6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

TreeTagger entraîné avec des données modernes face au *Critical Pronouncing Dictionary* de J. Walker

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

Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement supporté par le programme IdEx Université de Paris ANR-18-IDEX-0001

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 arrive-t-il à s'adapter à un état de langue plus ancien que celui avec lequel il a été entraîné ? 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.

Le résultat de l'expérience permet 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. Le taux de précision mesuré sur un échantillon de 200 mots est de 93,5%, résultat inférieur aux 96,36% officiels de TreeTagger (Schmid, 1994) sur des données de la même époque que celles sur lesquelles l'outil a été entraîné. Le taux de rappel, c'est à dire la capacité de reconnaissance d'une certaine catégorie grammaticale, est de 98% pour les prépositions, 91% pour les noms, 90% pour les adjectifs.

L'approche pour la détection des mots obsolètes s'est faite en observant les termes se voyant attribuer "unknown" comme lemme, ce qui fut le cas pour 9,2% de l'ensemble des tokens, plus précisément 35 748 sur 386 172 au total. Cependant, le jeu de données étant un dictionnaire, l'analyse du résultat peut être départagée entre celle des mots vedettes hors contexte d'une part et l'ensemble des définitions d'autre part. 89,0% des mots identifiés comme potentiellement obsolètes appartiennent à l'ensemble des mots vedettes et sur un échantillon de 200 tokens, seulement 38,5% se sont révélés l'être réellement. En revanche, le taux de réussite sur les mots contenus dans les définitions est bien plus prometteur : 78,0% sont en effet obsolètes, avec une précision de marquage morpho-syntaxique de 66,7%, expression de la capacité de TreeTagger à s'adapter à un vocabulaire désuet dans le cas de tokens contextualisés. Au total, 1,2% des mots en contexte du jeu de données complet a été identifié comme éventuellement obsolète. Une suite de la recherche pourrait conduire à élargir les échantillons étudiés et mesurer le pourcentage de mots reconnus comme obsolètes parmi un jeu de token certifié comme tel.

Finalement, avec une précision relativement élevée dans l'utilisation des lemmes marqués 'unknown'

pour identifier l'obsolescence des mots, TreeTagger semble être un outil pertinent d'évolution du langage entre deux périodes.

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

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

REMERCIEMENTS

Nous remercions en premier lieu Nicolas Ballier et Jean-Baptiste Yunès qui nous ont accompagnés tout au long de ce projet et introduits aux règles de la recherche académique. Nous remercions également Nicolas Trapateau pour son travail effectué sur le *Critical Pronouncing Dictionary* de John Walker que nous avons utilisé comme jeu de données dans nos recherches, ainsi que les organisateurs de la conférence des Apprenti-e-s Chercheur-euse-s 2020 de nous donner la possibilité de partager ce travail. Enfin, merci au programme IdEx de l'Université de Paris grâce à qui ce projet fut financé.

Références

HUANG F., MIRET B., SRINIVASAN P. & THAUVIN D. (2020). Github repository. https://github.com/BlancheMiret/TreeTagger_on_Walker.

KHURANA D., KOLI A., KHATTER K. & SINGH S. (2017). *Natural Language Processing : State of The Art, Current Trends and Challenges*. Thèse de doctorat, Manav Rachma International University.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, p. 44–49.

TRAPATEAU N. (2015). *Placement de l'accent et voyelles inaccentuées dans la prononciation de l'anglais du XVIIIe siècle sur la base du témoignage des dictionnaires de prononciation, des vers et de la musique vocale*. Thèse de doctorat, Université de Poitiers.

WALKER J. (1791). *A Critical Pronouncing Dictionary*. British Library.