PAISÀ: Corpus italiano

Im Projekt PAISÀ (Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati) wurden Texte aus dem Internet zusammengestellt und linguistisch dokumentiert, um authentische Texte für den Sprachunterricht zur Verfügung zu stellen. Das Korpus ist lemmatisiert und POS-getaggt und umfasst 250 Mio Tokens. Korpus und Frequenzlisten können im Volltext heruntergeladen werden.

Sprache	Italienisch
Varietät	Standard
Sprachliche Realisierung	schriftlich (Internet)
Umfang	ca. 250 Mio. Wörter
Medium	Texte mit einer Länge von mehr als 150 Wörtern, insgesamt ca. 380.000 Dokumente aus mehr als 1.000 Webseiten. 260.000 Dokumente aus Wikipedia, ca. 65.000 aus Blogs.
Geographischer Ursprung	[Italien]
Form der Daten	XML-Dokument, Frequenzlisten, Dokumentation zu Lemmata und POS
Annotation	lemmatisiert, part-of-speech-annotiert
Quelle /Herausgeber	PAISÀ-Projekt
Link	http://www.corpusitaliano.it/
Zum Zitieren:	Lyding, V. / Stemle, E. / Borghetti, C. / Brunello, M. / Castagnoli, S. / Dell'Orletta, F. / Dittmann, H. / Lenci, A. / Pirrelli, V. 2014. The PAISÀ Corpus of Italian Web Texts. *Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, April 2014.* 36-43.