itTenTen16

Das Korpus besteht aus online Texten (knapp 5 Mrd. Wörter), die durch *web crawling* gesammelt wurden. Die Texte wurden dann mit automatisierten Tools aufbereitet. Die ‘TenTen’ Korpora umfassen über 30 Sprachen.

Sprache	Italienisch
Sprachstufe	Standard
Sprachliche Realisierung	schriftlich
Umfang	ca. 5 Mrd. Wörter
Medium	Online-Texte
Geographischer Ursprung	Italien [Online]
Form der Daten	Texte aus dem Internet, die aufbereitet werden, indem unnötige Informationen gelöscht werden (URLs, Duplikate)
Format	online durchsuchbar
Annotation	lemmatiziert, POS-Tags
Mögliche Suchabfragen	SketchEngine hat mehrere Tools, die unterschiedliche Suchabfragen ermöglichen (Konkordanzen, Synonyme, n-grams usw.)
Quelle/Herausgeber	Jakubíek, M., Kilgarriff, A., Ková, V., Rychlý, P., & Suchomel, V., Masaryk University / Lexical Computing
Nutzungsvoraussetzungen	Anmeldung über SSO Universität Potsdam
Link	https://www.sketchengine.eu/ittenten-italian-corpus/
Zum Zitieren:	Jakubíek, M., Kilgarriff, A., Ková, V., Rychlý, P., & Suchomel, V. 2013. The TenTen corpus family. *7th International Corpus Linguistics Conference CL,* 125–127.