# 3 - CELEX

## Lemma

In some circumstances it can be useful to group together different forms of the same word; e.g. treat "damage", "damaged" and "damaging" as variants of the same thing for the purposes of frequency-counting and other analyses.

The demo database has been configured to tag each token with it's root form or 'Lemma'. To do this, LaBB-CAT has been integrated with the CELEX lexicon, which can be purchased from the Linguistic Data Consortium (LDC) and includes lemma, part of speech, morphological, phonological, and frequency information for English, German, and Dutch.

There is a *lemma* layer configured, which looks up each word token in the CELEX lexicon, and tags it with its lemma.

For words that are missing from CELEX, LaBB-CAT is configured to instead tag the word with it's 'stem' according to the Porter Algorithm (Porter, 1980, *An algorithm for suffix stripping*, Program, Vol. 14, no. 3, pp 130-137, or http://www.tartarus.org/~martin/PorterStemmer).

1. In the transcript you have open, under the "projects" heading, tick the *celex* project.
   A number of extra layer options will appear in the layer list.
2. Tick the *lemma* layer.
   When the transcript re-loads, you'll see that each word is tagged with its lemma; in some cases the lemma is the same as the word-form, and in other cases, the lemma has suffixes stripped off, etc.
3. Search for the word "damage" on the page (in most browsers, `Ctrl` + `F` or some similar keyboard combination allows you to search for text on the current page).
   You should see that variants like "damage", "damaged" and "damaging" are all tagged with the same lemma: "damage".
4. Now tick the *frequency* project.
   Three layer options appear; *word frequency*, *lemma frequency*, and *liwc*. You have already seen (and searched) the *word frequency* layer.
   The *lemma frequency* layer is similarly generated by the Frequency Layer Manager, but instead of counting up raw word forms from the *orthography* layer, it counts based on the *lemma* layer.
5. Tick both *word frequency* and *lemma frequency*
6. Find the word "damaging" in the text.
   You'll see that, although the word-form is very low frequency, the lemma frequency is somewhat higher (as you'd expect in a corpus of earthquake stories!).

The Frequency Layer Manager also keeps a straight word-list with word counts for each corpus…

7. Click the *home* menu option at the top.
8. Click the *Frequency Layer Manager* icon.
9. You will see a drop-down box with each frequency layer in it.
   Select *Lemma Frequency* and press *Select*.
10. Press the *Export* button at the bottom.
11. Save and open the resulting CSV file.
    You will see an alphabetical list of all the distinct lemmas in the database, and next to each, a count of the number of tokens of that type.

### Morphology

There is other information in CELEX that can be used to tag words. Let's say you're interested in the morphological suffix *~ing*. If we search for *.*ing* on the *orthoghraphy* layer, we'll get a number of false positives ...

1. Select the *search* menu option.
2. Search for `.*ing` on the *orthography* layer.
   You will see that the results include words like like "thing" and "everything" whose "ing" is part of the base word, not a morphological affix.
   Leave the results tab open, so you can compare these results with the next search ...
3. Swap back to the search matrix page, tick the 'celex' project, and add the *morphology* layer to the search matrix.
4. Now do a search on the *morphology* layer of words ending in *+ing*, and compare the results with the orthography-based search.

> **!** Important
>
> Remember that in regular expressions the 'plus' character + has a special meaning - it means "one or more of the previous thing". But we are now searching for actual literal + characters.
> In order to search for a literal "+" in the annotation, you have to 'escape' the +. Consult the *regular expressions* help page to figure out how to do that.

The results should now contain only words for which the *~ing* is a morphological suffix.

### Phonology

The CELEX lexicon includes phonological information, so we can tag each word with its phonemic transcription, and view/search the pronunciations of words.

1. Click the *transcripts* link on the menu.

2. Click the name of the first transcript listed, to display the transcript text.
3. Tick the *celex* project.
4. Tick the *phonemes* layer.

You will see that each word is tagged with its phonemic transcription using the International Phonetic Alphabet (IPA). However, CELEX doesn't use IPA symbols directly, it actually uses the 'DISC' encoding for phonemes, which uses ordinary 'typewriter' characters (ASCII), and uses exactly one character per phoneme.

The IPA symbols are being displayed by LaBB-CAT to provide a linguist-friendly representation of the phonemic transcription. But you can see the underlying DISC characters by selecting the 'ASCII' option on the layer in the transcript.

5. Select 'ASCII' on the *phonemes* layer, to see what CELEX is actually producing.
   You may find that this is somewhat harder to read. Diphthongs are generally represented by digits, schwa is "@", and various other characters are used to represent affricates, etc.

It's nice to display the IPA symbols, but it's important to understand the DISC symbols (shown in the table below), because they are what we have to use when searching on the *phonemes* layer, which we are going to try now.

| IPA | DISC | | IPA | DISC | |
|---|---|---|---|---|---|
| p | p | **p**at | ɪ | I | K**I**T |
| b | b | **b**ad | ɛ | E | DR**E**SS |
| t | t | **t**ack | æ | { | TR**A**P |
| d | d | **d**ad | ʌ | V | STR**U**T |
| k | k | **c**ad | ɒ | Q | L**O**T |
| g | g | **g**ame | ʊ | U | F**OO**T |
| ŋ | N | ba**ng** | ə | @ | **a**nother |
| m | m | **m**at | iː | i | FL**EE**CE |
| n | n | **n**at | ɑː | # | f**a**ther |
| l | l | **l**ad | ɔː | $ | TH**OU**GHT |
| r | r | **r**at | uː | u | G**OO**SE |
| f | f | **f**at | ɜː | 3 | N**UR**SE |
| v | v | **v**at | eɪ | 1 | F**A**CE |
| θ | T | **th**in | ɑɪ | 2 | PR**I**CE |
| ð | D | **th**en | ɔɪ | 4 | CH**OI**CE |
| s | s | **s**ap | əʊ | 5 | G**OA**T |
| z | z | **z**ap | ɑʊ | 6 | M**OU**TH |
| ʃ | S | **sh**eep | ɪə | 7 | N**EAR** |
| ʒ | Z | mea**s**ure | ɛə | 8 | SQU**ARE** |
| j | j | **y**ank | ʊə | 9 | C**URE** |
| x | x | lo**ch** | æ | c | t**i**mbre |

| | | | | | |
|---|---|---|---|---|---|
| h | h | **h**ad | ɑ̃: | q | dét**en**te |
| w | w | **w**et | æ̃: | 0 | l**in**gerie |
| ʧ | J | **ch**eap | ɒ̃: | ~ | bouill**on** |
| ʤ | _ | **j**eep | | | |
| ŋ | C | bac**on** | | | |
| m̩ | F | idealis**m** | | | |
| n̩ | H | burd**en** | | | |
| l̩ | P | dang**le** | | | |

6. Go to the *search* page.
7. Create a search matrix that's two words wide, and includes the *orthography* and *phonemes* layers.

Now we're going to do a search for the word "the" followed by a word that starts with schwa.

3. Type `the` in the first *orthography* box.
4. Click the second box on the *phonemes* layer, but don't enter anything in the box yet.
5. The box has a  button to the right of it.
   Hover the mouse over it to see what it says, and then click it.
   You will see that a section opens with a bunch of phoneme symbols on it.
6. Find the schwa symbol *ə* and click it.
   You will see that a *@* symbol appears in the box.*
   @* is the DISC symbol for *ə*, so in order to search for schwa, we have to use it in our search pattern.
7. We want words that *start with* schwa, so type `.*` after the `@` symbol.
8. Click *Search*.
   You will see that some of the words being matched are words that you might not normally think start with a schwa. LaBB-CAT is matching words against *all their possible phonemic transcriptions*, so if CELEX has multiple possible pronunciations for a word, and one of them starts with schwa, it will be matched.

With the phonemic transcriptions, we can do a better job of the search we tried in an earlier exercise - "the" followed by a word starting with a vowel…

9. Change your search so that, instead of just `@` at the beginning of the word, it matches any vowel.

> **Note**
>
> You could use the square-brackets `[]` at the start of your pattern, and type all vowel symbols inside them - Note that the vowels in the DISC representation extend beyond a, e, i, o, and u - you should add in all the vowels you see in the list that appears when you expand the Phoneme Symbol Selector, including all the diphthongs.
> **Alternatively**, you can simply click the *VOWEL* link in the Phoneme Symbol Selector, which will add all the DISC vowels for you, already enclosed in square-brackets.

10. Run the search and check that it's giving you what you expect. Notice that now there are no 'false positives' like "the one" that we were getting when searching by orthography alone.

Now that you've seen a few different layers, and how the search matrix works, you might want to try out some of the following searches, or invent some others:

- Words which have the vowel in DRESS as the second phoneme
- The word "the" followed by a word beginning with the phoneme /k/
- Words that begin with "k" in their spelling, but begin with the phoneme /n/
- Words that begin with "k" in their spelling, but *do not* begin with the phoneme /n/

---

In this worksheet you have seen that:

- The CELEX Layer Manager tags words with information from the CELEX lexicon.
- Phonemic transcription layers can be used to search on the basis of pronunciation.
- Although pronunciations can be displayed with IPA symbols, CELEX uses DISC to encode phonemes, so DISC must be used for searches.