Corpora from the Web (COW)

Sammlung extrem großer Webkorpora zu diversen romanischen Sprachen

Sprache	Niederländisch, Englisch, Französisch, Deutsch, Spanisch, Schwedisch
Sprachliche Realisierung	schriftlich
Umfang	Gigatoken-Webkorpora, die zum Teil 10-20 Milliarden Tokens umfassen
Medium	Sammlung von Texten aus dem Internet
Geographischer Ursprung	Europa
Zeitliche Einordnung	ab 2011
Form der Daten	gesammelte Texte aus dem Internet, online durchsuchbar mithilfe der Suchmaske Colibri²
Format	HTML, tab-separated files (TSV)
Annotation	tokenisiert, zum Teil lemmatisiert und part-of-speech-annotiert
Mögliche Suchabfragen	Suche nach Wörtern, Wortfolgen, Lemmata und Wortarten; die Ergebnisse sind exportierbar
Quelle /Herausgeber	Felix Bildhauer, Roland Schäfer, Freie Universität Berlin
Nutzungsvoraussetzungen	kostenlose Registrierung erforderlich
Link	http://corporafromtheweb.org/
Literatur	Schäfer, Roland (2015): "Processing and querying large web corpora with the COW14 architecture". In: Proceedings of Challenges in the Management of Large Corpora (CMLC-3). Download Weitere Literaturtips auf der Hompage