

SELEXINI – un grand corpus français, divers et parsé automatiquement

Manon Scholivet¹ Agata Savary¹ Louis Estève¹

Marie Candito² Carlos Ramisch³

(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France,

(2) Université Paris Cité, CNRS, LLF, 75013 Paris, France,

(3) Aix Marseille Univ, CNRS, LIS, 13009 Marseille, France

(1) nom@lisn.fr, (2) prenom.nom@u-paris.fr, (3) prenom.nom@lis-lab.fr

RÉSUMÉ

L'annotation de grands corpus de texte est essentielle pour de nombreuses tâches de Traitement Automatique des Langues. Dans cet article, nous présentons SELEXINI, un grand corpus français annoté automatiquement en syntaxe. Ce corpus est composé de deux parties : la partie BigScience, et la partie HPLT. Les documents de la partie HPLT ont été sélectionnés dans le but de maximiser la diversité lexicale du corpus total SELEXINI. Une analyse de l'impact de cette sélection sur la diversité syntaxique a été réalisée, ainsi qu'une étude de la qualité des nouveaux mots issus de la partie HPLT du corpus SELEXINI. Nous avons pu montrer que malgré l'introduction de nouveaux mots considérés comme intéressants (formes de conjugaison rares, néologismes, mots rares,...), les textes issus de HPLT sont extrêmement bruités. De plus, l'augmentation de la diversité lexicale n'a pas permis d'augmenter la diversité syntaxique.

ABSTRACT

SELEXINI – a large and diverse automatically parsed corpus of French

The annotation of large text corpora is essential for many tasks. We present here a large automatically annotated corpus for French. This corpus is divided into two parts : the first from BigScience, and the second from HPLT. The annotated documents from HPLT were selected in order to optimise the lexical diversity of the final corpus SELEXINI. An analysis of the impact of this selection was carried out on syntactic diversity, as well as on the quality of the new words resulting from the HPLT part of SELEXINI. We have shown that despite the introduction of interesting new words, the texts extracted from HPLT are very noisy. Furthermore, increasing lexical diversity did not increase syntactic diversity.

MOTS-CLÉS : Corpus arboré, Diversité, Syntaxe, Parsing Automatique, Échantillonnage.

KEYWORDS: Treebank, Diversity, Syntax, Automatic Parsing, Sampling.

ARTICLE : **Accepté à** 18th Workshop on Building and Using Comparable Corpora (BUCC)

<https://aclanthology.org/2025.bucc-1.10/>.
