Comportement de TreeTagger sur des données du 18ème

Francois Huang, Blanche Miret, Preethi Srinivasan, Dao Thauvin
RELIA Recherche En Licence Informatique et études Anglophones
Encadrants : Jean-Baptiste Yunès et Nicolas Ballier
Université de Paris, 5 rue Thomas Mann, 75013, Paris
blanche.miret@etu.univ-paris-diderot.fr,
francois.huang@etu.univ-paris-diderot.fr,preethi.lfp@gmail.com,
dao.thauvin@etu.univ-paris-diderot.fr

RÉSUMÉ

Peut-on utiliser un outil d'étiquetage morpho-syntaxique pour mesurer l'évolution d'une langue à travers les siècles, et notamment reconnaître les mots devenus obsolètes ? Dans quelle mesure cet outil, fondé sur l'apprentissage machine, arrive-t-il à s'adapter à une version plus ancienne du langage qu'il a été entrainé à reconnaître ? C'est pour répondre à ces interrogations que nous avons appliqué TreeTagger, exercé à identifier et catégoriser les mots de l'anglais moderne, sur le *Critical Pronouncing Dictionary* de John Walker datant de 1791. Les résultats nous permettent par exemple de retrouver la différence d'évolution attendue entre les différentes catégories grammaticales de la langue : les prépositions étant sujettes à peu de transformations, la reconnaissance de celles du 18e siècle ne pose pas de problème ; celle des noms communs ou adjectifs est moins évidente. Quant à la détection de l'obsolescence des mots, la majorité de ceux se voyant attribuer "unknown" comme lemme dans le résultat ne sont effectivement plus utilisés aujourd'hui. TreeTagger semble alors être une piste d'outil dans la mesure d'évolution d'un langage.

ABSTRACT

TreeTagger's behaviour towards 18th century's data

Can a morpho-syntactic labelling tool be used to measure the evolution of a language through the centuries, including the recognition of words that have become obsolete ? To what extent can this tool, based on machine learning, adapt to an older version of the language it has been trained to recognize ? It is to answer these questions that we applied TreeTagger, which is used to identify and categorize words in modern English, to John Walker's *Critical Pronouncing Dictionary* from 1791. The results allow us, for example, to find the expected difference in evolution between the different grammatical categories of the language : prepositions being subject to few transformations, the recognition of those of the 18th century does not pose a problem ; that of common nouns or adjectives is less obvious. As for the detection of word obsolescence, the majority of words that are attributed "unknown" as a lemma in the result are effectively no longer used today. TreeTagger then seems to be a potential tool in the measurement of the evolution of a language.

MOTS-CLÉS : TreeTagger, catégorie grammaticale, obsolescence, évolution, prédiction , 18ème siècle.

KEYWORDS: TreeTagger, Part-of-Speech tag, obsolescence, evolution, prediction, 18th century.