ItWac

Das "Italian Web-As-Corpus" besteht aus online verfügbaren Texten, die durch *web crawling* gesammelt wurden.

Sprache	Italienisch
Sprachstufe	Standard
Sprachliche Realisierung	schriftlich
Umfang	ca. 1.5–2 Milliarden Tokens
Medium	Texte von Webseiten, die ausschließlich die Domain .it haben
Geographischer Ursprung	Italien
Form der Daten	Online-Texte
Format	Text, XML
Annotation	lemmatisiert, POS-Tags (automatisch annotiert), das Subkorpus von italienischem Wikipedia wurde zusätzlich mit Semantik und Syntax annotiert
Mögliche Suchabfragen	Mit Sketch Engine oder NoSketchEngine können Wortfrequenzen, n-grams, Konkordanzen usw. erstellt werden
Quelle /Herausgeber	Università di Bologna
Nutzungsvoraussetzungen	Anmeldung erforderlich für Sketch Engine
Link	https://corpora.dipintra.it/ (NoSketchEngine) oder https://www.sketchengine.eu/itwac-italian-corpus/
Zum Zitieren:	Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3). 209–226.