Format download file *Leipzig Corpora Collection*

This file describes the format of the download corpus files of the Leipzig Corpora Collection. All files are encoded in UTF-8. Columns are separated by tabs.

Word list

The file contains the word list of all word forms of the corpus. Words are ordered by their frequency in descending order. The first 100 IDs of the word list are reserved for special characters.
Filename: *_words.txt
Format: Word_ID Word Frequency

Word list with POS and word stems (optional)

The file contains a list of word forms in the corpus with their POS tags (optionally with 'Universal POS tags' UD17[1]) and (optionally) stems. It is not available for all corpora.
Filename: *_words_pos_base.txt
Format: Word_ID Word POS POS_UD17 Baseform Frequency

Sentences list

The file contains all sentences of the corpus.
Filename: *_sentences.txt
Format: Sentence_ID Sentence

Sentences list with POS tags (optional)

The file contains sentences of the corpus with part of speech (POS) tags. It is not available for all corpora. Separation character between token and POS tag is the pipe character (e.g. 'car|NOUN'). The used tag set is language-dependent.
Filename: *_sentences_tagged.txt
Format: Sentence_ID Sentence

Sentences list with UD17 POS tags (optional)

The file contains sentences of the corpus with part of speech (POS) tags according to the "Universal POS Tags". It is not available for all corpora. Separation character between token and POS tag is the pipe character (e.g. 'car|NOUN').
Filename: *_sentences_tagged_ud17.txt
Format: Sentence_ID Sentence

Sources list

The file contains information about the used sources.
Filename: *_sources.txt
Format: Source_ID Source Date

Neighbourhood cooccurrences

The file contains information about how often two words occurred in direct neighbourhood in the the corpus and the significance of those cooccurrences based on log-likelihood. In the file, word1 occurrs immediately left of word2.

1 https://universaldependencies.org/u/pos/

Filename: *_co_n.txt
Format: Word1_ID Word2_ID Number_of_Cooccurrences Significance

Sentences cooccurrences

The file contains information about how often two words occurred in the same sentence and the significance of those cooccurrences based on log-likelihood.
Filename: *_co_s.txt
Format: Word1_ID Word2_ID Number_of_Cooccurrences Significance

Cooccurrences similarity

The file contains information on how similar two words are in terms of their sentence context. Both sentences co-occurrences and neighbourhood co-occurrences are taken into account. The similarity measure is based on the cosine similarity.
Filename: *_sim_w_co.txt
Format: Word1_ID Word_ID Cosine_Similarity

Inverted list

The file contains information about the occurrences of words in sentences (and optional their position in the sentence).
Filename: *_inv_w.txt
Format: Word_ID Sentence_ID (Position_in_Sentence)

Inverted source list

The file contains the mapping of sentences to the sources from which they were extracted.
Filename: *_inv_so.txt
Format: Source_ID Sentence_ID

Metadata

The file contains several metadata about the creation process of the corpus.
Filename: *_meta.txt
Format: Metadaten_ID Key Value

Import script

The import script can be used to import the files into a MySQL database.
Filename: *-import.sql
Example (Linux): $ mysql Database_Name < Database_Name-import.sql