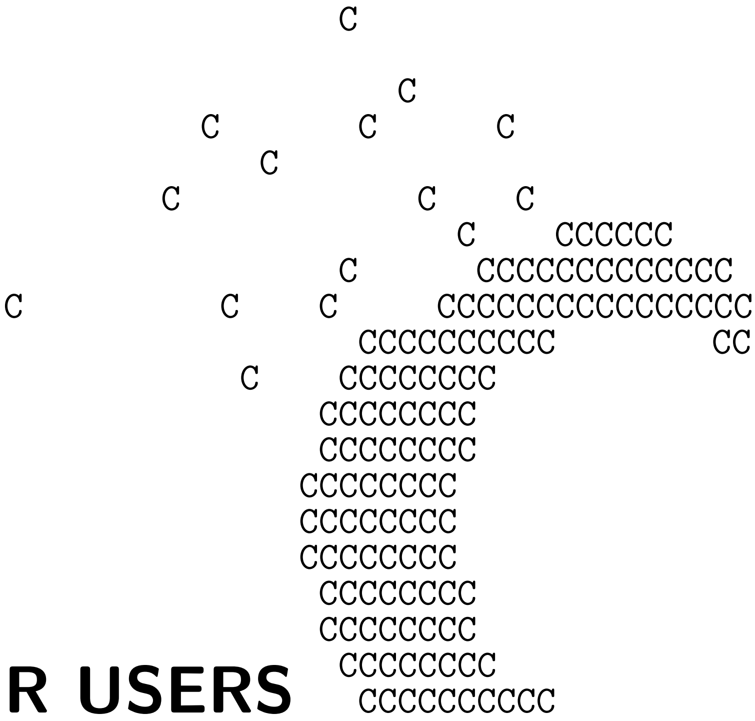
CELEX
A GUIDE FOR USERS

GAVIN BURNAGE

CELEX
CENTRE FOR LEXICAL INFORMATION



CELEX A GUIDE FOR USERS

CELEX – CENTRE FOR LEXICAL INFORMATION

Max Planck Institute for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

Telephone

31-(0)24-3615797
31-(0)24-3615751

Fax

31-(0)24-3521213

Electronic mail

INTERNET: `celex@mpi.nl`

First published in the Netherlands in 1990

ISBN 90 373 0063 4

Typeset using the TEX computer typesetting system.

Printed by drukkerij SSN, Nijmegen.

TEX is a trademark of the American Mathematical Society.

INTRODUCTION

There can be no doubt that lexicography is a
very difficult sphere of linguistic activity.
Many lexicographers have given vent to their feelings in this respect.
Perhaps the most colourful of these opinions
based on a lexicographer's long experience
is that of J.J Scaliger (16th–17th cent.)
who says in fine Latin verses that the worst criminals
should neither be executed nor sentenced to forced labour,
but should be condemned to compile dictionaries,
because all the tortures are included in this work.

— **LADISLAV ZGUSTA** Manual of Lexicography (1971)

The 1980s will one day be seen as a watershed in lexicography –
the decade in which computer applications began to alter radically
the methods and the potential of lexicography.
Gone are the days of painstaking manual transcription
and sorting on paper slips: the future is on disk,
in the form of vast lexical databases, continuously updated,
that can generate a dictionary of a given size and scope
in a fraction of the time it used to take.

— **DAVID CRYSTAL** The Cambridge Encyclopedia of Language (1988)

CONTENTS

1 DATABASES AND LEXICONS 1–1

1.1 Why use a database? 1–1

2 LEXICON TYPES 1–6

2.1 Dutch Lemmas 1–7
2.2 Dutch Wordforms 1–11
2.3 Dutch Abbreviations 1–11
2.4 Dutch INL Corpus Types 1–12
2.5 English Lemmas 1–14
2.6 English Wordforms 1–16
2.7 English COBUILD Corpus Types 1–17
2.8 German Lemmas 1–18
2.9 German Wordforms 1–21
2.10 German Mannheim Corpus Types 1–21

1 DATABASES AND LEXICONS

This introduction tries to do two things. In the first section, for those who aren't familiar with the ideas and possibilities of databases and lexicons, there is a description of the way in which a computer database and lexicon is like—and more importantly *unlike*—a traditional paper dictionary. If you're already familiar with such things, you may like to skip ahead to the second section, where there is a description of each of the main lexicon types available to you in FLEX. Fundamental to this description is the difference between *wordforms* (the words we use in everyday speech and writing) and *lemmas* (words used to represent families of wordforms, in the same way as bold-type dictionary headings, which take the form of stems or headwords). Since the linguistic information available to you depends on the type of your lexicon, you should make sure you understand the differences between the various lexicon types before beginning your work. And when you start work with FLEX, the special program which helps you build and use your lexicons, you'll be better off for having read these sections carefully. In the third and last section of this introductory chapter, you can find out how to log into CELEX using local, national, and international computer networks.

1.1 WHY USE A DATABASE?

Since we are dealing with words, we can start off by thinking of databases in terms of a paper dictionary. A book like the *Van Dale Groot Woordenboek der Nederlandse Taal* is essentially a long list of words with information supplied alongside each word. The key to a dictionary is the alphabetical order of its word entries: you can only look up one particular word at a time and examine the information given for it. If you've got time, you can look at every page to find all the words with a certain grammatical code or pronunciation, but, quite understandably, most people don't do this unless they're really desperate.

In its simplest form, a database can be like a dictionary: just a list of words, and some information alongside each word.

The first important difference between a computer database and a paper dictionary is that the database uses different *columns* to store separate types of information, whereas the dictionary uses one paragraph of text, and marks different sorts of information within that text by using different typefaces and coding systems, or by giving the information in a particular order. Dictionary text is fixed once it is printed. You can't move bits of an entry around, or miss them out: you are presented with everything at once, and you may have to read a lot of irrelevant information before you find what you're looking for. The columns which make up a database are much more regimented, but that, paradoxically, is what gives a database its flexibility. Each type of information keeps strictly to its own dedicated place, which means it's easier for the computer to locate and serve up one individual item, or several particular items, relating to each word that interests you. So, you can look up a word and its word class code and pronunciation, say, without even having to glance at all the other information. The diagram below is a simple representation of how information is held in a database.

Headword	**Class**	**Phonetics**	
aback	ADV	@-'b&k	
abacus	N	'&-b@-k@s	
abandon	N	@-'b&n-d@n	
abandon	V	@-'b&n-d@n	
abandoned	A	@-'b&n-d@nd	
abandonment	N	@-'b&n-d@n-m@nt	
abase	V	@-'beIs	
abasement	N	@-'beIs-m@nt	
abash	V	@-'b&S	
abate	V	@-'beIt	

The crucial difference between a database and a dictionary is the flexibility that a computer can achieve with the properly-defined rows and columns: you can gather together different parts of the database, and display the information in any way you like. This illustration shows you three vertical *columns*, which are entitled 'Headword', 'Class', and 'Phonetics', and ten horizontal *rows*, each of which displays information for each headword under the correct column heading. A row thus contains every type of information for one word, while a column contains one specific type of information for every headword.

The illustration is, of course, only a very simple example. To get an idea of what the whole CELEX Dutch, English or German database might look like, imagine three hundred or so more column headings added to the right hand side, and a hundred thousand or so more rows added at the bottom. This diagram would then represent a small part of the top left corner of an enormous grid packed with lexical information. Experts have calculated that if you printed out the rest of this table in full, you would end up with a piece of paper approximately 5.5m wide, and 2.4km long – so you could probably walk round it in just under an hour. Using FLEX, which itself uses a database management system to access the information in the grid, you can extract tiny bits of information, or long and detailed lists, just as you please. When you create a *lexicon*, you're essentially creating a little dictionary, designed to your own specifications.

Unlike a dictionary, you can use keys other than the headword when you look something up in a database. On a simple level, this means you can look up the verb *walk*, instead of the noun *walk*. On another level, it means that you can get a list of all the verbs in the database, excluding all the other words which are not verbs. The individual printed *paragraphs* for each word in a dictionary are fixed, but the corresponding *rows* in a computer database can be moved about and rearranged just as you want them. So, it's possible to create a lexicon like the one illustrated below by using FLEX restrictions. You simply state that you want to see all the words which have the word class code `V`, and you can then get as much information as you like about the verbs in your list. The example below shows a list of verbs with their pronunciations:

Headword	**Transcription**	
`abandon`	`@-'b&n-d@n`	
`abase`	`@-'beIs`	
`abash`	`@-'b&S`	
`abate`	`@-'beIt`	

Since you've specified that you only want verbs in your list, there's no need to put the word class code column on display. The computer uses it in preparing the list, but you don't have to look at it – you'd just get a list of V's. The possibilities for creating all sorts of lexicons are seemingly endless. You have hundreds of columns to choose from, most of which contain information you might want to inspect on your screen. Other columns contain information which can be used to *control* what is shown on your screen or in your file: the word class column in the illustration above, for example, or the Inflectional features columns under the morphology of Dutch wordforms which simply say 'yes' when a wordform does have a particular inflectional feature, and 'no' when it doesn't. A screen display of those columns isn't particularly interesting, but a file of wordforms created using the information they contain may well be very interesting.

And there are still more possibilities. If you want to build up a lexicon which contains words with, say, certain phonetic features in common, then FLEX lets you do it with the help of the *pattern matcher*. For example, you might want to see the words which contain in a non-initial position syllables beginning with a dental plosive or dental fricative. The required pattern is `@*-[tdTD]@*`, which when applied to a syllabified phonetic transcription column, tells FLEX to find transcriptions which consist of zero or more characters of any sort (`@*`) characters followed by a syllable marker (`-`) followed by one of the dental phonemes `t, d, T` or `D` followed by zero or more characters of any sort. The resulting lexicon would start off something like this:

Headword	**Class**	**Phonetics**	
abandon	N	@-'b&n-d@n	
abandon	V	@-'b&n-d@n	
abandoned	A	@-'b&n-d@nd	
abandonment	N	@-'b&n-d@n-m@nt	

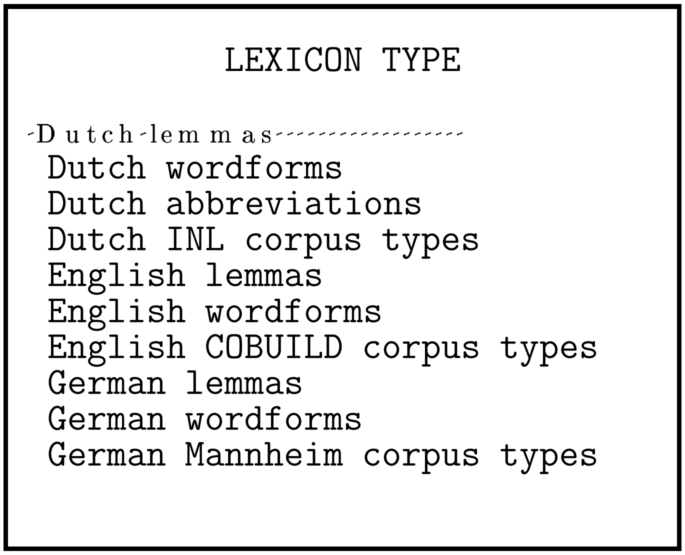
Compared to a dictionary, then, a lexicon-based database system like FLEX has significant advantages for the linguistic researcher. You have a great store of linguistic detail available, and the means to tailor and craft it according to the research you have to do, rather than being limited to the inflexible format of a dictionary. And while for a beginner the prospect of learning to use FLEX may at first seem daunting, the alternative—page by page inspection of a dictionary—should be enough to convince any who doubt the usefulness of it.

The next section describes the differences between the various *types* of lexicon you can develop within FLEX. The columns that you can add to your lexicons are described in the three *Linguistic Guides*. Read them carefully as you plan the construction of your personal lexicons.

2 LEXICON TYPES

When you work with FLEX, you create your own *lexicons*. The database which FLEX accesses is enormous, and you only ever see a tiny fraction of it on your screen at any one time. Lexicons allow you to narrow down the information you get on screen or in a file, so that you have a manageable view of the parts of the database which interest you most.

CELEX has several databases available for your use, and you can create lexicons using any one of them. When you're asked what type of lexicon you want, you are in fact being asked 'from which part of which database would you like your information?' The `LEXICON TYPE` menu is the menu screen that lets you choose:



```
          LEXICON TYPE

Dutch lemmas
 Dutch wordforms
 Dutch abbreviations
 Dutch INL corpus types
 English lemmas
 English wordforms
 English COBUILD corpus types
 German lemmas
 German wordforms
 German Mannheim corpus types
```

The most basic choice is, obviously, whether you want information on Dutch, English or German. After that, the choice you make depends on the work you want to do. For now, remember that in this context the terms *lemma*, *wordform*, *abbreviation* and *corpus type* refer to the database equivalent of a bold-type dictionary entry. Each database—and thus each of your lexicons—holds particular sorts of entries. So, if you choose a lemma lexicon, it is as if you are using a dictionary where every entry represents a full inflectional paradigm, making a lemma lexicon the closest thing to a normal dictionary that CELEX offers. If you choose a wordform lexicon on the other hand, the dictionary entries are the inflectional forms themselves: every entry (or *row*) deals specifically with one flection, something which conventional dictionaries never do. In effect you have a dictionary which contains all the words which are used in natural language. Naturally enough, an abbreviations lexicon is like a dictionary of abbreviations: the entry is always an abbreviated form of some sort. And a corpus types lexicon contains rows specific to each distinct item in one of the text corpora used to extract lexical features for each of the languages. Read on to discover more about each lexicon type. Once you've read and understood it, you'll be able to choose the lexicon type most appropriate to whatever task you have to carry out.

2.1 DUTCH LEMMAS

When you look up a word in a dictionary, you don't always find the exact word you want. Quite often you come across a shorter version in bold type, which represents the particular word you had in mind, as well as various other forms which you know intuitively 'belong' to the same 'word'. Thus when you're interested in a word like *loopt*, you know that you will find all the information you want under the bold-type entry for the verb *lopen*. These bold-type words in dictionaries are called *headwords* or *canonical forms*, since they represent what can be called the full *canon* or *paradigm* of inflections: *lopen* is the headword which stands for the wordforms *loop, loopt, lopen, liep, liepen, gelopen, lopend, lopende, lopenden,* and (occasionally) *lope*.

Because headword forms have become so firmly established, people often presume that there must be something linguistically special about them. Usually, the shortest form in the inflectional paradigm is the headword – but not always: for verbs, Dutch dictionaries use the present tense plural form (which is also the infinitive), even though the present tense first person singular is shorter: consider *openen* as against *open*. Many linguists in fact prefer to use the shorter form as the canonical form in their work, because all the other forms can be made from this basic form by adding inflectional affixes (though this is putting it very simply, of course). So, what form is used as the canonical form in the CELEX databases?

As far as CELEX is concerned, a *lemma* is an abstract way of representing a whole inflectional paradigm. The dictionary headword, as described above is one form a lemma can take to represent a 'word' in all its inflected forms. It is possible—but probably not very helpful for humans—to signify the 'word' by some completely different word, or even a number; anything will do, so long as it is understood to represent the whole inflectional paradigm. A lemma is that 'underlying' form; it doesn't really exist, except for use in databases and dictionaries. It looks like a real word, but in fact it's just a convenient way of expressing something bigger.

Since the lemma is an abstract notion, we need now to identify the more concrete forms it can take. Two are used in the databases, and you can choose for yourself which one you use. First, there is the *headword*, which corresponds exactly to the traditional lexicographic headword used in dictionaries. And second, there is the *stem*, the form which most linguists prefer. Since the forms headwords and stems take are often assumed rather than explicitly stated, table 1 defines what headwords and stems look like in the CELEX Dutch database. It holds true for just about every lemma; there are very few exceptions.

Table 1 shows that CELEX headwords and stems are very similar to the traditional lexicographic forms which are normally used in dictionaries. The major exception is the stem of a verb. One other important feature of stems is the possibility of using so-called *abstract stems*, which some linguists like to use in certain circumstances, again for reasons concerning the formation of flections. These forms are dealt with in the *Dutch Linguistic Guide*.

There is still one major difference between dictionary entries and CELEX lemmas, however: CELEX lemmas are never distinguished solely on the basis of meaning. In a dictionary, there might be two entries for the noun *bank*, one explaining that it means a sofa, the other that it means a financial institution. In the CELEX database, there is only *one* lemma for the noun *bank*, and thus it gets only *one* row in the database (which corresponds to an entry or sub-paragraph in a dictionary). On what basis, then, does CELEX differentiate between lemmas? There are six possible criteria. If two potential lemmas are the same on all six points, then they are taken to be one single lemma. This remains true *even if* the

Word Class	Lemma	
	Headword	Stem (where different from the headword)
Noun	Nominative singular, except for *pluralia tantum* which use the nominative plural. Diminutive forms are not treated as separate lemmas.	As for the headword, except that *pluralia tantum* are given a nominative singular form.
Adjective	The shortest positive form.	
Quantifier/Numeral	For a number, the cardinal and ordinal forms are two separate *lemmas*, and are used as the two headwords. For quantifiers, the shortest form is used.	
Verb	Infinitive.	First person singular present tense form (non-separated form as used in relative clauses).
Determiner	Nominative form.	
Pronoun	Shortest form.	
Adverb	Shortest form.	
Preposition	The only form.	
Conjunction	The only form.	
Interjection	The only form.	

Table 1: Celex canonical forms for Dutch

two words differ in meaning. If, however, they differ on any one criterion, *and* differ in meaning, then they are treated as two separate lemmas. The six distinguishing criteria are as follows:

1. Orthography of the wordforms. The adjectives *rauw* and *rouw* are two different lemmas because they are spelt differently and have a different meaning.

2. Syntactic class. The noun *wit* and the adjective *wit* are different lemmas because they each have a different word class. Sometimes the difference in word class is itself the only way a difference in meaning is indicated.

3. Gender. The noun *de pas* (meaning the pace) and the noun *het pas* (meaning a spirit level) are different lemmas because they differ in gender and also in meaning.

4. Inflectional paradigm. The verb *malen* (meaning to crush) and the verb *malen* (to be delirious) are two different lemmas because the first has the past participle *gemalen*, while the second has the past participle *gemaald*, and they differ in meaning.

5. Morphological structure. The noun *koker* (someone who cooks – *kook + er)* and the noun *koker* (a cylindrical object; a monomorphemic word) are two distinct lemmas because they differ in their derivational morphological structure and their meaning.

6. Pronunciation of the wordforms The noun *kip* (meaning chicken) and the noun *kip* (meaning the act of dumping) would be different lemmas because in standard Dutch the first is pronounced [kɪp] and the second is pronounced [kiːp], and they differ in meaning.

It should be clear by now that a lemma is a notional representation of an inflectional paradigm, and that the forms CELEX gives to a lemma are headwords and stems. These forms are convenient representations only, which exist to make life easier for dictionary users and computer lexicon builders. A lemma lexicon contains general information about an inflectional paradigm, similar to the way an ordinary dictionary does. Within a lemma lexicon, the lemmas are given as stems or headwords, and you can choose either form when you make your lexicon.

For lemmas, there is orthographic, phonetic, morphological, syntactic and frequency information available. In the appendices you can find diagrams which give an overview of the lemma columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of lemmas currently available in the CELEX databases.

2.2 DUTCH WORDFORMS

Wordforms can be thought of as *real* words: we use them every day, in speech and in writing. You are at this moment reading English wordforms. Even the shortest forms in an inflectional paradigm are wordforms (as opposed to lemmas) simply because they are working parts in the language. When you use a wordforms lexicon, it is as if you are looking in a dictionary which lists every possible word, instead of abstract forms which represent particular sets of words. For this reason, while a lexicon of type lemma only yields *kat*, a wordforms lexicon gives you all the occurring forms of the lemma – for example, both *kat*, *katje* and *katten* are wordforms.

Sometimes individual wordforms (in this case verbs) can be split into two distinct parts, depending on the way the sentence is formed. Both the whole form and the separated form are included in the wordforms information. For example, *bel op* (' ik bel jou op ') and *opbel* (' als ik jou opbel... ') can both be included in a wordforms lexicon.

Information about each wordform's lemma is supplied too, so that this lexicon type also covers all the information a normal lemma lexicon can contain. You can include such information by going to the morphology section of the `ADD COLUMNS` menus, where you can choose to include `Stem information` and/or `Inflectional features`.

For wordforms, there is orthographic, phonetic, morphological, and frequency information available. You can also use all the information relating to the lemma that each wordform belongs to. In the appendices you can find diagrams which give an overview of the wordform columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of wordforms currently available in the CELEX databases.

2.3 DUTCH ABBREVIATIONS

Abbreviations are shortened forms of words or names. For example, *gem.* is a shortened form of *gemiddeld* and *gemeubileerd.* Other abbreviations are composed of the first letter from each word in a name – *BBC* is thus an abbreviation for *British Broadcasting Corporation.* Many such abbreviations have two spellings, one with and one without dots, for example.

The abbreviations given are drawn from the *Van Dale Groot Woordenboek van Hedendaags Nederlands* and the sizeable text corpus of the INL (*Instituut voor Nederlandse Lexicologie* – the Institute for Dutch Lexicology based in Leiden).

For abbreviations, there is orthographic and frequency information available. In the appendices you can find diagrams which give an overview of the abbreviation columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of abbreviations currently available in the CELEX databases.

2.4 DUTCH INL CORPUS TYPES

INL *tokens* are strings in the large INL text corpus of modern Dutch, and here a string can be taken to mean at least one alphabetic character in series with zero or more other alphanumeric characters, delimited at either end by a space. (So, for example, *zes* is a token, and so is *6de*, but *6* by itself is not, because it does not contain at least one alphabetic character. This applies to all numerals.) INL corpus *types* are *distinct tokens*; that is, not a list of the many million tokens, but a representative list that includes once each separate token which occurs in the corpus.

In fact, the criteria for inclusion in the type list can be more closely defined. The INL corpus is made up of many different contemporary texts, or looking at it in another way, several millions of *tokens.* Included in the CELEX INL corpus type list, then, are all the *types* which occur in at least two different corpus texts.

Corpus types complement the lemma and wordform information; it's safe to say that amongst them, you can find almost every item which occurs in written text. Unlike the dictionary-style lemma and wordform lexicons, no syntactic, morphological, or phonetic information is available. What you do have is a database of real-life words, distinguished on the basis of their orthography, with detailed information on their frequency.

In the appendices there are diagrams which give an overview of the corpus type columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of types currently available in the CELEX databases, and the size and contents of the INL corpus.

2.5 ENGLISH LEMMAS

When you look up an English word in a dictionary, you don't always find the particular form you want. Instead, you come across a shorter version in bold type, which represents the particular form you had in mind, along with various other similar forms which you intuitively know 'belong' to the same 'word'. So when you're interested in a word like *walking*, you know that you can find lots of information about it under the bold-type entry for the verb *walk*. These bold-type words in dictionaries are called *headwords* or *canonical forms*, since they represent what can be called the full *canon* or *paradigm* of inflections: *walk* is the headword which stands for the wordforms *walk*, *walks*, *walking* and *walked*.

As far as CELEX is concerned, such a form is a *lemma*, an abstract way of representing a whole inflectional paradigm. The dictionary headword, as described above is one form a lemma can take to represent a 'word' in all its inflected forms. It is possible—but probably not very helpful for humans—to signify the 'word' by some other word, or even a number; anything will do, so long as it is understood to represent the whole inflectional paradigm. A lemma is that 'underlying' form; it doesn't really exist, except for use in databases and dictionaries. It looks like a real word, but in fact, it's just a convenient way of expressing something bigger.

In an English lemma lexicon, the lemma is given in the form of the traditional lexicographic headword. This is in contrast to Dutch lemma lexicons, where the lemma can take the form either of the traditional headword or of a stem, which is a form more suitable for most linguistic research. No such complications apply to English, however: the 'underlying' lemma always becomes the traditional headword when it comes to the 'surface'. Table 2 opposite sets out exactly which form that is for each lemma. It is almost always accurate; there are only a few exceptions, such as the verb *to be* which is given as *be* in accordance with the long-standing tradition.

There is one major difference between dictionary entries and CELEX English lemmas, however: CELEX lemmas are never distinguished solely on the basis of meaning. In a dictionary, there might be two entries for the noun *bank*, one explaining that it means the land at the side of a river, the other that it

Word Class	Headword
Noun	The singular form, except for *pluralia tantum* which use the plural form.
Adjective	The positive form.
Quantifier/Numeral	For a number, the cardinal and ordinal forms are two separate *lemmas*, and are used as the two headwords. For quantifiers, the only form is used.
Verb	First person singular present tense form.
Pronoun	The only form.
Adverb	The positive form.
Preposition	The only form.
Conjunction	The only form.
Interjection	The only form.

Table 2: Canonical forms for English lemmas

means a financial institution. In the CELEX database, there is only *one* lemma for the noun *bank*, and thus it gets only *one* row in the database (which corresponds to an entry or sub-paragraph in a dictionary). On what basis, then, does CELEX differentiate between lemmas? There are five possible criteria. If two potential lemmas are the same on all five points, then they are considered as belonging to one lemma. This remains true *even if* the two words differ in meaning. If, however, they differ on any one criterion, *and* differ in meaning, then they are treated as two separate lemmas. The five distinguishing criteria are as follows:

1. Orthography of the wordforms. The nouns *peek* and *peak* are two different lemmas because they are spelt differently and have a different meaning.

2. Syntactic class. The adjective *meet* and the adverb *meet* are different lemmas because they each have a different word class and a different meaning. Sometimes the difference in word class is itself the only way a difference in meaning is indicated, as with words like *water* (verb) and *water* (noun).

3. Inflectional paradigm. The noun *antenna* (meaning radio aerial) and the noun *antenna* (an anatomical feature of some insects) are two different lemmas because the first has the plural *antennas*, while the second has the plural *antennae*, and they differ in meaning.

4. Morphological structure. The noun *rubber* (someone or something that rubs – *rub + er)* and the noun *rubber* (the elastic substance; a monomorphemic word) are two distinct lemmas because they differ in their derivational morphological structure and their meaning.

5. Pronunciation of the wordforms. The verb *recount* (meaning count again) and the verb *recount* (meaning to tell a tale) would be different lemmas because the first is pronounced [ˈriː-kaʊnt] and the second is pronounced [rɪ-ˈkaʊnt], and they differ in meaning.

For lemmas, there is orthographic, phonetic, morphological, syntactic and frequency information available. In the appendices you can find diagrams which give an overview of the lemma columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of lemmas currently available in the CELEX databases, and the sources from which the information derives.

2.6 ENGLISH WORDFORMS

Wordforms can be thought of as *real* words: we use them every day, in speech and in writing. You are at this moment reading English wordforms. Even the shortest forms in an inflectional paradigm are wordforms (as opposed to lemmas) simply because they are working parts in the language. When you use a wordforms lexicon, it is as if you're looking in a dictionary which lists every possible word, instead of abstract forms which represent particular sets of words. For this reason, while a lexicon of type lemma only yields *dog*, a wordforms lexicon gives you all the occurring forms of the lemma – for example both *dog* and *dogs* are wordforms.

Information about each wordform's lemma is supplied too, so that this lexicon type also covers all the information a normal lemma lexicon can contain. You can include such information by going to the morphology section of the `ADD COLUMNS` menus, where you can choose to include `Lemma information` and/or `Inflectional features`.

For wordforms, there is orthographic, phonetic, morphological and frequency information available. You can also use all the information relating to the lemma that each wordform belongs to. In the appendices you can find diagrams which give an overview of the wordform columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of wordforms currently available in the CELEX databases, and the sources from which the information derives.

2.7 ENGLISH COBUILD CORPUS TYPES

COBUILD *tokens* are strings in the large COBUILD text corpus of modern English, and here a string can be taken to mean at least one alphabetic character in series with zero or more other alphanumeric characters, delimited at either end by a space. (So, for example, *six* is a token, and so is *6th*, but *6* by itself is not, because it does not contain at least one alphabetic character. This applies to all numerals.) COBUILD corpus *types* are *distinct tokens*; that is, not a list of the many million tokens, but a representative list that includes once each separate token which occurs in the corpus.

Corpus types complement the lemma and wordform information; it's safe to say that amongst them, you can find almost every item which occurs in written text. Unlike the dictionary-style lemma and wordform lexicons, no syntactic, morphological, or phonetic information is available. What you do have is a database of real-life words, distinguished on the basis of their orthography, with detailed information on their frequency.

In the appendices there are diagrams which give an overview of the corpus type columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of types currently available in the CELEX databases, and the size and contents of the COBUILD corpus.

2.8 GERMAN LEMMAS

When you look up a German word in a dictionary, you don't always find the particular form you want. Instead, you come across a different form of the word in bold type, which represents the particular form you had in mind, along with various other similar forms which you intuitively know 'belong' to the same 'word'. So when you're interested in a word like *gegangen*, you know that you can find lots of information about it under the bold-type entry for the verb *gehen*. These bold-type words in dictionaries are called *headwords* or *canonical forms*, since they represent what can be called the full *canon* or *paradigm* of inflections: *gehen* is the headword which stands for the wordforms *gehe, geht, gehst, gehest, gehen, gehn, gehet, gehend, ging, ginge, gingest, ginget, gingst, gingt* and *gegangen*. Many linguists however prefer to use the shorter form as the base form, the stem form, in their work, because all the other forms can be made from this basic form by adding inflectional affixes (though this is putting it very simply, of course).

We at CELEX use the notion *lemma* as an abstract way of representing a whole inflectional paradigm. Since the lemma is an abstract notion, we need now to identify the more concrete forms it can take. Two are used in the databases, and you can choose for yourself which one you use. First, there is the *headword*, which corresponds exactly to the traditional lexicographic headword used in dictionaries. And second, there is the *stem*, the form which most linguists prefer. Since the forms headwords and stems take are often assumed rather than explicitly stated, table 3 defines what headwords and stems look like in the CELEX German database.

There is one major difference between dictionary entries and CELEX German lemmas, however: CELEX lemmas are never distinguished solely on the basis of meaning. In a dictionary, there might be five entries for the noun *Absatz*, the first explaining that it means a piece of text, the second that it means a part of a shoe, the third that it means sedimentary deposit, the fourth that it means sales, and the fifth that it means the landing. In the CELEX database, there is only *one* lemma for the noun *Absatz*, and thus it gets only *one* row in the database (which corresponds to an entry or sub-paragraph in a dictionary). On what basis, then, does CELEX differentiate between lemmas? There are six possible

Word Class	Lemma	
	Headword	Stem (where different from the headword)
Noun	Nominative singular, except for *pluralia tantum* which use the nominative plural. Diminutive forms are not treated as separate lemmas.	As for the headword, except that *pluralia tantum* are given a nominative-singular-like form.
Adjective	The shortest positive form.	
Quantifier/Numeral	For a number, the cardinal and ordinal forms are two separate *lemmas*, and are used as the two headwords. For quantifiers, the shortest form is used.	
Verb	Infinitive.	Infinitive without the (e)n-ending.
Article	The only forms are 'der' and 'ein'	The only forms are 'der' and 'ein'
Pronoun	The nominative singular forms.	The shortest form
Adverb	Shortest form.	
Preposition	The shortest form.	
Conjunction	The only form.	
Interjection	The only form.	

Table 3: Celex canonical forms for German

criteria. If two potential lemmas are the same on all six points, then they are considered as belonging to one lemma. This remains true *even if* the two words differ in meaning. If, however, they differ on any one criterion, *and* differ in meaning, then they are treated as two separate lemmas. The six distinguishing criteria are as follows:

1. Orthography of the wordforms. The nouns *fallen* and *fällen* are two different lemmas because they are spelt differently and have a different meaning.

2. Syntactic class. The adjective *anderweitig* and the adverb *anderweitig* are different lemmas because they each have a different word class. Sometimes the difference in word class is itself the only way a difference in meaning is indicated, as with words like *ledern* (verb) and *ledern* (adjective).

3. Inflectional paradigm. The noun *Bank* (the bank in the park) and the noun *Bank* (die Deutsche Bank) are two different lemmas because the first has the plural *Bänke*, while the second has the plural *Banken*, and they differ in meaning.

4. Morphological structure. The noun *Messer* (knife) and the noun *Messer* (meter or measurer) are two distinct lemmas because they differ in their morphological structure and their meaning.

5. Pronunciation of the wordforms. The noun *Band* (meaning a group of people making music) and the noun *Band* (meaning the relationship between two people) would be different lemmas because the first is pronounced [bEnt] and the second is pronounced [bant], and they differ in meaning.

6. Gender of the wordforms. The noun *das Tor* (the gate) and *der Tor* (the mad person) will be different lemmas because the gender of the first noun is neuter and the gender of the second one is masculine.

For lemmas, there is orthographic, phonetic, morphological, syntactic and frequency information available. In the appendices you can find diagrams which give an overview of the lemma columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of lemmas currently available in the CELEX databases, and the sources from which the information derives.

2.9 GERMAN WORDFORMS

Wordforms can be thought of as *real* words: we use them every day, in speech and in writing. You are at this moment reading English wordforms. Even the shortest forms in an inflectional paradigm are wordforms (as opposed to lemmas) simply because they are working parts in the language. When you use a wordforms lexicon, it is as if you're looking in a dictionary which lists every possible word, instead of abstract forms which represent particular sets of words. For this reason, while a lexicon of type lemma only yields *Kind*, a wordforms lexicon gives you all the occurring forms of the lemma – for example *Kind*, *Kindes*, *Kinde*, *Kinder* and *Kindern* are all wordforms.

Information about each wordform's lemma is supplied too, so that this lexicon type also covers all the information a normal lemma lexicon can contain. You can include such information by going to the morphology section of the `ADD COLUMNS` menus, where you can choose to include `Lemma information` and/or `Inflectional features`.

For wordforms, there is orthographic, phonetic, morphological and frequency information available. You can also use all the information relating to the lemma that each wordform belongs to. In the appendices you can find diagrams which give an overview of the wordform columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of wordforms currently available in the CELEX databases, and the sources from which the information derives.

2.10 GERMAN MANNHEIM CORPUS TYPES

MANNHEIM *tokens* are strings in the MANNHEIM text corpus of modern German, and here a string can be taken to mean at least one alphabetic character in series with zero or more other alphanumeric characters, delimited at either end by a space. (So, for example, *fünfzehn* is a token, and so is *15jährige*, but *15* by itself is not, because it does not contain at least one alphabetic character. This applies to all numerals.) MANNHEIM corpus *types* are *distinct tokens*; that is, not a list of the many million tokens, but a representative list that includes once each separate token which occurs in the corpus.

In fact, the criteria for inclusion in the type list can be more closely defined. The MANNHEIM corpus is made up of many different contemporary texts, or looking at it in another way, several millions of *tokens*. Included in the CELEX MANNHEIM corpus type list, then, are all the *types* which occur in at least two different corpus texts.

Corpus types complement the lemma and wordform information; it's safe to say that amongst them, you can find almost every item which occurs in written text. Unlike the dictionary-style lemma and wordform lexicons, no syntactic, morphological, or phonetic information is available. What you do have is a database of real-life words, distinguished on the basis of their orthography, with detailed information on their frequency.

In the appendices there are diagrams which give an overview of the corpus type columns that are described in detail in the *Linguistic Guides*, as well as some basic information about the number of types currently available in the CELEX databases, and the size and contents of the MANNHEIM corpus.