CELEX frequencies – What do they mean and how to use them?

Billi Randall
Centre for Speech and Language, U of Cambridge

The database from which the frequencies are taken.
The frequencies are taken from the early 1991 version of the COBUILD corpus. There are about 17.9 million tokens in this corpus. The COBUILD programme is ongoing and currently has a Bank of English with over 330 million words in it. The COBUILD programme is based at the University of Birmingham, and now works in conjunction with Harper Collins.

Of the 17.9 million tokens, most are written and about 1 million are tokens of spoken English. Written sources include newspapers and books (both English and American, I think). Spoken sources include the BBC World Service and taped telephone conversations. Exact descriptions of the sources are not in the CELEX manual, and the information on the COBUILD website refers only to the current corpus.

The different types of frequencies
There are two CELEX databases devoted to frequency. The first is the wordform frequency database (efw) and the second is the lemma frequency database (efl). Each of these databases lists the same type of frequency information. However, these may be in different fields in the two databases.

Frequencies listed:
1. The overall frequency count from the corpus. Each identical string is counted. Disambiguation is done by hand. This is a way of getting the noun and verb frequencies. However, when the corpus count for a string is more than 100, disambiguation is estimated on the basis of the first 100 strings. When a disambiguation is estimated, there is an additional field, called Cobdev, which gives an indication of how good the estimate is. Generally, if the deviation is equal to or greater than the frequency listed, then the estimate is not good.

2. There is a field that gives the frequency per million for any given wordform or lemma. This is a figure that is easier to understand than the total CELEX frequency. Also, many other frequency databases are based on a million tokens.

3. Another frequency listing is the logarithmic frequency of each word in the database. This reduces the differences between high frequency words, while maintains the difference between low frequency words. This recognises the fact that the difference between a frequency of 1 and 2 is more important than the difference between a frequency of 2001 and 2002.

How are the frequencies worked out?
The wordform frequency lists the appropriate frequency for each individual word string. There is an entry for ‘eye’ and a separate entry for ‘eyes’. For verbs, there are separate entries for present and past tenses and for each person.

Let’s look at the word ‘arrested’
Here is the full entry in the word form database
And here we have selected just the per million frequency, which in this database is $6.

arrested:5
arrested:5
arrested:5
arrested:5
arrested:5

As you can see, this gives no clue as to what each of these different tokens of ‘arrested’ represent. If we look in the morphology of word forms database, more is revealed. Helpfully, the full listing frequency is given in field 3. Field 6 is the clue to what the word actually represents. We can see that there is the first, second and third person singular, a plural of the past tense (a= past tense; S = singular; P = plural; and a past participle (pa). I can find no justification for listing only one plural form (perhaps, it is because the 2nd person singular of the present tense differs from the 1st and 3rd person singulars.

3599\arrest\92\2074\a1S\@+ed
101147\arrest\92\2074\a2S\@+ed
118465\arrest\92\2074\a3S\@+ed
135375\arrest\92\2074\aP\@+ed
152031\arrest\92\2074\pa\@+ed

This means that each of the frequencies listed in the word form database has a meaning – they are not mere repetitions. If you are not sure which part of speech is shown in the wordform frequency database, you will have to cross reference it in the morphology of wordforms database.

Now let us turn to the stem of this verb. Here is it’s full listing in the wordform frequency database.

3597\arrest\2073\211\12\1.0792\209\13\1.1139\2\2.301
3598\arrest\2074\30\130\2\301\30\2.301\0\0\0
101146\arrest\2074\30\130\2\301\30\2\301\0\0\0
118464\arrest\2074\30\130\2\301\30\2\301\0\0\0
135374\arrest\2074\30\130\2\301\30\2\301\0\0\0

And here is the per million frequency only.

arrest\12
arrest\2
arrest\2
arrest\2
arrest\2

Once again we need to turn to the morphology database to disambiguate what these different tokens represent.

3597\arrest\211\2073\S@  
3598\arrest\30\2074\i@  
101146\arrest\30\2074\e1S@  

We can see that the first entry is a singular noun, the second is the infinitive, the third the first person singular of the present tense, the fourth is the second person singular and the fifth is the plural of the present tense.
This, of course, leaves out the third person singular.
Here is the entry for arrests, firstly, per million in the wordform frequency database

arrests\4
arrests\0

And, secondly, in the morphology wordform database.

arrests\3605\arrests\74\2073\P\@+s
arrests\3606\arrests\3\2074\e3S\@+s

This shows us that the first entry is the plural form of the noun and the second is the third person singular of the present tense of the verb.

Now, there is one part of the verb conjugation remaining. This is the present participle, arresting.

In the wordform frequency database, we get the following output.
arresting\1
arresting\1

Looking in the morphology database, they are disambiguated in the following way.

3603\arresting\18\2074\pe\@+ing
3604\arresting\21\2077\b\@

The first entry is the present participle and the second entry is an adjective.
We have now gone through all the forms of the verb ‘arrest’ that are listed in the CELEX database. From these listings we can see how the lemma frequency is worked out.

The lemma frequency of ‘arrest’ per million = 34;
Made up from wordform frequencies arrested x 5; arrest x 4; arresting x 1;

\[
\begin{align*}
5 \times 5 & = 25 \\
4 \times 2 & = 8 \\
1 \times 1 & = 1 \\
\text{Total} & = 34
\end{align*}
\]

Now we need to return to the wordform frequencies and look at how the creators of CELEX arrived at the different frequency listings for each part of speech.

For verbs, the number of times that a word string appears in the corpus is divided by the number of types of string that occurs with that verb. There are four for the person inflections, and possibly one for the past participle (remember that the past participle is not always the same as the simple past tense (took – taken; spoke – spoken, but made – made). This happens regardless of whether or not they actually occur an equal number of times. Also
remember that if a token appears more than 100 times in the corpus, its distribution across types (i.e. verbs or nouns) is estimated on the basis of the first 100 occurrences of that string. This means that the way that inflections are divided across parts of speech is an estimation of an estimation.

As you can see from all that has been explained above, it is important to decide at the beginning of setting up an experiment which frequency you are most interested in and whether or not differences in frequency between different parts of speech are important.

If at all possible, you should use the lemma frequency, with all types summed, as it's verbiness or nounness are not important, as this will be the most accurate. If you are interested in the difference between frequencies in nouns and verbs and you are also forced to use wordform frequencies, you will need to disambiguate the various types of frequency and then sum up the different verbal types.

Remember this database is nearly a decade old, and that language changes over time. There is no listing in CELEX for ‘world wide web’ or ‘internet’, while the word ‘policy’ has a frequency of 226 per million. I doubt whether our subjects use the word policy more often than they use the word internet.

Therefore, use CELEX as a tool and a very useful one, but do not take everything it says as gospel!!