Web as Corpus (WaCky)

Sammlung sehr großer Korpora aus Internettexten, auch für Französisch (1,6 Mrd. Wörter) und Italienisch (2 Mrd. Wörter).

Sprache	Englisch, Französisch, Deutsch, Italienisch
Sprachliche Realisierung	schriftlich
Umfang	bis zu 2 Milliarden Wörter pro Korpus
Medium	Acht Internet-Korpora in vier Sprachen, darunter ukWac, frWac, deWac und itWac. Diese Korpora wurden anhand von Stichwortlisten aus den Domains der jeweiligen Sprache extrahiert. Verfügbar sind auch annotierte Versionen der französischen und englischen Wikipedia.
Zeitliche Einordnung	aktuell
Form der Daten	große Textmengen aus dem Internet, online durchsuchbar. Download auf Anfrage möglich.
Format	HTML
Annotation	zum Teil Lemmatisierung und part-of-speech-Annotation, zum Teil syntaktisches Parsing
Mögliche Suchabfragen	Wörter, Sätze, Lemmata, Wortarten, reguläre Ausdrücke. Die Suchergebnisse werden mit Konkordanzen ausgegeben.
Quelle /Herausgeber	Universitäten Bologna, Pisa, Trento, Stuttgart, Darmstadt, Hildesheim, Naval, Oslo, Pecara, Leeds und Tokio
Nutzungsvoraussetzungen	Zugang frei
Link	Einführung: http://wacky.sslmit.unibo.it/doku.php Korpus-Übersicht: http://wacky.sslmit.unibo.it/doku.php Suchmaske: http://nl.ijs.si/noske/wacs.cgi/first_form?corpname=itwac;align=
Literatur	M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta (2009): "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora". In: *Language Resources and Evaluation* 43 (3), 209-226. Download