Wordforms such as *better* or *angrier* or *cannier* thus get the code Y, while all other non-comparative forms get the code N. There is also a small number of comparative adverbs which get the Y code, such as *further*. The FLEX name and description of this column are as follows:

Comp `Inflectional feature: comparative`

The fifth column marks all superlative forms, so that wordforms such as *best* or *angriest* get the code `Y`, and every other form gets the code `N`. There is also a small number of superlative adverbs which get the `Y` code, such as *furthest*. The FLEX name and description of this column are as follows:

Sup `Inflectional feature: superlative`

The sixth column marks the form of the verb usually known as the infinitive. It is used as a headword in the CELEX databases, and in most dictionaries. Words like *waffle* or *have* or *eat*, which can be used with the particle *to* in front of them, are infinitives. Any wordform which is an infinitive gets a `Y` code in this column; all the others get the code `N`. The FLEX column name and description for this column are as follows:

Inf `Inflectional feature: infinitive`

The seventh column marks any participles, past tense or present tense. Present participles are normally formed by adding *-ing* to the stem of the verb, with the exception of some irregular verbs. Past participles add a suffix ending in *-d* to the stem, and they are used in the formation of the perfect tense: 'I've *lived* in Nijmegen for four years'. Again, many irregular verbs don't match this rule (*gone* is the past participle of *go*, for example). Most past participles can also be used adjectivally, as in 'the *panelled* walls'. Any wordforms which are participles get the code `Y`, and all the rest get the code `N`. The FLEX name and description of this column are as follows:

Part `Inflectional feature: participle`

The eighth column identifies any present tense forms, including the present participles mentioned under ***Part***. Thus verb forms like *gleam, gleams* and *gleaming* get the code `Y`, while all other forms (including infinitives, which are marked in a different column) get the code `N`. The FLEX name and description of this column are as follows:

Pres `Inflectional feature: present tense`

The ninth column identifies any past tense forms, including the past participles mentioned under ***Part***. Thus forms like *occupied* and *elicited* get the code `Y`, while all other forms (including infinitives, which are marked in a different column) get the code `N`. The FLEX name and description of this column are as follows:

Past `Inflectional feature: past tense`

The tenth column marks first person singular forms of verbs, whether present tense or past tense. So, all first person singular forms, like 'I *go*' or 'I *finished off*', are given the code `Y`, and every other form gets the code `N`. The FLEX column name and description of this column are as follows:

Sin1 `Inflectional feature: 1st person verb`

The eleventh column marks second person singular forms of verbs, whether present tense or past tense forms. For most verbs, the second person form is the same as the first person form, but some irregular verbs are exceptions. So all second person forms like 'you *are*' or 'you *shout*' are given the code `Y`, and every other form gets the code `N`. The FLEX column name and description of this column are as follows:

Sin2 `Inflectional feature: 2nd person verb`

The twelfth column identifies third person singular forms of verbs, whether present tense or present tense forms. For most verbs, the third person present tense form consists of the stem plus the suffix *-s*. Thus forms like 'he *stood up*' or 'Gilbert *acts*' get the code `Y` while every other form gets the code `N`. The FLEX name and description for this column are as follows:

Sin3 `Inflectional feature: 3rd person verb`

The thirteenth and last column marks rare forms – normally forms which have become outdated like *brethren*, *shouldst*, or *wert*. Such forms have the code `Y` in this column, while every other wordform gets the code `N`. The FLEX name and description of this column are as follows:

Rare `Inflectional feature: Rare form`

3.4.2 TYPE OF FLECTION

In the 'Inflectional Features' section above, thirteen different inflectional features are distinguished, and assigned to thirteen separate 'yes/no' columns. The same information is also available in one single column, using combinations of single-letter codes to show all the features each wordform has. The 'yes/no' columns are useful for constructing restrictions on your lexicon, whereas the 'type of flection' column described here provides you with a label that identifies at a glance all the features each wordform has. Table 22 below sets out all the combinations of single-letter codes that occur.

Inflectional feature	Label	'yes/no' column name
Singular	S	***Sing***
Plural	P	***Plu***
Positive	b	***Pos***
Comparative	c	***Comp***
Superlative	s	***Sup***
Infinitive	i	***Inf***
Participle	p	***Part***
Present tense	e	***Pres***
Past tense	a	***Past***
1st person verb	1	***Sin1***
2nd person verb	2	***Sin2***
3rd person verb	3	***Sin3***
Rare form	r	***Rare***
Headword form (not nouns, verbs adjectives or adverbs)	X	

Table 22: Type of flection labels

For a full definition of these flection types, read the details given for the appropriate 'yes/no' columns in section 3.4.1 above. However, note that there is one type of flection label which does not correspond to a 'yes/no' column. The X label identifies many forms not covered by the other labels, including prepositions like *among* or *less*, pronouns like *that* or *hers*, conjunctions like *immediately* or *that*, numerals like *fifth* or *thousand*, contracted forms like *I'll* or *hadn't*, and interjections like *phew* or *amen*. These forms are always the same as those used as the headword form of the lemma (thus the very few inflected adverbial forms do not get the code

X). No nouns, verbs, adjectives or adverbs ever get the code X.

Each wordform may have more than one code attached to it. Thus the wordform *boasted* has the code a3S: a means it is a past tense form, 3 means that it is a third person form, and S means that it is singular.

The FLEX name and description of this column are as follows:

FlectType `Type of flection`

3.4.3 INFLECTIONAL TRANSFORMATION

The last column shows how the orthographic form of a stem is altered when a flection is formed. Each string of letters in the stem is shown by the symbol @, so the first person present tense form of the verb whose stem is *abide* is simply given as @. Any blanks or hyphens in the stem are shown as a blank, so *abide by* is shown as @ @. Letters removed from the front or back of a string are prefixed by a minus sign −, and letters added are prefixed by a plus sign +, so *abiding by* is given as @-e+ing @: first the final *-e* of *abide* is removed, and then the suffix *-ing* is added. This formalism is an unambiguous way of showing the inflectional transformations that occur in the orthographic formation of wordforms.

Whenever the inflectional transformation is irregular (as with the past tense forms of the verb *sing* for example – *sang* and *sung*) no transformation is given; the field remains empty.

The tables below show all the lettergroups represented in the database which can be subtracted from or added to a headword to make a wordform.

Lettergroups removed from a headword				
e	ey	f	fe	y

Table 23: Inflectional transformation codes (letters removed)

Lettergroups added to a headword				
bed	ber	best	bing	d
ded	der	dest	ding	ed
er	es	est	ged	ger
gest	ging	ied	ier	ies
iest	ing	ked	king	led
ler	lest	ling	med	mer
mest	ming	ned	ner	nest
ning	ped	ping	r	red
ring	s	sed	ses	sing
st	ted	ter	test	ting
ved	ving	ves	zed	zes
zing				

Table 24: Inflectional transformation codes (letters added)

The FLEX name and description of this column are as follows:

TransInfl
(TransInflLemma)

`Inflectional transformation`