Extraction dans des textes anciens d'entités nommées de type binômes de la classification linnéenne du vivant : une étude de cas

Clément Morand*, Olivier Ridoux**

* École Normale Supérieure de Rennes
clement.morand@ens-rennes.fr,
** Université de Rennes - IRISA
olivier.ridoux@irisa.fr

Résumé. Les binômes linnéens, ou taxons, sont un type d'entités nommées rarement étudié, et pas du tout dans le cadre de l'enrichissement d'archives anciennes. Nous introduisons *l'hypothèse du lecteur compétent* qui sait reconnaître un taxon, même obsolète ou mal composé. Cette hypothèse est la base des évaluations présentées. Nous comparons plusieurs approches pour la reconnaissance des taxons : dictionnaires, règles, et une forme d'apprentissage par généralisation. Nous montrons que ressembler à du latin est un critère trop peu précis. Enfin, nous montrons que combiné à un critère de rareté, le critère du latin permet une reconnaissance de bonne qualité : une f-mesure d'environ 70 %.

1 Introduction

La revue La Nature (Tissandier, 1873; Vautrin, 2018) est une revue de vulgarisation scientifique et technique qui a été publiée de 1873 à 1960 (on abrégera son nom en *LN*). Son contenu *prima facie* est obsolète, mais l'étude de ce que cette archive révèle de presque un siècle d'évolution de la société relève de ce qu'on appelle les humanités numériques (Burdick et al., 2012) et intéresse les historiens, sociologues, etc., mais aussi le citoyen curieux.

La revue LN avait une publication hebdomadaire. Tous les semestres, les numéros étaient rassemblés dans des volumes qui étaient publiés séparément. Il y a 155 volumes disponibles, d'environ 500 pages chacun, pour un total de plus de 80 000 pages. Chaque volume contient une centaine d'articles avec une très grande variabilité de contenus et de styles, plus des notes, comptes-rendus, etc. Ces volumes sont composés selon des règles typographiques qui varient avec le temps. Ce sont ces volumes qui ont été numérisés par le Conservatoire numérique du CNAM (`http://cnum.cnam.fr/CGI/redira.cgi?4KY28`) et qui sont disponibles sous forme de scans de très faible résolution.

Mettre à la disposition du public ces archives demande d'en baliser la structure et les contenus. La reconnaissance d'entités nommées (NER, pour *Named Entities Recognition* (Jurafsky et Martin, 2009; Ehrmann et al., 2021; Nadeau et Sekine, 2007; Nasar et al., 2021)) peut être utilisée pour cela car les entités reconnues donnent une idée des sujets traités, et leur répartition dans le temps donne une idée de l'évolution des sujets. Les principaux types d'entités étudiés