Un corpus de référence pour le français

Une ressource lexicale et syntaxique richement annotée (et validée manuellement) pour les linguistes, utilisable en TAL.

- Projet initié en 1997, avec le soutien de l'IUF, du CNRS et du CNRTL
- 21 550 phrases (environ 664 500 tokens) du journal Le Monde (1990-1993)
- métadonnées : auteur, date, domaine (par article)
- Annotations lexicales (catégories, sous-catégories, flexion, mots composés avec composants) et syntaxiques (constituants majeurs, fonctions grammaticales) validées
- Corpus annoté téléchargeable (version 1.0 2016) en plusieurs formats (xml, Tiger-xml, PTB, ConNL)

Le corpus arboré est diffusé gratuitement à des fins de recherche, sous réserve de la signature des conditions d'utilisation

Nous contacter pour obtenir une licence permettant une utilisation commerciale et le développement de produits dérivés

Citation : Abeillé, A., L. Clément, and F. Toussenel. 2003. "Building a treebank for French", in A. Abeillé (ed) Treebanks, Kluwer, Dordrecht. (p.165-187)

Le corpus a été annoté par des outils automatiques dédiés (Clément 2001) et corrigé à la main par plusieurs passages successifs sur les différentes annotations (mots composés, catégories lexicales, flexion, constituants majeurs, fonctions syntaxiques...)

Il est toujours possible que des erreurs subsistent. Si vous repérez une erreur potentielle, merci de vérifier dans les guides qu'il ne s'agit pas d'un choix d'annotation ; sinon, merci de nous signaler.

Exemples d'annotation syntaxique

Sélectionnez une phrase

- Une quinzaine de militaires libériens ont été transférés à Abidjan.
- Aussi s'est-elle évertuée à torpiller tous les projets en faveur de Rhône-Rhin.
- La diminution paraît, toutefois, moins nette en France et en Italie.

Sélectionnez le format de sortie

Texte | XML | PTB | Tiger | CoNNL

```
Une quinzaine de militaires libériens ont été transférés à Abidjan.
```