SOFA : Une plateforme d'analyse syntaxique en ligne pour l'ancien français

Gaël Guibon
(1) LaTTiCe CNRS, 92120 MONTROUGE, France
gael.guibon@gmail.com

RÉSUMÉ

SOFA une application web dédiée à l'étiquetage syntaxique de l'ancien français. Cette plateforme est une démonstration permettant d'appliquer sur n'importe quel texte, ou sur un des textes d'ancien français, des modèles de lemmatisation, d'annotation morpho-syntaxique, et d'analyse syntaxique, en plus d'en visualiser les performances.

ABSTRACT

SOFA : An online Syntactic Old French Annotator

SOFA is a web application for Old French dependency parsing. It provides lemmatization, part-of-speech tagging, and dependency parsing specialized for Old French. These annotations can be applied on a set of medieval texts, or any other text from the user. Performance are also displayed.

MOTS-CLÉS : analyse syntaxique, ancien français, corpus arboré, étiquetage morpho-syntaxique, apprentissage automatique.

KEYWORDS: Dependency Parsing, Old French, treebank, POS labelling, machine learning.

1 Introduction

Les corpus de langues anciennes constitués à des fins d'annotation automatique sont de plus en plus exploités (Celano & Crane, 2015). Mais ils ne bénéficient pas d'une grande visibilité. C'est pourquoi nous avons développé une application web dédiée à l'analyse syntaxique de l'ancien français. Cette application a été développée dans le cadre de travaux visant à l'exploitation d'un corpus arboré d'ancien français, et a pour but de permettre à une communauté plus large d'y accéder. L'utilisateur peut en effet étiqueter syntaxiquement et morpho-syntaxiquement des textes d'ancien français, mais également appliquer cet étiqueteur sur n'importe quel autre texte, qu'il soit d'ancien français ou non.

2 Corpus arboré d'ancien français

Le corpus utilisé est le *Syntactic Reference Corpus of Medieval French* [1] (SRCMF) (Stein & Prévost, 2013), le premier corpus arboré d'ancien français. Issu d'un projet ANR-DFG dirigé par Achim

1. http://srcmf.org/

Stein [2] (ILR, U. Stuttgart) et Sophie Prévost [3](Lattice, CNRS/ENS/Paris3), ce corpus apporte une annotation syntaxique sur des textes de la Base de Français Médiéval (Guillot *et al.*, 2007) (BFM [4]) et le Nouveau Corpus d'Amsterdam (Stein & others, 2006) (NCA [5]).

Les dix textes du SRCMF sélectionnés pour l'apprentissage varient en date (10ème au 13ème siècle), en forme (vers ou prose), en domaine (religieux, littéraire, didactique, historique), et en dialecte (normand, picard, champenois, anglo-normand).

3 Architecture de l'application

Composantes techniques. L'application est développée en JavaEE et utilise les JavaServer Faces (JSF) à l'aide du *framework* Primefaces [6]. Son déploiement requiert un serveur Tomcat et une version pré-compilée ainsi que le code source sont disponibles sur github : `https://github.com/gguibon/sofa/`.

Annotation. L'annotation du texte sélectionné, ou du fichier téléversé, permet de prédire trois types d'annotations pour chaque mot de chaque phrase. Chaque annotation a été optimisée pour l'ancien français lors de précédents travaux (Guibon *et al.*, 2015). Elles suivent la chaîne de traitement suivante : 1) Lemmatisation avec *TreeTagger* [7] (Stein, 2014) ; 2) Etiquetage morpho-syntaxique à base de *Conditional Random Fields* (Lafferty *et al.*, 2001) ; 3) Analyse syntaxique en dépendance avec *Mate* (Bohnet, 2010).

Utilisation. L'application permet deux types d'utilisation. La première est la sélection d'un modèle parmi ceux du SRCMF pour s'en servir afin d'annoter en syntaxe et en morpho-syntaxe un texte du SRCMF. La seconde est le chargement d'un fichier au format TSV (tabulated separated format) qu'il sera alors possible d'annoter à l'aide d'un des modèles du SRCMF. Le format d'entrée TSV est en fait une modification du format CoNLL [8] qui se justifie par l'ajout d'informations supplémentaires dans le corpus initial. A la fin de l'annotation la vue se met à jour et présente les résultats.

L'application elle-même est hébergée à l'adresse suivante : `http://sofa.ilpga.fr/`.

4 Conclusion

Afin de mettre en avant l'étiquetage automatique de l'ancien français et de le rendre plus accessible, nous avons présenté une plateforme en ligne qui permet d'appliquer des modèles d'analyse syntaxique et morpho-syntaxique appris sur la toute dernière version du SRCMF. Ces modèles pouvant ainsi être plus facilement appliqués sur des textes du SRCMF ou sur un nouveau texte d'ancien français dont l'utilisateur aurait la possession, qu'il soit déjà annoté ou non.

2. `http://www.uni-stuttgart.de/lingrom/stein/`
3. `http://www.lattice.cnrs.fr/Sophie-Prevost,229`
4. `http://bfm.ens-lyon.fr/`
5. http ://www.uni-stuttgart.de/lingrom/stein/corpus/
6. `http://primefaces.org/`
7. `http://bfm.ens-lyon.fr/article.php3?id_article=324`
8. `http://ilk.uvt.nl/conll/#dataformat`

Références

BOHNET B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *The 23rd International Conference on Computational Linguistics (COLING 2010)*.

CELANO G. G. & CRANE G. (2015). Semantic role annotation in the ancient greek dependency treebank. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, p.2̃6.

GUIBON G., TELLIER I., PRÉVOST S., CONSTANT M. & GERDES K. (2015). Searching for discriminative metadata of heterogenous corpora. In *Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, p. 72–82.

GUILLOT C., LAVRENTIEV A. & MARCHELLO-NIZIA C. (2007). La base de français médiéval (BFM) : états et perspectives. p. 143–152.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, p. 282–289.

STEIN A. (2014). Parsing heterogeneous corpora with a rich dependency grammar. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* : European Language Resources Association.

STEIN A. & OTHERS (2006). *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gle\s sgen*. Stuttgart : Institut für Linguistik/Romanistik.

STEIN A. & PRÉVOST S. (2013). Syntactic annotation of medieval texts : the syntactic reference corpus of medieval french (SRCMF). In T. NARR, Ed., *New Methods in Historical Corpus Linguistics*. Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3, P. Bennett, M. Durrell, S. Scheible and R. Whitt (eds).