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

dans la littérature sont les entités géographiques, de personnes et d'institutions, mais par son thème général de vulgarisation scientifique et technique, d'autres types sont intéressants pour instrumenter les archives de LN : ex. les noms d'espèces chimiques et les noms d'espèces biologiques (appelés aussi *taxons*). C'est à ces dernières que nous nous intéressons ici.

Les entités nommées de type taxon ont été étudiées pour des recherches en biodiversité (Koning et al., 2005; Sautter et al., 2006; Little, 2020), en biomédecine (Akella et al., 2012; Pafilis et al., 2013), ou en microbiologie (Nédellec et al., 2006), mais pas pour des corpus historiques aussi multi-thématiques que celui de LN. Au contraire, beaucoup des travaux de ce domaine ont une finalité extrêmement précise, comme celle d'étudier la biodiversité des Philippines (Nguyen et al., 2019) ou les plantes médicinales du Maghreb (Seideh et al., 2016), et définissent un réseau de tâches très spécifique, comme reconnaître un taxon et la localisation géographique ou temporelle de la découverte d'un spécimen, ou reconnaître un taxon et son rôle thérapeutique.

Notre objectif est de reconnaître dans un corpus qui n'est que marginalement à contenu biologique des dénominations d'espèces biologiques, même si celles-ci sont obsolètes ou mal composées. Nous formulons à ce sujet *l'hypothèse du lecteur compétent* qui sait reconnaître cette intention dans un texte. La connexion avec d'autres analyses sera faite ailleurs, un peu dans le style de Voyant Tools (Rockwell et Sinclair, 2016).

Les méthodes d'apprentissage neuronal sont de plus en plus utilisées pour les tâches NER, mais ne nous semblent pas adaptées à notre objectif. Au delà de LN, nous souhaitons proposer à un public qui n'est ni informaticien ni très équipé, des outils sobres et faciles à expliquer qui lui permettraient d'incorporer ses propres corpus. De plus, les taxons linéens obéissent à des règles précises et explicites données par les codes de nomenclature. Enfin, de par sa grande variété de sujets, les taxons sont rares dans LN, mais significatifs là où ils sont présents. Par conséquent un grand nombre de pages annotées ne correspond qu'à un faible nombre d'occurences positives. Pour toutes ces raisons, il nous semble plus intéressant d'explorer d'autres méthodes étudiées pour réaliser des tâches NER, comme utiliser un dictionnaire, des règles, apprises ou données a priori ; et montrer que certaines répondent à nos objectifs.

La suite de l'article compare ces différentes méthodes : utiliser un dictionnaire (Section 3), apprendre le langage des entités recherchées en généralisant une base d'exemples (Section 4), et apprendre le langage des entités recherchées en codant les règles de formation de leurs noms (Section 5). Dans tous les cas, les résultats obtenus sont médiocres. La Section 6 introduit un nouveau critère, la rareté des noms recherchés relativement au vocabulaire courant. Mais avant cela, la Section 2 présente la stratégie d'évaluation utilisée dans les Sections 3 à 6.

2 Stratégie d'évaluation

Afin d'évaluer les différentes approches étudiées nous avons annoté à la main quatre volumes, suffisamment écartés dans le temps pour représenter des situations variées. Ces volumes sont le 12 (1er semestre de 1879), le 83 (2nd semestre de 1912), le 126 (1er semestre de 1934), et le 155 (année 1960, dernière année de publication). L'annotation consiste à identifier les articles, et pour chacun d'eux à identifier les occurrences de taxons linnéens. Chaque occurrence est représentée par sa chaîne source et sa position dans l'article.

C'est dans l'annotation que réside l'hypothèse du lecteur compétent. Celle-ci conduit à accepter comme positives des chaînes de caractères fautives : ex. *Wus* pour *Mus*, le genre des

souris - une erreur d'océrisation typique ; *Chamaerops fortunei* pour *Trachycarpus fortunei*, un palmier - un changement de classification, très fréquent en biologie ; *Chamærops Fortunei* pour *Chamaerops fortunei* - une infraction au code de nomenclature qui stipule que les épithètes (le second terme d'un taxon) commencent par une minuscule, même quand ils dérivent d'un nom propre (ici, celui de Robert Fortune, un botaniste), et qui stipule aussi de n'employer ni lettre accentuée ni ligature. Le lecteur compétent reconnaîtra aussi les formes abrégées, comme *C. fortunei*, ainsi que leurs variantes fautives, comme *C. Fortunei*.

Concernant les infractions à la règle qui veut qu'un nom de genre commence par une majuscule, et un épithète par une minuscule, on observe toutes les combinaisons possible : *genre espèce*, *genre Espèce* et *Genre Espèce*, que l'on notera par la suite par *mm*, *mM* et *MM*, à côté de *Mm* qui est le seul autorisé aujourd'hui. Le *lecteur compétent* reconnaît l'intention de la désignation formelle d'une espèce à d'autres traits, comme l'usage de l'italique et le ton général de l'article, dont ne rend pas compte l'océrisation. L'entreprise de reconnaître ces entités nommées est donc vouée à l'erreur.

Les erreurs d'une heuristique de reconnaissance des taxons consistent en des faux positifs et des faux négatifs. Les vrais positifs et vrais négatifs constituent les cas où l'heuristique a réussi. Les erreurs seront mesurées en *précision* et *rappel*, et synthétisées dans un indicateur unique, la *f-mesure*. Certains travaux utilisent l'exactitude (*accuracy*), mais c'est inadapté pour mesurer une tâche NER (Jurafsky et Martin, 2009). Dans la suite, les annotations manuelles seront utilisées des deux façons suivantes. Quand il s'agira d'évaluer une heuristique nécessitant une calibration (Section 6), les volumes 12 et 126 serviront de données de calibration, et les volumes 83 et 155 de données de test de la calibration. Quand il s'agira d'évaluer une heuristique sans calibration (Sections 3, 4 et 5) les quatre volumes serviront de données de test.

3 Utiliser littéralement un référentiel contemporain

Une des méthodes classiques de reconnaissance d'entités nommées consiste à utiliser un dictionnaire de noms d'entités (Ehrmann et al., 2021; Nadeau et Sekine, 2007; Nasar et al., 2021). L'intuition biologique prévoit que cette heuristique provoquera de nombreux faux négatifs puisque la nomenclature du vivant évolue constamment, que ce soit dans la structure même de la classification (ex. des genres qui sont scindés ou d'autres qui sont fusionnés) ou dans les conventions typographiques (ex. l'usage des ligatures). Nous avons conduit deux expériences appliquant cette méthode : l'une qui utilise un outil existant qui suit cette approche, l'autre où nous avons codé un reconnaisseur qui exploite un référentiel taxonomique.

LINNAEUS : utiliser un reconnaisseur basé sur un dictionnaire L'outil LINNAEUS (Gerner et al., 2010) réalise en fait deux tâches : reconnaître les taxons linéens et reconnaître les noms vernaculaires anglais correspondant (ex. *apple-tree* pour le genre *Malus*). Nous faisons donc un sous-emploi de cet outil car sa reconnaissance des noms vernaculaires ne marche pas du tout pour le français. LINNAEUS utilise le référentiel du NCBI (*National Center for Biotechnology Information*). Sur nos données, cet outil a une précision à peine supérieure à 16 %, un rappel de presque 20 %, et une f-mesure de presque 18 %. Mais l'objectif de LINNAEUS est d'explorer la littérature médicale contemporaine, pas d'explorer des archives d'une revue de vulgarisation. Nous allons donc préférer développer un reconnaisseur original.

Classifieur	Précision (%)	Rappel (%)	F-mesure (%)
TAXREF strict sans abréviation	100.00	33.8	50.6
TAXREF strict avec abréviation	100.00	37.1	54.1
TAXREF abstrait rang 7	100.00	33.6	50.3
TAXREF abstrait rang 5	99.8	40.6	57.7
TAXREF abstrait rang 4	96.3	50.3	66.1
TAXREF abstrait rang 3	31.7	61.6	41.9
TAXREF abstrait rang 2	4.5	67.0	8.45
TAXREF abstrait *Mm A* (rang 3)	90.32	65.20	75.73
LATIN *Mm A*	70.10	69.76	69.93
TAXREF abstrait *Mm MM A* (rang 3) + seuil	42.81	70.73	53.34
LATIN *Mm MM A* + seuil	63.49	75.77	69.09
TAXREF abstrait *Mm MM mm A* (rang 3) + seuil	38.12	74.63	50.47
LATIN *Mm MM mm A* + seuil	60.97	79.51	69.02

TAB. 1 – Évaluation des différents classifiers

TAXREF strict : calquer un reconnaisseur sur un référentiel La seconde expérience utilise le référentiel taxonomique du Muséum d'histoire naturelle, TAXREF (Gargominy et al., 2021). Ce référentiel prend la forme d'un tableau de plus de 700 000 lignes avec environ une ligne par taxon. Le reconnaisseur cherche dans le texte des occurrences de taxons du référentiel en forme longue ou abrégée. Les deux premières lignes de la Table 1 présentent les résultats obtenus avec cette méthode. Le rappel quantifie ce que prévoyait l'intuition biologique ; presque $\frac{2}{3}$ des dénominations taxonomiques utilisées dans LN ne sont plus en usage ou ne respectent pas les conventions modernes. Si on omet la reconnaissance des formes abrégées, le rappel diminue un peu et donc la f-mesure aussi.

4 Abstraire un référentiel contemporain

TAXREF contient un ensemble de taxons qui sont des chaînes de caractères ; il détermine donc un langage. Avec une précision de 100 %, un rappel de 33-37 %, et une f-mesure de 50-54 %, on peut espérer relaxer ce langage sans trop perdre en précision. L'utilisation d'une distance d'édition s'est révélée décevante, et nous proposons une autre forme de relaxation.

TAXREF abstrait : abstraction par relaxation ciblée du référentiel Un point important des dénominations binômiales est leur air latin. C'est ça qu'il faut préserver dans une relaxation du langage de TAXREF. Nous proposons donc un schéma de relaxation qui ne conserve que les terminaisons. Ex. en posant le niveau d'abstraction à trois lettres terminales, le binôme *Trachycarpus fortunei* devient le modèle de tous les binômes en *pus* et *nei*. Ce faisant, chaque taxon du référentiel devient le modèle d'une famille infinie de binômes qui ont les mêmes terminaisons. Les contraintes sur le début des noms de genre et des épithètes sont complètement relaxées, mais celles qui portent sur les terminaisons sont conservées. De plus, non seulement

la terminaison de chaque terme est correcte, mais les terminaisons des deux termes d'un même binôme sont en accord ; le système « apprend » un latin empirique en lisant le référentiel.

On appellera *rang* la longueur des terminaisons retenues pour l'abstraction. La Table 1 présente les résultats de cette expérience pour les rangs de 7 à 2. De 7 à 4, on n'observe pas de gain de rappel vraiment spectaculaire. Les choses évoluent brutalement au rang 3. Le rappel grimpe alors à 62 %, mais la précision chute à 32 %, et avec elle la f-mesure à 42 %. On voit que ce qui est gagné du côté rappel est perdu du côté précision. Le résultat n'est donc pas globalement meilleur que celui de TAXREF strict.

Il reste à examiner une autre façon d'apprendre le latin, non pas en généralisant un langage donné, mais en consultant la partie de la grammaire latine que le code de nomenclature utilise.

5 LATIN : encoder le code de nomenclature taxinomique

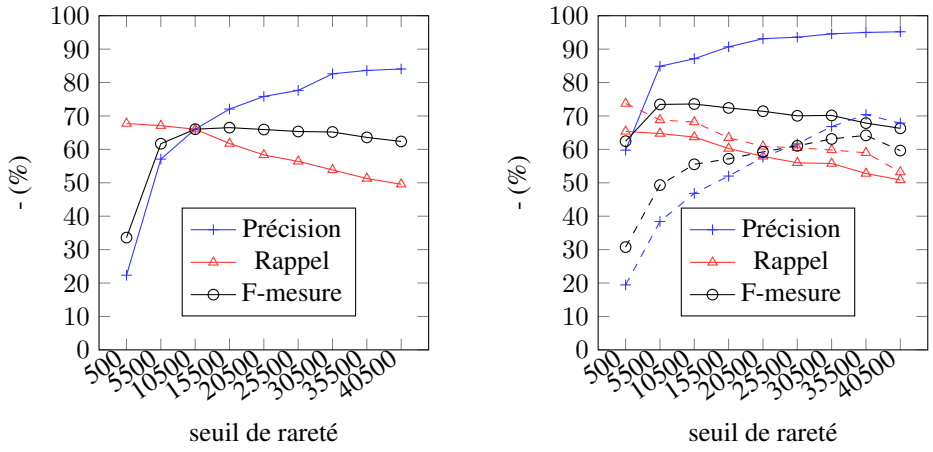
Nous rappelons les grandes lignes des codes de nomenclature puis leur implémentation.

Principaux traits des codes de nomenclature Les taxons sont écrits en italique. Malheureusement, l'océrisation ne préserve pas ce trait. Les taxons sont des binômes : c-à-d. qu'ils consistent en deux parties. La première désigne le genre et s'écrit avec une majuscule initiale. La seconde partie, l'épithète, désigne l'espèce au sein du genre et s'écrit avec une minuscule initiale. Le genre taxinomique et son épithète doivent être écrits en « latin », avec accord en genre grammatical (masculin, féminin et neutre) entre les deux. Les taxons sont soit des paires nominatif-adjectif (ex. *Helleborus niger* ou hellébore noire), nominatif-nominatif (ex. *Panthera leo*, le lion) ou nominatif-génitif (ex. *Trachycarpus fortunei*, le trachycarpus de Robert Fortune). En réalité, les choses sont beaucoup plus complexes. Ex. des genres et espèces sont divisés en sous-genres ou sous-espèces, des codes de nomenclature différents sont utilisés en botanique, zoologie, virologie, et agriculture, et ils sont régulièrement mis à jour.

Mise en œuvre des contraintes Il est facile de trouver les tables des déclinaisons et accords du latin. Les appliquer à des phrases de deux mots selon trois structures grammaticales est aussi facile en calculant une expression régulière à partir de ces tables. Le calcul ne peut guère être fait à la main, mais il peut facilement être automatisé. Cela revient essentiellement à faire une jointure relationnelle selon le nom de genre entre la table du nominatif et elle-même, la table du nominatif et celle du génitif, et la table du nominatif et celle des adjectifs.

Le critère *Mm* (défini en Section 2) est satisfait par beaucoup de débuts de phrase alors que le critère *MM* l'est aussi par beaucoup de noms propres de personne (Prénom Nom). Beaucoup des digrammes du français satisfont le critère *mm*. On voit alors que les règles de base sont très imprécises par nature, alors qu'il va falloir les relaxer encore plus pour s'adapter aux usages observés dans les dénominations binômiales. Afin de pouvoir observer ces phénomènes, nous avons rendu les classifieurs paramètrables par les critères *Mm*, *MM* et *mm*.

La situation ne semble donc pas meilleure qu'avec l'apprentissage empirique du latin, sauf pour le rappel qui devient excellent, 79,1 % (mais, précision 2,4 % et f-mesure 4,7 %). C'est donc une situation duale de celle de TAXREF strict, mais cela rejoint les résultats de TAXREF abstrait pour le rang 2. Utilisé naïvement, le critère de ressembler à du latin n'est donc pas assez précis. Toutefois, on peut ajouter un nouveau critère en observant que les mots des taxons apparaissent assez rarement par rapport à la distribution de tous les mots dans le corpus.



(a) LATIN (b) TAXREF abstrait

FIG. 1 – Évolution des indicateurs de performance en fonction du seuil de rareté

6 Prendre en compte la rareté des entités recherchées

Le critère de rareté des taxons linnéeens est réalisé par un filtre qui exclut les R mots les plus fréquents. On appelle R le *seuil de rareté*. Il est défini comme suit. Considérant l'ensemble des mots de la collection (environ 100 000 mots uniques) triés par ordre décroissant de fréquence, le seuil de rareté est le rang R qui sépare les mots qui sont exclus, c-à-d. tout mot w de rang $r_w < R$, des mots qui restent candidats pour former des taxons, $R \leqslant r_w$.

Afin de calibrer le seuil de rareté, nous avons utilisé les classifieurs LATIN *Mm + A* et TAXREF abstrait *Mm + A* (rang 2 en trait plein et 3 en pointillés) itérativement dans une séquence d'expériences avec des seuils de rareté croissants. Les Figures 1a et 1b présentent les évolutions des indicateurs de performance dans ces expériences. Un choix optimal du seuil de rareté peut être lu au maximum de la courbe de f-mesure. On peut voir qu'il semble raisonnable de choisir un seuil à 15 000 mots pour LATIN et TAXREF abstrait de rang 3.

Les six dernières lignes de la Table 1 présentent les indicateurs de performance pour les différentes méthodes explorées avec un seuil de rareté fixé à 15 000 mots. On peut comparer les performances des différentes approches avec des niveaux de relaxation des contraintes syntactiques équivalentes. Les lignes « TAXREF abstrait *Mm A* (rang 3) » et « LATIN *Mm A* » comparent l'apprentissage par abstraction d'un dictionnaire et le codage d'une grammaire latine en respectant strictement le code actuel. Ici, TAXREF domine LATIN en précision et f-mesure. Les deux lignes suivantes (avec *MM*) étendent la comparaison au cas où les épithètes peuvent être capitalisés. Les deux reconnaisseurs gagnent 5 points de rappel, TAXREF abstrait perd en qualité globale, mais pas LATIN. Enfin, les deux dernières lignes étendent la comparaison au cas *mm*. Les deux reconnaisseurs gagnent encore 4 points de rappels, TAXREF abstrait perd encore plus de qualité globale, alors que LATIN n'en perd presque pas.

En conclusion, LATIN semble une méthode robuste au sens où sa qualité globale résiste à de nombreuses variantes de paramètrage. Au contraire, TAXREF abstrait résiste moins

bien aux relâchements des règles taxonomiques, même si c'est cette méthode qui donne les meilleurs résultats dans le cas des taxons qui respectent les règles contemporaines.

7 Conclusion

L'étude de cas de la reconnaissance des entités nommées de type taxons permet d'aboutir aux conclusions suivantes. L'utilisation d'un référentiel contemporain (TAXREF) ne permet pas de reconnaître efficacement des taxons dans une archive historique. Par construction, la précision est alors de 100 %, mais le rappel n'est que d'environ 30 %. La relaxation ciblée du langage de TAXREF, en préservant les terminaisons, ou le codage des quelques règles de grammaire du latin qui est utilisé dans les nomenclatures du vivant, donnent des résultats inverses de l'exploitation stricte de TAXREF : un bon rappel, mais une précision très faible. Les taxons sont le plus souvent des mots rares dans la distribution des mots du français. En tenir compte en combinaison avec une méthode d'abstraction du latin permet d'augmenter considérablement la précision, sans trop faire baisser le rappel. C'est confirmé par la f-mesure qui croît aussi. On arrive ainsi à des précisions, rappels et f-mesures de l'ordre de 70 %.

Les évaluations présentées ici sont du genre *intrinsèque* (Clark et al., 2012). Elles fournissent des indicateurs formels, mais ne disent rien des inconvénients empiriques à avoir 30 % d'imprécision ou d'oubli dans le cadre de l'application envisagée, c-à-d. l'exploration d'archives par des historiens, philosophes ou journalistes. Il nous semble donc plus important de donner suite à ce travail en se donnant les moyens de faire une évaluation *extrinsèque*. Ce travail est en cours sous la forme d'un navigateur spécialisé.

Nous pensons que les stratégies employées ici pourraient être employées pour la reconnaissance d'entités nommées du domaine de la chimie, où des codes de nomenclature existent aussi, et dans une moindre mesure, du domaine technologique, qui ne possède pas de tels codes, mais procède souvent à des assemblages de mots d'un vocabulaire assez spécifique.

Les scripts des expériences ainsi que les anotations décrites ici sont disponibles dans le dépôt `https://github.com/oridoux/TAXONER`.

Références

Akella, L. M., C. Norton, et H. Miller (2012). Netineti : Discovery of scientific names from text using machine learning methods. *BMC Bioinformatics 13*, 211.

Burdick, A., J. Drucker, P. Lunenfeld, T. Presner, et J. Schnapp (2012). *Digital Humanities*. The MIT Press.

Clark, A., C. Fox, et S. Lappin (2012). *The handbook of computational linguistics and natural language processing*, Volume 118. John Wiley & Sons.

Ehrmann, M., A. Hamdi, E. L. Pontes, M. Romanello, et A. Doucet (2021). Named entity recognition and classification on historical documents : A survey. *CoRR abs/2109.11406*.

Gargominy, O., S. Tercerie, C. Régnier, T. Ramage, Dupont, P. P., Daszkiewicz, et L. Poncet (2021). TAXREF v15, référentiel taxonomique pour la France : méthodologie, mise en œuvre et diffusion.

Gerner, M., G. Nenadic, et C. M. Bergman (2010). Linnaeus : a species name identification system for biomedical literature. *BMC Bioinformatics 11*(1), 1–17.

Jurafsky, D. et J. H. Martin (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

Koning, D., I. N. Sarkar, et T. Moritz (2005). Taxongrab : Extracting taxonomic names from text. *Biodiversity Informatics 2*, 79–82.

Little, D. (2020). Recognition of Latin scientific names using artificial neural networks. *Applications in Plant Sciences 8*.

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*, 3–26.

Nasar, Z., S. W. Jaffry, et M. Malik (2021). Named entity recognition and relation extraction : State of the art. *ACM Computing Surveys 54*.

Nédellec, C., P. Bessières, R. R. Bossy, A. Kotoujansky, et A.-P. Manine (2006). Annotation guidelines for machine learning-based named entity recognition in microbiology. In *Workshop of Data and Text Mining for Integrative Biology*. Springer - Verlag.

Nguyen, N. T. H., R. Gabud, et S. Ananiadou (2019). Copious : A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*.

Pafilis, E., S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, et L. J. Jensen (2013). The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one 8*(6), e65390.

Rockwell, G. et S. Sinclair (2016). *Hermeneutica : Computer-Assisted Interpretation in the Humanities*. The MIT Press.

Sautter, G., K. Böhm, et D. Agosti (2006). A combining approach to find all taxon names (FAT). *Biodiversity Informatics 3*.

Seideh, M., H. Fehri, et K. Haddar (2016). Named entity recognition from arabic-french herbalism parallel corpora. Volume 607, pp. 191–201.

Tissandier, G. (1873). *LA NATURE : Revue des Sciences et de leurs applications aux arts et à l'industrie*. Masson.

Vautrin, G. (2018). *Histoire de la vulgarisation scientifique avant 1900*. EDP sciences.

Summary

Linnean binoms (aka. taxons) are rarely studied as a type of named entities, and so is their extraction from archival texts. We introduce the *competent reader hypothesis*, i.e., the ability to recognize a taxon, even if it is deprecated or ill-composed. This hypothesis is the key to our evaluation process. We compare several approaches for recognizing taxons: dictionary-based, rule-based, and a form of generalization learning. We show that the criteria of looking Latin used alone lacks precision. Finally, we show that a rarity criteria, when combined with the Latin criteria, yields a high quality recognizer with an f-measure of about 70 %.