

FRENCH TREEBANK

Corpus of French lexical and syntactic resources richly annotated for linguists, Natural Language Processing (NLP) - ready. Corpus for French made up of more than 20,000 sentences extracted from the newspaper *Le Monde* (1990-1993), richly annotated and validated manually

ERG\NEO

L'AVENIR EST FAIT D'AUDACE

PRESENTATION

Analyzing and reproducing natural language requires an understanding of the meaning of the sentence. To meet this need, the corpus made up of more than 20,000 richly annotated sentences in French constitutes a lexical and syntactic resource of reference for linguists and computer scientists, in particular in the case of use in automatic natural language processing.



© Inés Estébanez

APPLICATIONS

- Automatic natural language processing
- Semantic web, search engine
- Human-machine dialogue (chatbots)
- Spellchecking
- Automatic translation
- Language teaching

CONTACT

- ☎ +33 (0)1 44 23 21 50
- ✉ industriels@erganeo.com
- Ref. project : 519

Lexical ressources - Syntatic ressources - NLP

COMPETITIVE ADVANTAGES

- Quality of the corpus: annotation by automatic tools and corrected by hand by several successive passages on the different annotations
- Available in four formats: xml (original format), Tiger-xml (the most complete format with compound components), PTB (constituent annotations), CoNLL (dependencies annotations)
- Rich annotation : domain, author, date; compound words (and components), 218 morpho-syntactic labels, grammatical functions and trees of syntactic constituents

INTELLECTUAL PROPERTY

Corpus filing on 01/28/2018 with APP, n°IDDDN FR 001 050008 000 D C 2008 000 10300

PUBLICATIONS

[Abeillé, A., L. Clément, F. Toussenet. 2003. «Building a treebank for French», in A. Abeillé \(ed\) Treebanks, Kluwer, Dordrecht. \(p. 165-187\)](#)
[A.Abeillé, Clément L., Liégeois L. \(2019\), Un corpus annoté pour le français: le French Treebank, TAL, 60 :2](#)

DEVELOPMENT PHASE

- ☑ Commercial version and documentation in French available.