

PERCEO : un Projet d'Etiqueteur Robuste pour l'Ecrit et pour l'Oral

Sammlung von Texten und Transkriptionen verschiedener Korpora, die mithilfe von *TreeTagger* POS-annotiert und Lemmatisiert wurden.

Sprache	Französisch
Sprachliche Realisierung	schriftlich und mündlich
Umfang	gesprochene Sprache: ca. 102.000 Tokens geschriebene Sprache: ?
Medium	Lemmatisiertes und morphosyntaktisch annotiertes Korpus, bestehend aus den Teilkorpora <i>L'Est Républicain</i> , <i>TCOF</i> und <i>Frantext</i> . Das verwendete Annotationsschema ist ebenfalls verfügbar und kann für das Tagging ähnlicher Korpora verwendet werden.
Geographischer Ursprung	Frankreich
Zeitliche Einordnung	Teilkorpus <i>L'Est Républicain</i> : 1999, 2002, 2003 Teilkorpus <i>Frantext</i> : 18.-20. Jh. Teilkorpus <i>TCOF</i> : 1980er und 1990er-Jahre
Form der Daten	PERCEO Oral ist im Textformat mit dem dazugehörigen Annotationsschema und der <i>TreeTagger</i> -Datei herunterladbar
Format	txt, par
Annotation	mehrfach korrigierte Lemmatisierung und POS-Annotation mit <i>TreeTagger</i>
Mögliche Suchabfragen	Lemma, Wortart
Quelle/Herausgeber	ATILF (CNRS - Université de Lorraine) & INIST (CNRS) - LIPN / CNRTL
Nutzungsvoraussetzungen	Zugang frei
Link	http://cnrtl.fr/corpus/perceo/