Format Download „Deutscher Wortschatz“

Im Folgenden werden alle Angaben kurz erläutert, die Bestandteil einer Downloaddatei des Wortschatz-Projektes sind. Alle Dateien sind in UTF-8 kodiert. Spalten werden durch Tabulatoren getrennt.

Wortliste

Die Datei enthält die Wortliste aller im Korpus vorkommenden Wörter. Wörter sind absteigend sortiert nach ihrer Häufigkeit. Die ersten 100 Wort-IDs sind für Sonderzeichen reserviert.
Dateiname: *_words.txt
Format: Wort_ID Wort Häufigkeit

Wortliste mit POS und Grundformen (optional)

Die Datei enthält eine Liste im Korpus vorkommender Wörter mit ihren jeweiligen POS-Tags (optional zusätzlich POS-Tags nach „Universal POS Tags“ UD17[1]) und (optional) Grundformen. Sie liegt nicht für alle Korpora vor.
Dateiname: *_words_pos_base.txt
Format: Wort_ID Wort POS POS_UD17 Grundform Häufigkeit

Satzliste

Die Datei enthält alle Sätze des Korpus.
Dateiname: *_sentences.txt
Format: Satz_ID Satz

Satzliste mit POS (optional)

Die Datei enthält alle Sätze des Korpus mit POS-Tagging, liegt aber nicht für alle Korpora vor. Die Trennung zwischen Token und POS-Tag erfolgt über das Pipesymbol (z.B. ‚car|NOUN‘). Das verwendete Tagset ist sprachabhängig.
Dateiname: *_sentences_tagged.txt
Format: Satz_ID Satz

Satzliste mit POS nach UD17 (optional)

Die Datei enthält alle Sätze des Korpus mit POS-Tagging entsprechend der „Universal POS Tags“, liegt aber nicht für alle Korpora vor. Die Trennung zwischen Token und POS-Tag erfolgt über das Pipesymbol (z.B. ‚car|NOUN‘).
Dateiname: *_sentences_tagged_ud17.txt
Format: Satz_ID Satz

Quellenliste

Die Datei enthält Angaben zu den verwendeten Quellen.
Dateiname: *_sources.txt
Format: Quellen_ID Quelle Datum

Nachbarschaftskookkurrenzen

Die Datei enthält Informationen wie oft zwei Wörter in unmittelbarer Nachbarschaft im Korpus vorkommen und die Signifikanz dieser Auftreten auf Basis von Log-Likelihood. Wort1 steht dabei

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

unmittelbar links von Wort2.
Dateiname: *_co_n.txt
Format: Wort1_ID Wort2_ID Anzahl_Vorkommen Signifikanz

Satzkookkurrenzen

Die Datei enthält Informationen wie oft zwei Wörter im gleichen Satz im Korpus vorkommen und die Signifikanz dieser Auftreten auf Basis von Log-Likelihood.
Dateiname: *_co_s.txt
Format: Wort1_ID Wort2_ID Anzahl_Vorkommen Signifikanz

Kookkurrenz-Ähnlichkeit

Die Datei enthält Informationen wie ähnlich sich zwei Wörter bezüglich ihres Satz-Kontextes sind. Dabei werden sowohl Satzkookkurrenzen als auch Nachbarschaftskookkurrenzen berücksichtigt. Das verwendete Ähnlichkeitsmaß basiert auf der Kosinus-Ähnlichkeit.
Dateiname: *_sim_w_co.txt
Format: Wort1_ID Wort2_ID Kosinus_Ähnlichkeit

Inverse Liste

Die Datei enthält die Zuordnung eines Wortes zu den Sätzen in welchem es (und optional an welcher Position in diesem) vorkommt.
Dateiname: *_inv_w.txt
Format: Wort_ID Satz_ID (Position_im_Satz)

Inverse Quellenliste

Die Datei enthält die Zuordnung der Sätze zu den Quellen aus denen sie extrahiert wurden.
Dateiname: *_inv_so.txt
Format: Quellen_ID Satz_ID

Allgemeine Metadaten

Die Datei enthält verschiedene Metadaten zum Erstellungsprozess des Korpus.
Dateiname: *_meta.txt
Format: Metadaten_ID Key Value

Importskript

Das Importskript kann zum Import der Dateien in eine MySQL-Datenbank genutzt werden.
Dateiname: *-import.sql
Beispielaufruf (Linux): $ mysql Datenbank_Name < Datenbank_Name-import.sql