Wikicorpus

Dreisprachiges (Spanisch, Katalanisch, Englisch), lemmatisiertes und morphosyntaktisch annotiertes Korpus, bestehend aus einem Großteil der 2006 verfügbaren Wikipedia-Inhalte.

Sprache	Spanisch, Englisch, Katalanisch
Varietät	Standard
Sprachliche Realisierung	schriftlich
Umfang	ca. 750 Mio. Wörter
Medium	Wikipedia-Inhalte in drei Sprachen, sprachanalytisch aufbereitet
Zeitliche Einordnung	2006
Form der Daten	anntotierte Wikipedia-Texte, zum Download verfügbar
Format	XML
Annotation	lemmatisiert, part-of-speech-annotiert, semantisch annotiert
Quelle /Herausgeber	Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, German Rigau
Nutzungsvoraussetzungen	Zugang frei
Link	http://www.cs.upc.edu/~nlp/wikicorpus/
Literatur	Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, German Rigau (2010): "Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus". In: Proceedings of 7th Language Resources and Evaluation Conference (LREC'10). La Valleta, Malta. Download