Traitement automatique des langues

TAL et humanités numériques

sous la direction de
Jean-Gabriel Ganascia
Francesca Frontini

Vol. 60 - n°3 / 2019

ATALA

TAL et humanités numériques

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
60

n°3
2019

TAL et humanités numériques

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

http ://atala.org/revuetal

Traitement automatique des langues

Secrétaire
Peggy Cellier - IRISA, INSA Rennes

Traitement automatique des langues

Volume 60 – n°3 / 2019

TAL ET HUMANITÉS NUMÉRIQUES

Table des matières

TAL et humanités numériques

Jean-Gabriel Ganascia* — **Francesca Frontini** **

* *LIP6 - Sorbonne Université et CNRS*
Jean-Gabriel.Ganascia@lip6.fr
** *Laboratoire Praxiling - Université Paul Valéry Montpellier 3 et CNRS*
francesca.frontini@univ-montp3.fr

1. Introduction

C'est un peu avant le tournant du millénaire, dans la fin des années quatre-vingt-dix, que le terme « humanités numériques » (HN) (*digital humanities* en anglais) fait son apparition de façon massive. Il évoque le virage moderniste et informatique des humanités consécutif à la numérisation des contenus, le terme « humanités » étant entendu au sens anglo-saxon d'étude des œuvres humaines. Pour autant, l'utilisation de l'informatique dans les disciplines relevant des humanités est bien antérieure. Elle remonte à la fin des années quarante, avec le projet *Index Thomisticus* du père Roberto Busa qui se proposait, dès 1949, d'indexer la *Somme théologique* de Thomas d'Aquin à l'aide d'ordinateurs. Quant à l'utilisation de statistiques et de nombres pour étudier les textes littéraires, elle est plus ancienne encore. Ainsi, évoque-t-on parfois les travaux d'Augustus de Morgan qui proposa, dès 1851, une étude quantitative de la fréquence des mots pour caractériser le style des auteurs. En somme, cela fait longtemps que les humanités sont numériques, et ce, dans le double sens du terme, à la fois parce qu'elles emploient des nombres et parce qu'elles recourent aux technologies de l'information et de la communication.

Initialement, ces travaux ont porté sur le texte à la fois par commodité, parce que ce sont les contenus les plus faciles à numériser, et par habitude, parce que cela reconduisait les anciennes pratiques de la philologie, avec la construction d'index, et de la linguistique, avec l'établissement de lexiques. Le champ de l'« informatique littéraire et linguistique » (*Literary and Linguistic Computing* en anglais) (Hockey, 2004) résume bien, dans son intitulé, ce croisement entre le calcul, les disciplines d'érudition, dont les études littéraires font partie, et les sciences du langage. Toutefois, avec le

TAL. Volume 60 – n°3/2019, pages 7 à 12

temps, ce domaine a subi de multiples évolutions. D'un côté, les études sur la langue prirent leur autonomie et s'agrégèrent aux efforts très précoces d'automatisation de la traduction, ce qui donna naissance au traitement automatique des langues (TAL), d'un autre côté les travaux dans les HN s'étendirent à d'autres contenus, en particulier à des contenus multimodaux, images bi et tridimensionnelles, vidéos, sons, etc.

Depuis une vingtaine d'années, le champ des HN est en rapide expansion et ses frontières sont à la fois difficiles à identifier et en constante évolution (Dacos et Mounier, 2015 ; Terras *et al.*, 2013 ; Ganascia, 2015). Du fait de la numérisation des contenus et de la possibilité de les traiter avec des ordinateurs, les humanités se transforment, en particulier les études littéraires, l'histoire, l'archéologie, la sociologie, et cela ouvre la voie à l'émergence de nouvelles pratiques scientifiques que l'on range sous le vocable d'humanités numériques.

Dans ce contexte, même si un certain nombre d'œuvres humaines, qu'il s'agisse de tableaux, d'objets, par exemple de poteries, nous sont données sous forme multimodale, la plupart d'entre elles, que ce soit en littérature, en philosophie, en archéologie ou en histoire, nous parviennent sous forme textuelle. De ce fait, les techniques de traitement des textes sont essentielles pour les HN. Et, parmi ces techniques beaucoup recourent aux techniques du TAL, qui paraissent dès lors d'un immense profit pour les HN et en particulier pour le très vaste sous-domaine des HN que l'on qualifie d'« humanités numériques textuelles ».

Cependant, alors que la recherche actuelle en TAL s'articule généralement autour de tâches bien identifiées et plus ou moins complexes (comme la reconnaissance d'entités nommées, l'analyse syntaxique, l'extraction d'informations, les systèmes questions-réponses, le résumé de texte, etc.), les HN utilisent des techniques et des méthodes de TAL comme outils, en les hybridant à d'autres techniques issues de la fouille de données, de l'algorithmique des chaînes de caractères ou de la théorie des graphes, et en les intégrant dans des scénarios de recherche complexes, allant de l'acquisition à l'annotation et à l'analyse de textes, ces derniers incluant aussi bien des collections de textes bruts, que des éditions numériques hautement encodées. En conséquence, les défis ultimes des HN ne visent pas uniquement à améliorer les performances des outils de TAL, mais aussi, et surtout, leur utilisation dans les différents champs des humanités. Au-delà, la taille des corpus varie considérablement, depuis de grandes bibliothèques comprenant des centaines de milliers d'ouvrages numérisés – avec malheureusement de trop fréquentes erreurs – à de petits ensembles de dizaines ou de centaines de livres. Citons, à titre d'illustration, les travaux d'attribution d'auteurs appliqués aux manuscrits médiévaux de (Pinche *et al.*, 2019), de stylistique outillée sur la poésie espagnole (Ruiz *et al.*, 2017), ou le développement d'études de narratologie dans les romans fondées sur l'emploi de techniques de détection d'entités nommées (voir entre autres (de Does *et al.*, 2017) et (Alex *et al.*, 2019)).

À ces différences de finalité, s'ajoutent la très grande variété et complexité des textes traités. La diversité des types de textes communément traités par les HN, diversité d'époques, de registres ou de genres (poésie, théâtre, etc.), constitue souvent, par sa nature, un défi supplémentaire pour les outils et algorithmes courants. En particu-

lier, les documents historiques consignés dans des variantes linguistiques anciennes peuvent poser des problèmes tant d'un point de vue linguistique que pour la complexité de leur contenu. Et, il en va de même avec les textes littéraires, en particulier avec la poésie, du fait des contraintes métriques et des licences grammaticales qu'elle autorise. Il s'ensuit que des opérations désormais assez bien maîtrisées dans le champ du TAL, comme l'étiquetage syntaxique, la lemmatisation ou la racinisation (*stemming*), présentent, dans le contexte des HN, de nouveaux défis, lorsque les corpus annotés se font rares.

Enfin, malgré, ou plutôt du fait de toutes ces difficultés, les applications des HN peuvent se présenter elles-mêmes comme un banc d'essai idéal pour évaluer les dernières avancées dans le TAL. Cela paraît crucial aujourd'hui, à l'heure où la reproductibilité des résultats se pose avec acuité en sciences en général, et en TAL en particulier (Kovár *et al.*, 2016 ; Cohen *et al.*, 2018). En effet, du fait de la variété des corpus et de la profonde connaissance que les spécialistes des disciplines d'érudition ont de leurs propres corpus, on peut tester efficacement des techniques de TAL dans ces domaines et s'interroger sur la pertinence des méthodologies que l'on met en œuvre.

2. Présentation des articles

Ce numéro spécial de la revue TAL présente une petite anthologie des recherches situées à la croisée des chemins entre les HN et le TAL, un accent particulier étant mis sur des projets dans lesquels les outils du TAL sont développés et/ou appliqués pour annoter, traiter et étudier des contenus textuels provenant de différentes disciplines des humanités. Le parcours que nous proposons vise à mettre en évidence l'apport du TAL en montrant qu'il peut servir à maints égards dans le champ des HN[1] – cela va de la transcription automatique à l'annotation, à l'exploration, à l'analyse sémantique, à l'extraction et à la modélisation de connaissances. Nous verrons aussi que le TAL sert de support à des approches très variées des HN, où l'on aborde toutes sortes de genres littéraires, depuis la poésie homérique, jusqu'aux traductions et aux textes géographiques, écrits dans de multiples variétés linguistiques depuis les textes médiévaux, écrits en latin ou en langue vernaculaire, jusqu'au français moderne. Si la sélection n'a évidemment pas de présomption d'exhaustivité, ne pouvant pas représenter toutes les tendances actuelles, elle offrira sans doute au lecteur une idée de l'ampleur des recherches qui sont en train de définir ce secteur.

Transcription automatique et segmentation thématique de livres d'heures manuscrits (Daille et al.*)*

Ce premier article présente un exemple paradigmatique en ce qu'il montre comment une équipe de recherche, composée de philologues computationnels et de ta-

1. Pour un approfondissement voir le recensement systématique des « *Research Activities* » dans la taxonomie TaDiRAH - *Taxonomy of Digital Research Activities in the Humanities* - `http://tadirah.dariah.eu/vocab/index.php`

listes, s'attaque à la première étape de la chaîne d'extraction d'information textuelle à partir de documents numérisés, à savoir à la segmentation du texte. Il s'agit de procéder à l'analyse de la mise en page et à la reconnaissance de l'écriture dans les manuscrits des livres d'heures, « plus grand best-seller de tout le Moyen Âge », pour citer nos auteurs. Ce travail recourt à de nombreux algorithmes, depuis les réseaux de neurones profonds jusqu'aux chaînes de Markov cachées, et à des approches semi-supervisées. Dans tous les cas, la connaissance approfondie des textes et de leur mise en page facilite la mise en œuvre des algorithmes et permet de choisir la solution la plus appropriée.

Édition comparative intermédiale de séries traductives : exploiter les homologies pour créer des visualisations modulables (Suchecka et al.*)*

La numérisation des textes et l'identification de leur structure sont évidemment préalables à toute analyse outillée des textes. Dans cet article les auteurs nous proposent de comparer différentes traductions françaises du dixième livre des *Métamorphoses* d'Ovide à l'aide de deux outils d'alignement textuel ; ces traductions, équivalentes quant au contenu, ou supposées telles, mais différentes du point de vue linguistique et lexical, permettent de tester les algorithmes d'alignement. L'analyse montre aussi comment les outils TAL proposés aux humanistes doivent tenir compte de contraintes pratiques propres aux HN, notamment de l'encodage en TEI qui se superpose au contenu textuel pur pour y ajouter des éléments structuraux de mise en page et d'édition, dont les systèmes de comparaison devraient pouvoir tirer profit.

Vector space models of Ancient Greek word meaning, and a case study on Homer (Rodda et al.*)*

En poursuivant notre parcours ascendant vers des niveaux d'analyse linguistique de plus en plus élevés, nous arrivons à cette contribution de philologie numérique en langue anglaise. Il s'agit là d'utiliser la sémantique distributionnelle pour explorer des aspects de type lexical, notamment liés à la phraséologie de la langue grecque homérique et en particulier à la récurrence de formules plus ou moins figées comme « Achille aux pieds légers » et aux structures annulaires que leur répétition produit. Dans l'expérience proposée, différents modèles distributionnels sont confrontés à une référence « idéale » dérivée du travail des lexicographes anciens et modernes, afin de pouvoir identifier le paramétrage optimal pour cette tâche spécifique. Si la connaissance philologique nécessaire pour adapter les algorithmes est considérable, l'approche quantitative montre néanmoins ses avantages, car elle permet d'évaluer systématiquement différents aspects - tels la flexibilité sémantique des expressions – qui n'avaient jamais été pris en considération jusque là.

Chronique d'un échec : identification des métaphores dans les écrits des géographes (Mpouli)

Si les articles présentés jusqu'ici offrent une perspective plutôt positive des apports du TAL aux HN, il est important de souligner aussi les difficultés, en particulier

celles qui tiennent à des spécificités textuelles. Cet article explore la délicate question de la détection et de l'annotation des métaphores – et plus généralement de toutes les figures tropes – qui restent l'un des défis les plus complexes en TAL. L'approche choisie se fonde sur l'allocation de Dirichlet latente et vise à identifier les contextes métaphoriques dans lesquels un écart sémantique entre le domaine cible et le domaine source est identifiable. Cette méthode, très utilisée en extraction d'informations, ne semble pas toutefois donner les résultats attendus à cause de la haute spécificité des typologies et sous-typologies textuelles dans le domaine de la géographie, ce qui rend difficile l'apprentissage automatique des domaines alignés. La conclusion est que la transposition immédiate aux HN d'algorithmes et de solutions existantes n'est pas toujours possible et que seule une analyse approfondie permet de circonscrire la question de recherche afin de proposer des solutions plus adéquates.

The names of lighting artefacts : extraction and representation of Portuguese and Spanish terms in the archaeology of al-Andalus (Almeida et al.*)*

Le dernier des cinq articles proposés dans ce numéro spécial nous présente une tout autre perspective, qui croise la linguistique de corpus, l'utilisation de ressources linguistiques numériques, la modélisation des connaissances et les humanités numériques. L'article propose un schéma qui est désormais de plus en plus typique dans les HN : un corpus de textes anciens est constitué puis traité à l'aide d'algorithmes et d'outils d'extraction lexicale avant qu'une terminologie du domaine en soit extraite et que celle-ci soit ensuite modélisée par les experts à l'aide d'une ontologie formelle. Au-delà du sujet fascinant sur lequel porte cet article, à savoir les luminaires dans la culture andalouse, ce travail est représentatif d'une approche de la modélisation qui établit un lien entre l'information linguistique et l'information conceptuelle, tout en préservant la distinction entre les deux plans, celui de la langue et celui des concepts. Cette approche, qui est largement utilisée tant dans les HN qu'en linguistique informatique, est à l'origine des schémas du modèle *Ontolex-Lemon*, qui est maintenant très utilisé pour la représentation des ressources lexicales et leur exploitation avec le TAL.

3. Remerciements

Nous remercions le comité éditorial et scientifique de la revue TAL, ainsi que le comité scientifique invité, en particulier les relecteurs, qui ont contribué par leur temps et leurs efforts à la qualité de ce numéro.

Comité de lecture : Adrien Barbaresi (Berlin-Brandenburg Academy of Sciences), Valérie Beaudouin (Télécom ParisTech), Federico Boschetti (Istituto di Linguistica Computazionale « A. Zampolli » CNR, Pisa), Sascha Diwersy (Université Paul-Valéry Montpellier 3), Antoine Doucet (Université de La Rochelle), Maud Ehrmann (École polytechnique fédérale de Lausanne), Clovis Gladstone (University of Chicago), Agata Jackiewicz (Université Paul-Valéry Montpellier 3), Adam Jatowt (Kyoto University), Mike Kestemont (University of Antwerp), Anas Fahad Khan (Istituto di

Linguistica Computazionale « A. Zampolli » CNR, Pisa), Thomas Lebarbé (Université Grenoble – Alpes), Dominique Legallois (Sorbonne Nouvelle - Paris 3), Dominique Longrée (Université Saint-Louis, Bruxelles), Robert Morrissey (University of Chicago), Małgorzata Niziołek (Pedagogical University, Kraków), Rachel Panckhurst (Université Paul-Valéry Montpellier 3), Javier Perez Guerra (University of Vigo), Michael Piotrowski (Université de Lausanne), Thierry Poibeau (Laboratoire LATTICE, CNRS), Marianne Reboul (École Normale Supérieure de Lyon), Glenn Roe (Sorbonne Université), Laurent Romary (INRIA / Berlin-Brandenburgische Akademie der Wissenschaften, Berlin), Christof Schöch (University of Trier), Sara Tonelli (Fondazione Bruno Kessler, Trento)

4. Bibliographie

Alex B., Grover C., Tobin R., Oberlander J., « Geoparsing Historical and Contemporary Literary Text Set in the City of Edinburgh », *Language Resources and Evaluation*, vol. 53, n° 4, p. 651-675, December, 2019.

Cohen K. B., Xia J., Zweigenbaum P., Callahan T. J., Hargraves O., Goss F., Ide N., Névéol A., Grouin C., Hunter L. E., « Three Dimensions of Reproducibility in Natural Language Processing », *Proceedings of the International Conference on Language Resources & Evaluation (LREC 2018)*, vol. 2018, p. 156-165, May, 2018.

Dacos M., Mounier P., Humanités Numériques : État des lieux et positionnement de la recherche française dans le contexte international., Research Report, Institut français, March, 2015.

de Does J., Depuydt K., van Dalen-Oskam K., Marx M., « Namescape : Named Entity Recognition from a Literary Perspective », *in* J. Odijk, A. van Hessen (eds), *CLARIN in the Low Countries*, Ubiquity Press, p. 361-370, 2017.

Ganascia J.-G., « The Logic of the Big Data Turn in Digital Literary Studies », *Frontiers in Digital Humanities*, vol. 2, p. 7, 2015.

Hockey S., « The History of Humanities Computing », *in* S. Schreibman, R. Siemens, J. Unsworth (eds), *A Companion to Digital Humanities*, Blackwell, Oxford, 2004.

Kovár V., Jakubícek M., Horak A., « On Evaluation of Natural Language Processing Tasks - Is Gold Standard Evaluation Methodology a Good Solution ? », *Proceedings of the* 8^{th} *International Conference on Agents and Artificial Intelligence (ICAART 2016)*, p. 540-545, 2016.

Pinche A., Camps J.-B., Clérice T., « Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis », *Digital Humanities Conference 2019 - DH2019*, ADHO and Utrecht University, Utrecht, Netherlands, July, 2019.

Ruiz P., Martínez Cantón C., Poibeau T., González-Blanco E., « Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets », *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Association for Computational Linguistics, Vancouver, Canada, p. 27-32, August, 2017.

Terras M., Vanhoutte E., Nyhan J., *Defining Digital Humanities : A Reader*, Routledge, London/New York, 2013.

Transcription automatique et segmentation thématique de livres d'heures manuscrits

Béatrice Daille[*] — Amir Hazem[*] — Christopher Kermorvant[†‡] — Martin Maarand[†] — Marie-Laurence Bonhomme[†] — Dominique Stutzmann[§] — Jacob Currie[§] — Christine Jacquin[*]

[*] *LS2N - Université de Nantes, Nantes* *prenom.nom@ls2n.fr*
[§] *Institut de recherche et d'histoire des textes (IRHT), Paris*
prenom.nom@irht.cnrs.fr
[†] *TEKLIA, Paris* *nom@teklia.com*
[‡] *LITIS, Université de Rouen-Normandie, Rouen*

RÉSUMÉ. *Les livres d'heures sont le plus grand best-seller de tout le Moyen Âge, avec plus de 10 000 témoins conservés. Incontournables pour comprendre l'univers mental médiéval, leurs textes ont été très peu étudiés. Ils sont très longs et ont une structure complexe correspondant à l'organisation liturgique médiévale et la prière quotidienne de l'office. Cet article décrit les méthodes et les traitements automatiques mis en œuvre sur les livres d'heures : la reconnaissance de l'écriture manuscrite et la segmentation adaptées à ces manuscrits. L'approche de segmentation semi-supervisée proposée tire profit de la constitution spécifique du manuscrit pour mieux retrouver leur structure malgré le bruit engendré par la reconnaissance de l'écriture.*

ABSTRACT. *Books of Hours are the number one best seller of the Middle Ages, with more than 10 000 copies preserved. They are a crucial witness to the medieval mindset, but their textual contents have been very scarcely studied. They are very long and offer a complex hierarchical entangled structure, with several characteristics specific to medieval daily Prières office. This paper presents the methods and processing applied to books of hours: handwritten text recognition and text segmentation adapted to medieval manuscripts. We propose a weak supervised approach, based on the overarching structure of the manuscripts, that provides the first state-of-the-art results on transcript texts and despite remaining errors for this new challenging task.*

MOTS-CLÉS : *reconnaissance de l'écriture manuscrite, segmentation thématique, livre d'heures.*

KEYWORDS: *handwritten text recognition, text segmentation, Book of hours.*

TAL. Volume 60 – n° 3/2019, pages 13 à 36

1. Introduction

Les livres d'heures sont un recueil de prières à l'usage des fidèles (Leroquais, 1927 ; Wieck *et al.*, 1988). Souvent richement enluminés, et répandus dès le XIII^e siècle en France, au sud des Pays-Bas, en Angleterre et plus tard en Italie et en Espagne, ils constituent une part importante de l'ensemble des manuscrits médiévaux préservés et sont une source d'information sur la vie et la chrétienté au Moyen Âge. Ils font partie des textes les plus lus au Moyen Âge et, par la richesse de leur décor, plusieurs livres d'heures font aussi partie des objets d'art les plus connus du Moyen Âge français, comme le livre d'heures d'Étienne Chevalier peint par Jean Fouquet et les *Très Riches Heures du duc de Berry*, peint par les frères de Limbourg. En empruntant leurs principaux éléments au bréviaire, l'un des types de livres liturgiques que la religion chrétienne utilise pour régler son culte, et en reproduisant ainsi partiellement le contenu de livres destinés aux prêtres et au clergé, ils permettent aux laïques de prier, comme ceux-ci, selon les heures canoniales tout au long de la journée (matines, laudes, prime, tierce, sexte, nones, vêpres, complies), d'où leur nom générique de « livres d'heures ». Leur noyau est en latin (Heures de la Vierge, Heures de la Croix et du Saint-Esprit, office des morts) et présente des additions en latin et dans des langues vernaculaires (souvent en français). Malgré leur succès à l'époque, leur contenu textuel reste actuellement très peu étudié, alors que la production d'un si grand nombre de manuscrits est un phénomène culturel et industriel capital qui manifeste les profonds changements du monde religieux du bas Moyen Âge, avec, à la fois, le développement d'une production livresque proto-industrielle et le passage de l'économie de la demande à celle de l'offre, mais aussi ce que J. Burckhart a nommé « l'éveil de l'individu » (Rosenwein, 2005) et surtout l'intériorisation de la foi, à une époque où l'encadrement ecclésial devient de plus en plus contraignant.

Malgré plus de 10 000 manuscrits témoins conservés, il existe très peu de livres d'heures transcrits en entier et annotés d'un point de vue linguistique. Comme l'affirme Christopher De Hamel (1994) : « *It sometimes seems surprising, therefore, that there is still no critical edition of the text [...]. Its cultural impact (if that is not too pompous a term for an illuminated prayer-book) was wider and deeper than that of many rare literary texts worked over and over again by modern editors. It reached people too with no other knowledge of literacy. Anyone who could be encouraged to edit the first proper printed edition of the Book of Hours since the sixteenth century would win the gratitude of all historians of manuscripts. The task, however, will be made immensely complicated by the number of surviving manuscripts and their endless subtle differences.* » L'une des rares ressources sur le texte des livres d'heures est la base *Beyond Use*, qui contient, en particulier, une section sur l'*Obsecro Te* (Plummer et Clark, 2015). Cette prière à la Vierge a été transcrite et annotée manuellement à partir de plus de 772 livres d'heures (Plummer et Clark, 2015)[1].

Les livres d'heures sont très longs, avec une moyenne de 300 pages. Ils ont une structure complexe correspondant à l'organisation liturgique médiévale et, en parti-

1. `http://www6.sewanee.edu/beyonduse/`

culier, à la prière quotidienne de l'office, avec plusieurs parties, sections et sous-sections et de nombreux textes dits « accessoires ». À l'heure actuelle, la majorité des livres d'heures sont faiblement catalogués. L'étude de l'usage liturgique qui en résulte repose en conséquence sur un faible nombre de points de repère textuels parmi les plus courts (antiennes, versets et répons) (Leroquais, 1927 ; Ottosen, 1993 ; Ottosen, 2008 ; Drigsdahl, 2013). Or, ceux-ci ne reflètent pas la structure globale et n'empêchent pas les ambiguïtés. Un même texte biblique peut apparaître dans des sections ou sous-sections différentes d'un livre d'heures à l'autre. Parmi les textes accessoires, les prières latines et vernaculaires ont, certes, fait l'objet de repérages, mais presque tout reste à faire. De très nombreux textes restent à découvrir : les textes latins sont surtout repérés pour les manuscrits les plus anciens ; les prières françaises vernaculaires font l'objet de recensements (Sonet, 1956 ; Sinclair, 1978 ; Sinclair, 1987 ; Sinclair, 1979 ; Sinclair, 1982 ; Sinclair, 1988 ; Rézeau, 1986), voire d'éditions (Rézeau, 1983), mais, latins comme vernaculaires, de nombreux textes sont inédits et la diffusion de ces textes par les livres d'heures reste à explorer.

Cet article présente nos premiers travaux pour identifier automatiquement la structure logique des livres d'heures, une étape nécessaire pour permettre une analyse textuelle complète par les historiens médiévistes. L'accès au texte à partir des images de livres d'heures numérisés nécessite une transcription automatique de l'écriture manuscrite. Dans un premier temps, une analyse automatique de la mise en page est réalisée pour identifier les différents éléments présents : iconographie, décoration et zones de texte. La transcription des lignes de texte est ensuite réalisée par un système automatique entraîné spécifiquement sur le type d'écriture manuscrite présent dans les livres d'heures.

2. Composition et structure du livre d'heures

Le livre d'heures, apparu au XIII^{e} siècle en se détachant du psautier dont il était un appendice, a évolué au cours du temps, et des textes non présents dans les premières versions ont été ajoutés.

2.1. *Composition du livre d'heures*

Le livre d'heures inclut un certain nombre de textes de référence, possédant les caractéristiques suivantes (Lebigue, 2007).

Antienne (*antiphona*) est une pièce chantée courte (une à deux lignes) dont le texte est généralement d'origine biblique. Dans le cadre des livres d'heures, ces textes apparaissent principalement autour d'un texte central qui peut être un psaume, un groupe de psaumes ou un cantique, généralement, pour les offices présents dans les livres d'heures, avec l'intonation (début du chant pour « imposer l'antienne ») avant le texte central, puis la pièce entière après le texte central. D'autres formes de l'antienne sont possibles : sous une forme complexe dans

l'invitatoire, où elle est dite en entier avant et après le psaume et la doxologie ; après les versets impairs, où seule la fin de l'antienne est chantée après les versets pairs. Chaque suffrage comporte aussi une antienne.

Absolution et bénédiction sont des textes prononcés avant les leçons de matines.

Cantique (*canticum*) est un chant tiré de la Bible, utilisé comme les psaumes, dont sept, dits « bibliques » sont tirés de l'Ancien Testament tels que le Cantique des trois enfants (*Benedicite*) et le Cantique d'Isaïe (*Confitebor tibi*) et sept sont tirés du Nouveau Testament, dont le Cantique de Zacharie (*Benedictus*), le Cantique de Vierge (*Magnificat*) et le Cantique de Siméon (*Nunc dimittis*).

Capitule (*Capitulum*) est une lecture brève tirée de la Bible, présente dans toutes les heures sauf les matines où contiennent des lectures longues.

Doxologie est une formule conclusive de prières. On distingue notamment la grande doxologie « *Gloria in excelsis...* » de la petite doxologie *Gloria Patri...*, qui est en particulier récitée à la fin des psaumes, cantiques et dans les répons ; dans l'office des morts, le verset *Requiem aeternam...* est utilisé comme doxologie.

Hymne (*hymnus*) est un chant métrique ou rythmique d'origine non biblique.

Invocation est la première partie des heures et comprend un ou deux versets de psaumes de la Bible qui invitent à la prière (Ps. 50,17 et Ps. 69,2 à matines, puis seulement Ps. 69,2 aux autres heures), puis la doxologie (*Gloria patri*).

Invitatoire (*invitatorium*) suit, à matines, l'invocation et comprend un psaume, lui-même dit « invitatoire », avec son antienne intercalée ; le psaume invitatoire le plus courant est le psaume 94 (*Venite exultemus*).

Leçon ou Lecture (*lectio*) est extraite de la Bible (dans certains offices, on trouve également des extraits d'œuvres patristiques ou hagiographiques) et lue au sein des nocturnes de matines.

Oraison (*oratio*) est une prière. Des oraisons forment la conclusion des offices (sauf matines), suffrages et litanies.

Preces est une partie de l'office et de la litanie rassemblant des formules de supplication et principalement constituées d'un ou de deux versicules et de leurs réponses, du *Kyrie eleison* et du *Pater noster*.

Psaume (*psalmus*) est un chant de louange. Au nombre de cent cinquante dans la tradition latine, ils sont regroupés, au sein de la Bible, dans le livre des Psaumes. Il s'agit de textes poétiques divisés en versets ; ils constituent le fondement de la liturgie chrétienne et de la prière continuelle de l'Église.

Répons (*responsorium*) est un court chant de méditation après une lecture. Il est composé de (1) un répons proprement dit (anglais « *respond* »), lui-même divisé en (1a) une première partie du répons et (1b) une « réclame » ou « reprise », en

anglais « *partial respond* »), puis (2) un « verset » (lat. *Versus* ou *Versus responsorii*).Il existe des répons de deux sortes : les « répons prolixes », longs, utilisés après les longues leçons de matines, et les « répons brefs » après les capitules. On dit le répons proprement dit (1) en entier (une fois pour le répons prolixe, deux pour le répons bref), le verset (2), la réclame (1b), puis la doxologie et, pour le répons bref, le verset (2).

Verset (*versus*) désigne soit (1) un vers d'un psaume ou d'une hymne, soit (2) la deuxième partie d'un répons (lat. *responsorium*, angl. « *verse* »). Le mot français est parfois utilisé à égalité avec « versicule » ou pour désigner l'ensemble formé par le versicule et la réponse.

Salutation conclut les offices et est constituée du verset *Benedicamus Domino*, de son acclamation *Deo gratias* et d'un verset (généralement *Fidelium animae*).

Versicule (*versiculus*, abrév. *Vers.*) est un vers suivi d'une « réponse » (lat. *responsio* ou *responsum versiculi*, abrév. *Resp.*, angl. *response*). Dans la liturgie collective, le versicule est chanté par les solistes et la réponse par le chœur. Il intervient au début d'un office, après une hymne et dans les *preces*.

Ainsi, la majeure partie des textes constituant les livres d'heures sont extraits de la Bible. D'autres textes, tels que le Notre Père (*Pater noster*), les doxologies, ou d'autres prières interviennent dans la composition des offices ou dans l'agencement des livres d'heures, en particulier la prière *Obsecro Te*, une supplication à la Vierge afin de recevoir son assistance au moment de la mort.

2.2. *Structure du livre d'heures*

Le livre d'heures comporte une structure complexe qui peut varier selon le lieu d'origine, la destination liturgique (« usage liturgique ») et les désirs du commanditaire.
Traditionnellement, le livre d'heures débute avec le calendrier liturgique. Celui-ci permet au fidèle d'avoir connaissance des fêtes religieuses et, le cas échéant, du ou des saints à célébrer selon le jour. Le calendrier peut être suivi d'extraits de chacun des quatre Évangiles (« péricopes évangéliques »), puis par les « Heures de la Vierge » ou petit office de la Vierge (*Officium parvum beatae Mariae Virginis*), office votif en l'honneur de la Vierge Marie. Ces Heures de la Vierge constituent la section la plus importante du livre d'heures. Divisées en huit sections selon chacune des heures de la journée, elles sont composées de psaumes, de cantiques et d'hymnes, tous thématiquement liés, au moins partiellement, à la Vierge. Ces textes sont eux-mêmes séparés par des antiennes, versets et répons. Généralement placées à la suite des Heures de la Vierge, parfois divisées par heures et intercalées à l'intérieur de celles-ci, surtout dans les manuscrits de l'Ouest de la France, se trouvent les Heures de la Croix et les Heures du Saint-Esprit. Deux autres parties sont presque systématiquement présentes : d'une

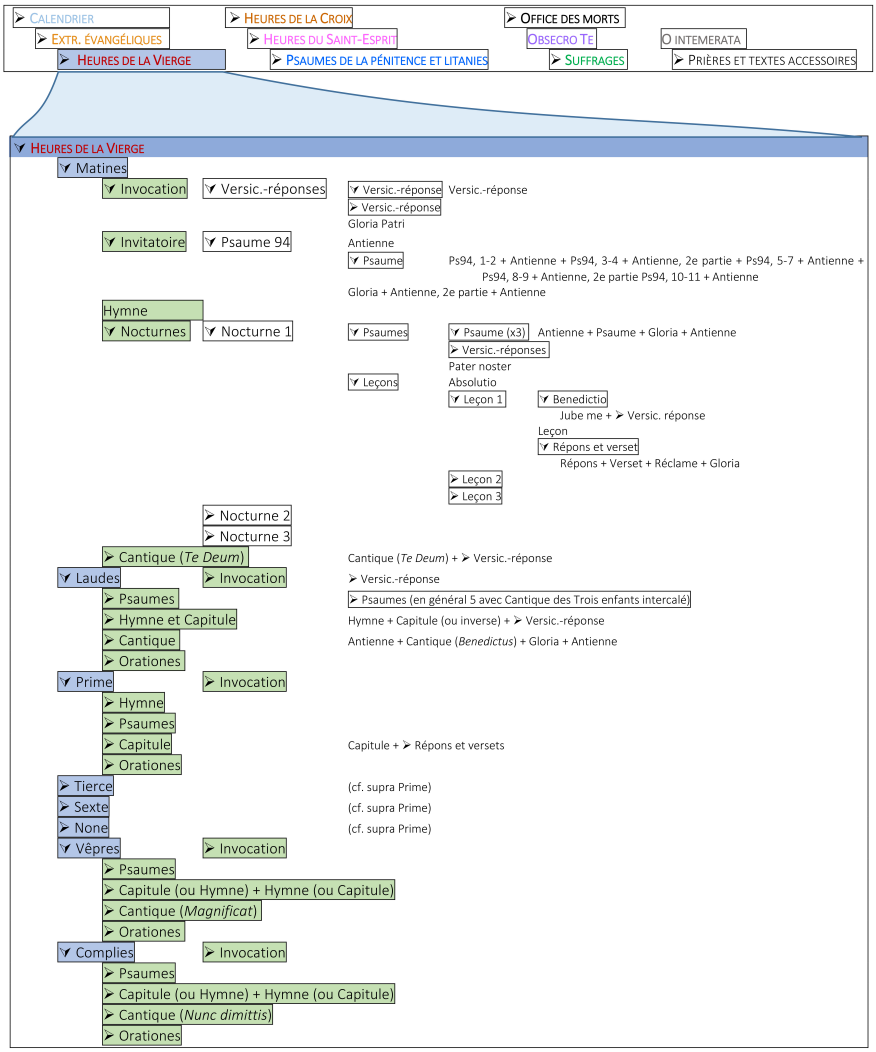

Figure 1. *Structure d'un livre d'heures : parties principales et subdivisions des Heures de la Vierge*

part, l'office des morts, traditionnellement appelé « office » et non « heures » car il ne se compose pas des huit heures, mais seulement de vêpres, matines et laudes, contenant les prières récitées par le clergé pour le salut de l'âme des défunts, et, d'autre part, les sept psaumes de la pénitence ou psaumes pénitentiels (psaumes 6, 31, 37, 50, 101, 129 et 142 dans la numérotation de la Vulgate) qui sont complétés par les

Niveau 1	Nombre de textes de niveau 3
Heures de la Vierge	228
Heures de la Croix	85
Heures du Saint-Esprit	68
Office des morts	156
Suffrages	101

Tableau 1. *Nombres de textes de niveau 3 pour chaque partie constitutive de niveau 1 du livre d'heures ms. Paris, bibliothèque de l'Arsenal, 1194.*

litanies et prières adressées à Dieu, aux anges, à la Vierge et aux saints appelés hiérarchiquement (apôtres, martyrs, confesseurs, etc.). De nombreux autres textes peuvent également être copiés : offices complets, comme les Heures de la Passion ; suffrages, c'est-à-dire des mémoires votifs, composés d'une antienne, d'un verset et d'une oraison ; prières additionnelles, en latin ou en vernaculaire, parfois liées à des indulgences, et dont les deux plus fréquentes sont les prières *Obsecro Te* et *O intemerata*. L'ordre de chacune des parties peut être interverti et tous les offices appelés « heures » se subdivisent en huit sections selon les heures de la journée.

La figure 1 récapitule la structure générique d'un livre d'heures. Plusieurs niveaux de structures sont distingués. Le premier niveau correspond aux grandes catégories de prières comme les péricopes évangéliques ou les Heures de la Vierge. Le second niveau décline les huit prières selon l'échelle temporelle, des matines à complies. Au troisième niveau apparaissent l'invitatoire, les hymnes, les nocturnes et cantiques.

Idéalement, une segmentation automatique doit pouvoir identifier ces trois niveaux. Le tableau 1 indique le nombre de textes présents dans le livre d'heures conservé à la bibliothèque de l'Arsenal (Ms-1194 réserve). Le tableau 2 indique la structuration au niveau 1 de huit livres d'heures. Cet examen comparatif montre que l'ordre type de succession des grandes prières n'est pas stable et qu'elles ne sont pas toutes présentes. Ce tableau illustre les difficultés qui vont être rencontrées pour la segmentation automatique.

3. Reconnaissance du texte manuscrit des livres d'heures

Nous décrivons dans cette section le système de reconnaissance d'écriture manuscrite utilisé pour obtenir une transcription automatique d'un corpus de livre d'heures numérisé sous forme d'images.

Harvard Lat 251	Harvard Lat 253	Harvard Typ 32	Harvard Typ 1000	Harvard Typ 464	Poitiers 1097	Poitiers 43	Poitiers 46
Calendrier	Calendrier	Calendrier	Calendrier	Suffrages	Extr. Évangiles	Calendrier	Calendrier
Extr. Évangiles	Extr. Évangiles	Extr. Évangiles	Vierge	Vierge	Obsecro Te	Extr. Évangiles	Extr. Évangiles
Vierge	Obsecro Te	Obsecro Te	Litanies de la Vierge	Croix	O Intemerata	Obsecro Te	Obsecro Te
Croix	Vierge	Vierge	Psaumes, litanies	Esprit	Prières	Vierge	O Intemerata
Psaumes, litanies	Psaumes, litanies	O Intemerata	Morts	Psaumes, litanies	Heures mêlées (Vierge	Croix	Vierge
Suffrages	Croix	Psaumes, litanies	Croix	Morts	+ Esprit	Esprit	Croix
Prières	Esprit	Croix	Esprit	Extr. Évangiles	+ Croix)	Psaumes, litanies	Esprit
	Morts	Esprit	Prières	Obsecro Te	Psaumes, litanies	Morts	Psaumes, litanies
		Morts			Morts	Suffrages	Morts
		Joies de la Vierge			Suffrages Prières	Sept requêtes à N.S.	Versets de s. Bernard
		Sept requêtes à N.S.			Extr. Évangiles (Passio)	Suffrages Prières	Suffrages

Tableau 2. *Exemples de la segmentation obtenue pour le niveau 1 pour huit livres d'heures (Harvard : Cambridge, Ma., Harvard University, Houghton Library, et Poitiers, médiathèque François-Mitterrand). Une couleur différente est attribuée à chaque élément de niveau 1.*

3.1. *Reconnaissance automatique de documents et d'écritures*

La reconnaissance de l'écriture imprimée (OCR), (*Optical Character Recognition*) sur des documents récents est considérée comme un problème résolu : des systèmes disponibles dans le commerce ou d'accès libre (*open source*) atteignent des taux d'erreurs bien inférieurs à 1 %. La situation est différente en ce qui concerne la reconnaissance de l'écriture manuscrite (HTR) (*Handwritten Text Recognition*) : il existe peu de systèmes commerciaux et les taux d'erreurs restent très variables et bien supérieurs aux taux obtenus par les OCR. La reconnaissance des écritures médiévales, dans une langue très éloignée de la langue actuelle et avec des spécificités sur les formes de lettres et l'usage des abréviations, est encore plus complexe. Il est dans ce cas nécessaire d'entraîner des systèmes spécifiques à la fois à la graphie et au contenu textuel. Les récentes avancées apportées par les techniques d'apprentissage statistique à base de réseaux de neurones profonds ont grandement amélioré les performances des systèmes qui sont maintenant capables de lire une grande diversité d'écritures anciennes après entraînement, comme l'ont montré de récentes applications (Bluche *et al.*, 2017a ; Lang *et al.*, 2018) et les compétitions internationales (Sánchez *et al.*, 2016 ; Sánchez *et al.*, 2017 ; Strauß *et al.*, 2018).

Avant d'appliquer un système de reconnaissance d'écriture, il est d'abord nécessaire d'analyser la structure de l'image des documents afin d'en extraire les zones de texte. Cette étape d'analyse de la mise en page (DLA) (*Document Layout Analysis*) est elle aussi beaucoup plus simple sur les documents imprimés que sur les documents manuscrits. L'extraction des lignes de texte dans un document manuscrit est généralement rendue plus complexe par les variations de taille d'écriture et l'inclinaison des lignes. Là encore, les modèles les plus performants actuellement sont les modèles par apprentissage automatique à base de réseaux de neurones profonds (Diem *et al.*, 2017 ; Renton *et al.*, 2018 ; Moysset *et al.*, 2018 ; Ares Oliveira *et al.*, 2018 ; Grüning *et al.*, 2018). Un exemple d'analyse d'une page de livre d'heures est présenté sur la figure 2.

3.2. *Description du système de reconnaissance d'écriture*

Un système complet de transcription automatique de document est composé d'un certain nombre d'étapes exécutées séquentiellement. Premièrement, les lignes de texte sont localisées dans chaque image de page du manuscrit numérisé. Ces lignes de texte sont ensuite extraites, et le système de reconnaissance d'écriture est appliqué sur chacune des imagettes de ligne. La reconnaissance d'écriture comprend elle-même deux étapes, l'application d'un modèle optique qui reconnaît des caractères, des fragments de caractères ou de mots, et l'application d'un modèle de langue qui détermine les séquences de caractères et de mots les plus vraisemblables.

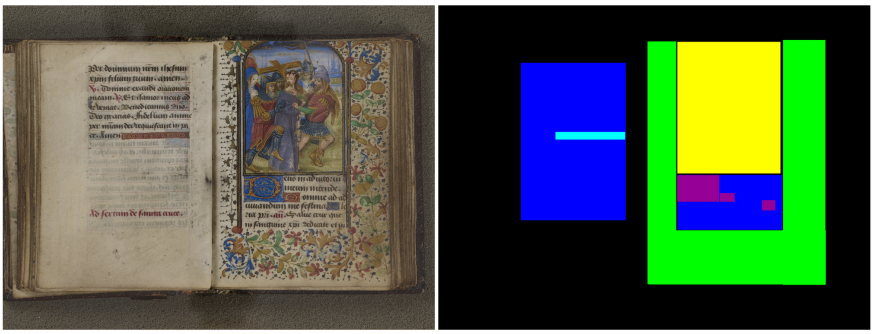

Figure 2. *Analyse d'une double page de livre d'heures : texte (bleu foncé), marge ornée (vert), miniature (jaune), lettrine (violet) et bout de ligne (bleu clair)*

3.2.1. *Détection des lignes de texte*

La détection des lignes de texte a été réalisée avec le logiciel Transkribus[2], une plate-forme de traitement de documents développée principalement à destination des chercheurs en humanités et pour le traitement des documents anciens. Transkribus permet à la fois de réaliser des opérations automatiques sur les images, comme la localisation des régions de texte et l'identification des lignes, mais aussi d'annoter manuellement les documents. La constitution d'un corpus de pages de livres d'heures annotées a ensuite permis d'entraîner un système spécifique de détection des zones et lignes de texte (Boillet *et al.*, 2019).

Un exemple d'extraction des lignes de texte est présenté sur la figure 3. Le texte est écrit de différentes couleurs, présente des lettrines (initiales décorées) et une marge ornée.

3.2.2. *Reconnaissance de l'écriture manuscrite*

Nous avons développé un système de reconnaissance de l'écriture manuscrite des livres d'heures fondé sur la librairie logicielle KALDI[3]. Bien qu'initialement développée pour la reconnaissance de la parole, elle peut être également utilisée pour la reconnaissance d'écriture car les deux applications partagent de nombreux points communs, surtout depuis la généralisation de l'utilisation des réseaux de neurones profonds pour ces deux applications (Bluche *et al.*, 2017b).

Le système développé repose sur une combinaison de réseaux de neurones profonds et de modèles de Markov cachés (HMM) (Peddinti *et al.*, 2015). Le modèle optique, en charge de modéliser la forme des lettres, est composé de plusieurs couches de réseaux de neurones à convolution suivies de couches TDNN (*Time Delay Neural*

2. `https://transkribus.eu`
3. `https://github.com/kaldi-asr/kaldi`

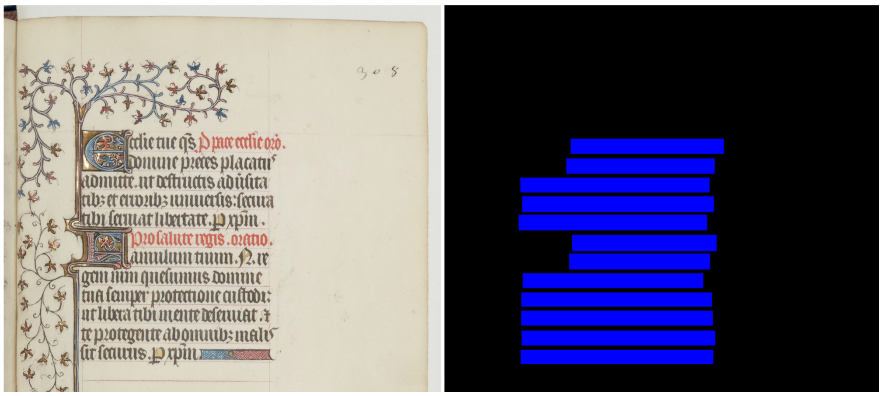

Figure 3. *Extraction des lignes de texte (en bleu) sur une page de livre d'heures présentant une marge ornée et des lettrines*

Networks) qui permettent de modéliser les caractères en contexte. Les prédictions sont ensuite utilisées par un modèle HMM qui modélise les mots comme des séquences de HMM de caractères. Un modèle de langue statistique de type n-gramme est ensuite utilisé pour modéliser les séquences de mots.

Le modèle a été entraîné sur des données provenant de trois corpus de documents médiévaux manuscrits, Psautiers, ECMEN et Fontenay, constitués dans le cadre de précédents projets de recherhche : ORIFLAMMS[4] (*Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts*) et ECMEN (écriture médiévale et outils numériques). Les corpus Psautiers et Fontenay sont en latin, tandis que le corpus ECMEN est en ancien français. Dans un premier temps, il n'a pas été réalisé de détection de la langue des livres d'heures, ce qui aurait permis d'utiliser un modèle spécialisé soit en latin soit en français, car la quantité de données d'entraînement et de test est trop faible. Cette approche sera testée lorsque plus de documents annotés seront disponibles.

Le modèle a été évalué sur 247 lignes transcrites manuellement, issues des *Obsecro Te* de huit livres d'heures du corpus cible. Aucune transcription complète de livre d'heures n'étant disponible dans notre corpus, le système n'a été appris sur aucune donnée issue du corpus cible afin de les réserver pour l'évaluation des performances de la transcription automatique. Les données sont réparties en trois ensembles : les données d'entraînement, de validation et de test, comme présenté dans le tableau 3.

Ces documents sont assez hétérogènes, tant par la qualité des images, leur type (en couleur ou en niveaux de gris), la précision du découpage des lignes et les choix de transcriptions opérés. Cette hétérogénéité se manifeste dans les résultats de recon-

4. `https://oriflamms.hypotheses.org/`

	Entraînement	Validation	Test
Psautiers	4500	660	0
ECMEN	2000	542	0
Fontenay	723	0	0
Obsecro Te	0	0	247

Tableau 3. *Répartition des données et tailles en nombre de lignes des différents ensembles pour l'entraînement et l'évaluation du modèle de reconnaissance d'écriture manuscrite*

	Entraînement	Validation	Test
WER	14.51	24.46	34.19
WER (rescored)	–	32.36	26.32
CER	8.93	11.07	11.21
CER (rescored)	–	14.10	9.90

Tableau 4. *Évaluation des performances de la reconnaissance d'écriture manuscrite sur les différents ensembles, selon les taux d'erreurs mots (WER) et caractères (CER), avec (rescored) ou sans application du modèle de langue*

naissance obtenus et présentés dans le tableau 4. Ces résultats sont évalués selon deux métriques : le taux d'erreurs mots (WER) (*word error rate*) qui mesure le nombre de mots incorrects dans la transcription fournie par le système par rapport à la transcription humaine. Ce taux prend en compte les substitutions, insertions, suppressions et peut donc être supérieur à 100 %. Le taux d'erreurs caractères (CER) (*character error rate*) est son équivalent mesuré au niveau des caractères. Pour mesurer l'impact du modèle de langue, les taux WER et CER sont mesurés avant (*WER, CER*) et après réestimation des hypothèses de séquences de mots par le modèle de langue (*WER rescored, CER rescored*). Cette évaluation est réalisée en ignorant les confusions entre majuscules et minuscules, la ponctuation, et en assimilant les paires de lettres *u* et *v* ainsi que *j* et *i* qui sont des lettres identiques à cette époque.

Les taux d'erreurs sur le corpus de validation, qui contient des lignes de texte issues du même type de documents que l'ensemble d'entraînement (Psautiers et ECMEN), sont plus faibles que les taux d'erreurs sur l'ensemble de test, qui contient des lignes de texte issues de documents complètement disjoints (*Obsecro Te*). Les taux d'erreurs du modèle de reconnaissance d'écriture sont assez élevés sur le corpus de test, car aucune donnée d'entraînement n'est disponible pour le corpus cible. Des exemples d'erreurs sont présentés sur la figure 4. Cependant, ces taux élevés n'empêchent pas l'identification des textes comme il sera montré dans les sections suivantes.

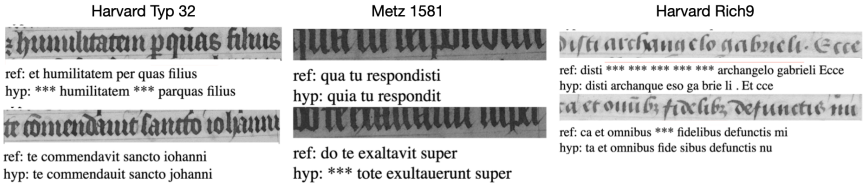

Figure 4. *Exemples de résultats de reconnaissance (*hyp*) et d'annotations manuelles (*ref*) sur les manuscrits Harvard Typ 32, Metz 1581, Harvard Rich 9. À gauche et au centre, écriture de type « Textualis », à droite de type « Cursiva ». À gauche, présence d'abréviations (*et *non reconnu,* per *reconnu comme 'par' qui est correct dans d'autres contextes,* m *correctement reconnu), confusion i/j et u/v sur la dernière ligne (les lettres ramistes ne sont pas distinguées au Moyen Âge et sont une restitution des annotateurs). Au centre, erreurs de segmentation, le haut de la ligne de texte est tronqué. À droite, erreurs de segmentation, confusion c/t, abréviations correctement reconnues.*

4. Segmentation de textes

La segmentation de textes est une problématique proche de l'analyse thématique et recouvre trois tâches distinctes (Ferret *et al.*, 1998) :

1) la segmentation thématique où le texte est découpé en segments thématiquement homogènes ;

2) l'identification thématique qui assigne aux segments de textes un thème. Les thèmes sont préalablement connus ;

3) le suivi thématique qui analyse les relations existant entre les thèmes des segments. Si les thèmes sont très différents, la segmentation thématique est conservée, si à l'inverse ils constituent une spécialisation ou une généralisation, des liens hiérarchiques peuvent être créés entre segments.

4.1. *Segmentation thématique*

La segmentation thématique vise à retrouver la structure sous-jacente d'un document. La segmentation thématique est considérée comme linéaire quand le texte est segmenté en sous-thèmes successifs (Skorochod'ko, 1972) ou comme hiérarchique quand il s'agit de distinguer thèmes et sous-thèmes (Grosz et Sidner, 1986). Les approches état de l'art font l'hypothèse d'une corrélation entre segments et thèmes. Deux segments adjacents seront réunis s'ils sont corrélés. Inversement, si la corrélation calculée apparaît comme faible, une frontière sera insérée entre les deux segments (Hearst, 1997 ; Choi, 2000 ; Riedl et Biemann, 2012). La segmentation thématique est nécessaire pour effectuer l'identification et le suivi thématique. Elle peut

s'appuyer sur le contenu textuel où chaque thème est caractérisé par un vocabulaire spécifique. La cohésion lexicale s'appuie sur la distribution des mots afin d'identifier les changements significatif de vocabulaire révélateurs de changements de thèmes (Hearst, 1994). Des marqueurs de rupture de thème peuvent être utilisés comme, à l'oral, l'intonation ou le silence et, à l'écrit, les caractères de mise en page (titres, espacements, séparateurs, liste d'éléments), les connecteurs de discours ou encore les expressions typiques fortement corrélées avec des frontières de segments thématiques. Les méthodes fondées sur la cohésion lexicale fonctionnent bien pour une segmentation linéaire (Hearst, 1994 ; Choi, 2000). Pour une segmentation hiérarchique, des méthodes plus élaborées requérant une analyse en sous-thèmes sont nécessaires (Yaari, 1997 ; Eisenstein, 2009).

Les principales approches utilisées pour segmenter un texte effectuent soit une analyse lexicale pour détecter les changements de thèmes à l'aide de patrons de co-occurrences (Hearst, 1997) comme les marqueurs de discours (Nomoto et Nitta, 1994), soit calculent la cohésion lexicale (Morris et Hirst, 1991) en exploitant des récurrences lexicales (Hearst, 1994) ou la présence de relations sémantiques fournies par un thésaurus (Morris et Hirst, 1991), un dictionnaire (Kozima, 1993) ou un réseau de collocations construit automatiquement (Ferret *et al.*, 1998).

La cohésion lexicale a inspiré un nombre important d'approches non supervisées. Ces méthodes sont les plus populaires, car elles sont indépendantes du document et ne nécessitent pas de phase d'apprentissage. Les principales approches sont TextTiling (Hearst, 1994), qui identifie localement les ruptures de la cohésion lexicale à l'aide de la mesure statistique du tf-idf et du cosinus, C99 (Choi, 2000), qui mesure globalement la cohésion lexicale au sein de chaque segment et cherche à en maximiser la cohésion lexicale, LSeg (Galley *et al.*, 2003), qui exploite les récurrences lexicales, U00 (Utiyama et Isahara, 2001) fondé sur les modèles probabilistes à facteurs latents, TopicTiling (Riedl et Biemann, 2012), exploitant l'allocation de Dirichlet latente *Latent Dirichlet Analysis (LDA)*, TOPICOLL (Ferret, 2002), exploitant conjointement la récurrence lexicale et des co-occurrences lexicales pour l'analyse thématique, etc. La cohésion lexicale a été appliquée sur deux genres principaux de textes : les documents scientifiques et techniques (Hearst, 1997) où la répétition de termes spécifiques du domaine constitue un indice fiable et les textes narratifs (Morris et Hirst, 1991 ; Kozima, 1993) où il est nécessaire d'utiliser des ressources lexicales pour identifier des relations sémantiques, la récurrence lexicale n'étant pas suffisante. Ferret *et al.* (1998) sont les premiers à proposer une approche mixte en combinant récurrence lexicale et identification de relations sémantiques.

Un ensemble de méthodes employant l'apprentissage supervisé a aussi été proposé pour traiter des discours (Joty *et al.*, 2015), des dialogues multiparties et des forums de chats (Hsueh *et al.*, 2006 ; Hernault *et al.*, 2010) ou pour segmenter des textes au niveau phrastique de manière à identifier les « unités élémentaires du discours » (Hernault *et al.*, 2010 ; Joty *et al.*, 2015). Ces approches combinent des traits de nature différente comme les indices de cohésion lexicale et les caractéristiques de dialogue dans différents classifieurs : arbre de décision (Hsueh *et al.*, 2006), champs

conditionnels aléatoires (CRF) (Hernault *et al.*, 2010 ; Joty *et al.*, 2015). Les travaux les plus récents utilisent des réseaux profonds : TextTiling intègre des plongements lexicaux pour la segmentation de dialogues de questions-réponses (Song *et al.*, 2016), des modèles séquentiels pour la segmentation de dialogues multiparties pour identifier les unités élémentaires de discours (Shi et Huang, 2019) ou encore des modèles d'apprentissage par renforcement (Takanobu *et al.*, 2018). Récemment, Li *et al.* (2018) ont proposé SegBot, un réseau neuronal récurrent (RNN) bidirectionnel couplé avec un mécanisme d'attention qui peut segmenter soit en unités élémentaires de discours, soit en unités thématiques.

Les méthodes ci-dessus s'appliquent principalement pour une segmentation linéaire. L'une des premières approches effectuant une segmentation hiérarchique a utilisé un algorithme de *clustering* hiérarchique (Yaari, 1997). Eisenstein (2009) a proposé un modèle génératif bayésien avec programmation dynamique. Enfin, pour inférer la structuration logique du texte, des traits additionnels de marques caractéristiques de changement de thèmes ont été inclus au sein du CRF (Fauconnier *et al.*, 2014).
Toutes ces méthodes ont été appliquées sur des textes scientifiques, narratifs ou des dialogues écrits ou retranscrits. Dans cet article, nous nous attelons à la segmentation automatique du livre d'heures après sa retranscription par les méthodes de reconnaissance de l'écriture manuscrite qui présentent un taux d'erreurs important.

4.2. *Approche semi-supervisée de segmentation de livres d'heures*

Nous proposons une approche semi-supervisée fondée sur une représentation par plongements de mots des parties du livre d'heures. Notre approche exploite l'idée de

Algorithm 1 Approche semi-supervisée

$Refs = ObsecroTe, Psalm6, Psalm50...$
$Blocks = block1, block2, block3...$
$bestblocks \leftarrow Empty$
for $doc \in Refs$ **do**
  $Max \leftarrow 0$
  for $block \in Blocks$ **do**
    if $sim(doc, block) < Max$ **then**
      $bestblocks[doc] \leftarrow block$
      $Max \leftarrow sim(doc, block)$
    end if
  end for
end for
print $bestblocks$

décomposition en blocs et de mesure de similarité comme utilisée dans (Choi, 2000 ; Utiyama et Isahara, 2001). Elle diffère, cependant, dans la manière de représenter les blocs d'un texte et dans la prise de décision quant à la sélection ou non d'un

bloc candidat, comme segment à part entière. En effet, les approches traditionnelles utilisent au sein d'un même document à segmenter, une similarité interbloc pour, soit les fusionner s'ils sont suffisamment similaires, soit, dans le cas contraire, décider d'une rupture de cohésion lexicale. Notre approche, en revanche, utilise une similarité, non pas entre blocs d'un même document, mais entre un bloc d'un document et une liste de parties du livre d'heures. Ainsi, la cohésion lexicale n'est plus détectée au niveau interne du texte à segmenter, mais au niveau externe, et ceci en s'appuyant sur une base de référence externe contenant des textes préalablement annotés. Notre approche peut être vue comme un alignement de textes de référence des livres d'heures et des textes transcrits découpés arbitrairement en blocs distincts. Nous illustrons notre démarche dans l'algorithme 1.

Premièrement, nous avons à disposition une liste de textes présents dans les livres d'heures qui vont nous servir de liste de référence ($Refs$). À partir de $Refs$, nous construisons une représentation par plongements de mots[5] de chaque texte. Le texte *Obsecro Te*, par exemple, sera représenté par un vecteur de plongements de mots. Ce vecteur est calculé à partir d'une combinaison linéaire des vecteurs de plongements des mots qui le composent (Arora *et al.*, 2017).

Deuxièmement, nous effectuons autant de segmentations de la transcription d'un livre d'heures que de textes présents dans $Refs$. Par exemple, pour le texte *Obsecro Te*, nous découpons le livre d'heures en blocs de taille égale à celle de l'*Obsecro Te*. Nous réitérons cette segmentation en blocs pour chaque texte de $Refs$. Chaque bloc (représenté dans $Blocks$) est représenté par un vecteur de plongements de mots de la même manière que pour $Refs$.

Enfin, pour chaque document référence, nous mesurons la similarité entre le vecteur de plongements de mots de celui-ci et les vecteurs de plongements de mots de tous les blocs de l'ensemble $Blocks$. Le bloc le plus similaire sera considéré comme étant la section correspondante du livre d'heures. Cette procédure est répétée pour chaque section et sous-section du livre d'heures.

La structure des livres d'heures n'étant pas identique d'un livre à l'autre comme l'illustre le tableau 2, un seuil appris sur un corpus d'entraînement est utilisé pour éliminer les documents de référence qui présentent une similarité faible au regard des blocs de la transcription. Une des limites de notre approche est l'hypothèse que la taille de la section de référence et de celle de la section présente dans la transcription sont similaires. Cependant, nous considérons que l'impact de la variabilité en termes de taille est faible vis-à-vis de la représentation par plongements de mots des blocs. Aussi, nous laissons une analyse fine de détection de ruptures entre les sections et sous-sections pour des travaux futurs.

5. Nous utilisons le modèle préentraîné latin de FastText qui s'appuie sur Wikipédia `https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md`

4.3. *Expériences et résultats*

Dans une première expérience, nous évaluons plusieurs approches état de l'art ainsi que notre approche semi-supervisée sur un livre d'heures construit artificiellement et composé de plusieurs sous-sections d'un livre d'heures. La motivation de l'utilisation de données artificielles est de comparer les performances des méthodes sur des données propres, dépourvues d'erreurs de transcription, et sur les données bruitées proposées par la reconnaissance des caractères. Dans une seconde expérience, nous évaluons les différents systèmes sur la transcription du livre d'heures Harvard253[6], extrait de la collection digitale de Harvard (Harvard digital collections).

4.3.1. *Données expérimentales*

Le tableau 5 résume le nombre de segments de niveau 1, le nombre de segments de niveau 2 ainsi que la somme des nombres de segments des deux niveaux de segmentation notée « Niveaux 1 et 2 ». Le livre d'heures artificiel est construit aléatoirement à partir de 29 sous-sections de référence comportant entre autres les prières *Obsecro Te* et *O Intemerata* ainsi que plusieurs psaumes. La transcription du livre d'heures Harvard253 est composée de huit sections au premier niveau (comme l'illustre le tableau 1) et de 38 sections au niveau 2.

	Nombre de sections par niveau		
Livre d'heures	#Niveau 1	#Niveau 2	# Niveaux 1 et 2
Artificiel	-	29	29
Harvard253	8	38	46

Tableau 5. *Représentation du nombre de sections par niveau de la transcription Harvard253 ainsi que du livre d'heures artificiel*

4.3.2. *Mesures d'évaluation*

Les approches sont évaluées en termes de score d'erreurs par P_k (Beeferman *et al.*, 1999) et Windowdiff (WD) (Pevzner et Hearst, 2002). P_k est une mesure d'erreurs qui combine le rappel et la précision pour estimer la contribution relative de différents types de traits. Cependant, elle possède plusieurs désavantages et quelques points faibles épinglés par Pevzner et Hearst (2002). P_k est sensible à la variabilité de la taille des segments. Elle pénalise plus sévèrement les faux négatifs que les vrais positifs. Enfin, elle surpénalise les erreurs de segmentation qui sont proches des frontières (*near misses*) correctes. Pour pallier les manques de P_k, une seconde mesure, aussi état de l'art, WindowDiff (WD) est considérée. Cette mesure qui est une variante

6. `https://curiosity.lib.harvard.edu/medieval-renaissance-manuscripts/catalog/34-990094032810203941`

de P_k, pénalise de la même manière les faux positifs et les segmentations proches des vraies frontières.

4.3.3. *Approches état de l'art*

Nous évaluons, plusieurs approches état de l'art qui sont TextTiling (Hearst, 1994), le modèle par *clustering* (C99) (Choi, 2000), le modèle probabiliste par programmation dynamique (U00) (Utiyama et Isahara, 2001), le modèle de graphes partitionné par rupture minimale (MinCut) (Malioutov et Barzilay, 2006) et un modèle hiérarchique Bayésien (HierBays) (Eisenstein, 2009). Nous évaluons aussi une approche par modèle thématique TopicTiling (Riedl et Biemann, 2012). D'autres modèles par apprentissage comme dans (Koshorek *et al.*, 2018) auraient pu être considérés mais le manque de données d'apprentissage rend leur utilisation inefficace à ce stade.

4.4. *Résultats*

Le tableau 6 illustre les performances des différentes approches lors de la première expérience menée sur les données artificielles. Comme le montrent les résultats, aucune des approches état de l'art n'obtient des performances satisfaisantes. L'approche pionnière TextTiling obtient des résultats très faibles (P_k = 54 % et WD = 55,7 %). Une modélisation thématique par TopicTiling, bien que meilleure en termes de P_k avec un score de 42,5 %, obtient de moins bons résultats que TextTiling en termes de WD avec un score de 58,3 %. Ceci laisse à penser que le modèle thématique construit avec la LDA n'a pas permis une segmentation efficace. Un manque de données d'entraînement pourrait expliquer ces résultats. La meilleure approche état de l'art, MinCut, obtient un score P_k de 39 % et un score WD de 42,6 %. Enfin, notre approche semi-supervisée obtient les meilleurs résultats avec un score P_k de 27,5 % et un score WD de 29,2 %.

Le tableau 7 illustre les résultats des expériences menées sur la transcription du livre Harvard253 à la fois au premier niveau, au deuxième niveau ainsi que sur les deux niveaux de segmentation notés « Niveaux 1 et 2 ». La encore, nous observons les mêmes comportements que ceux constatés dans l'expérience précédente. Les résultats des approches état de l'art sont décevants. Cependant, une exception est à noter au niveau 1 concernant les approches U00, avec un score P_k de 23,6 % et HierBays, avec un score P_k de 14,2 % et un score WD de 25,3 %. De manière générale, notre approche obtient les meilleurs résultats au niveau 2 et au niveau d'une évaluation sur les deux niveaux (Niveaux 1 et 2) avec un score P_k de 29,9 % et un score WD de 31,4 % au niveau 1 et un score global de 31,4 % en termes de P_k et 32,6 % en termes de WD.

Ces résultats montrent d'une part, la pertinence d'utiliser une approche semi-supervisée dans ce type de textes que sont les livres d'heures et, d'autre part, que les erreurs de segmentation ont un impact faible sur la segmentation. Il est à noter que l'approche HierBays obtient les meilleurs résultats au niveau 1. Ainsi, une com-

	Data	
Approche	P_k	WD
TextTiling	54,0	55,7
C99	44,0	44,2
U00	47,6	51,7
MinCut	39.0	42.6
HierBays	41,3	50,3
TopicTiling	42,5	58,3
Approche proposée	**27,5**	**29,2**

Tableau 6. *Analyse de différentes méthodes de segmentation avec P_k et WindowDiff (WD) sur le livre d'heures artificiel. P_k et WD sont de taux d'erreurs, les scores les plus faibles caractérisent les meilleures méthodes.*

binaison de notre approche semi-supervisée avec l'approche HierBays pourrait être envisagée afin d'améliorer les résultats au niveau 1.

	Niveaux de segmentation					
	Niveau 1		**Niveau 2**		**Niveaux 1 et 2**	
Approche	P_k	WD	P_k	WD	P_k	WD
TextTiling	66,9	99,9	48, 1	60,4	46,0	57,5
C99	68,7	96,8	60,0	676	59,2	66,1
U00	23,6	39,5	38,0	39,4	35,6	38,7
MinCut	40,9	49,2	48,4	52,1	45,2	48,7
HierBays	**14,2**	**25,3**	36,7	39,9	32,9	38,5
TopicTiling	60,3	87,0	42,0	48,3	42,0	47,4
Approche proposée	27,2	33,5	**29,9**	**31,4**	**31,4**	**32,6**

Tableau 7. *Analyse de différentes méthodes de segmentation avec P_k et WindowDiff (WD) pour les deux premiers niveaux de segmentation du livre d'heures. P_k et WD étant des taux d'erreurs, les scores les plus faibles caractérisent les meilleures méthodes.*

5. Conclusion

Nous avons présenté dans cet article une chaîne globale de traitements automatiques du livre d'heures, allant de la reconnaissance de l'écriture manuscrite médiévale sur parchemin à sa segmentation en parties. Les traitements consistent en une première étape dédiée à la transcription du livre d'heures et en une seconde qui vise à fournir une segmentation hiérarchique de plusieurs niveaux. Le taux d'erreurs mots et caractères de la reconnaissance du document reste important, il est dû au manque de données d'entraînement. Nous avons évalué la segmentation au premier et second niveau en appliquant les principales approches état de l'art de segmentation thématique. Les résultats ne sont pas satisfaisants à l'exception de l'approche HierBays pour le seul premier niveau. Nous avons proposé une approche semi-supervisée entraînée sur les textes de référence connus, constitutifs du livre d'heures. Cette approche produit des résultats encourageants sur les deux niveaux de segmentation du livre d'heures. Ils laissent à penser que notre approche de segmentation n'est que peu sensible, voire insensible, aux erreurs de transcription puisqu'un même comportement à été observé à la fois sur un livre d'heures artificiel propre et sur le livre d'heures transcrit bruité. À visée globale, ce travail s'inscrit dans un continuum d'analyses et d'interprétations des textes liturgiques anciens que sont les livres d'heures. Si pour l'heure, nous ne sommes qu'au début de la constitution d'une chaîne de traitement robuste, nous envisageons, comme prochain travail, d'étendre l'évaluation à une base de données plus conséquente de livres pour ensuite permettre aux historiens d'interpréter nos résultats à des fins historiques et anthropologiques.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet HORAE (Hours - Recognition, Analysis, Editions) a bénéficié d'une aide de l'Agence nationale de la recherche portant la référence ANR-17-CE38-0008.

6. Bibliographie

Ares Oliveira S., Seguin B., Kaplan F., « dhSegment : A generic deep-learning approach for document segmentation », *International Conference Frontiers in Handwriting Recognition*, 2018.

Arora S., Yingyu L., Tengyu M., « A Simple but Tough to Beat Baseline for Sentence Embeddings », *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, p. 1-11, 2017.

Beeferman D., Berger A., Lafferty J., « Statistical Models for Text Segmentation », *Mach. Learn.*, vol. 34, n° 1-3, p. 177-210, February, 1999.

Bluche T., Hamel S., Kermorvant C., Puigcerver J., Stutzmann D., Toselli A. H., Vidal E., « Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript

Collection in the HIMANIS Project », *International Conference on Document Analysis and Recognition*, 2017a.

Bluche T., Kermorvant C., Ney H., *How to design deep neural networks for handwriting recognition*, 2017b.

Boillet M., Bonhomme M.-L., Stutzmann D., Kermorvant C., « HORAE : an annotated dataset of books of hours », *International Workshop on Historical Document Imaging and Processing*, 2019.

Choi F. Y. Y., « Advances in Domain Independent Linear Text Segmentation », *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 26-33, 2000.

De Hamel C., *A history of illuminated manuscripts*, 2nd ed. rev., enl. and with new ill edn, Phaidon P., London, 1994.

Diem M., Kleber F., Fiel S., Grüning T., Gatos B., « cBAD : ICDAR2017 Competition on Baseline Detection », *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

Drigsdahl E., *Late Medieval and Renaissance Illuminated Manuscripts - Books of Hours 1300-1530*, 2013.

Eisenstein J., « Hierarchical Text Segmentation from Multi-Scale Lexical Cohesion », *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, p. 353-361, 2009.

Fauconnier J.-P., Sorin L., Kamel M., Mojahid M., Aussenac-Gilles N., « Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux », *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, Association pour le Traitement Automatique des Langues, Marseille, France, p. 340-351, July, 2014.

Ferret O., « Using Collocations for Topic Segmentation and Link Detection », *19th International Conference on Computational Linguistics, COLING*, 2002.

Ferret O., Grau B., Masson N., « t : Two Methods for Two Kinds of Texts », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 392-396, 1998.

Galley M., McKeown K., Fosler-Lussier E., Jing H., « Discourse Segmentation of Multi-party Conversation », *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, p. 562-569, 2003.

Grosz B. J., Sidner C. L., « Attention, Intentions, and the Structure of Discourse », *Computational Linguistics*, vol. 12, n° 3, p. 175-204, 1986.

Grüning T., Leifert G., Strauß T., Labahn R., « A Two-Stage Method for Text Line Detection in Historical Documents », 2018.

Hearst M. A., « MULTI-PARAGRAPH SEGMENTATION EXPOSITORY TEXT », *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Las Cruces, New Mexico, USA, p. 9-16, June, 1994.

Hearst M. A., « TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages », *Comput. Linguist.*, vol. 23, n° 1, p. 33-64, March, 1997.

Hernault H., Bollegala D., Ishizuka M., « A Sequential Model for Discourse Segmentation », *Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings*, p. 315-326, 2010.

Hsueh P.-y., Moore J. D., Renals S., « Automatic Segmentation of Multiparty Dialogue », *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Joty S., Carenini G., Ng R. T., « CODRA : A Novel Discriminative Framework for Rhetorical Analysis », *Computational Linguistics*, vol. 41, n° 3, p. 385-435, 2015.

Koshorek O., Cohen A., Mor N., Rotman M., Berant J., « Text Segmentation as a Supervised Learning Task », *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, p. 469-473, June, 2018.

Kozima H., « Text Segmentation Based on Similarity Between Words », *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 286-288, 1993.

Lang E., Puigcerver J., Toselli A. H., Vidal E., « Probabilistic Indexing and Search for Information Extraction on Handwritten German Parish Records », *International Conference on Frontiers in Handwriting Recognition*, 2018.

Lebigue J.-B., *Initiation aux manuscrits liturgiques*, 2007.

Leroquais V., *Les Livres d'heures manuscrits de la Bibliothèque nationale*, [s. n.], Paris, 1927.

Li J., Sun A., Joty S., « SegBot : A Generic Neural Text Segmentation Model with Pointer Network », *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, IJCAI-ECAI-2018, Stockholm, Sweden, p. xx - xx, July, 2018.

Malioutov I., Barzilay R., « Minimum Cut Model for Spoken Lecture Segmentation », *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, 2006.

Morris J., Hirst G., « Lexical Cohesion Computed by Thesaural Relations As an Indicator of the Structure of Text », *Comput. Linguist.*, vol. 17, n° 1, p. 21-48, March, 1991.

Moysset B., Kermorvant C., Wolf C., « Learning to detect, localize and recognize many text objects in document images from few examples », *International Journal on Document Analysis and Recognition*, vol. 21, p. 161-175, 2018.

Nomoto T., Nitta Y., « A Grammatico-statistical Approach to Discourse Partitioning », *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1145-1150, 1994.

Ottosen K., *The responsories and versicles of the Latin office of the dead*, Aarhus university press, Aarhus, 1993.

Ottosen K., « Responsories of the Latin Office of the Dead », 2008.

Peddinti V., Povey D., Khudanpur S., « A time delay neural network architecture for efficient modeling of long temporal contexts », *INTERSPEECH*, 2015.

Pevzner L., Hearst M. A., « A Critique and Improvement of an Evaluation Metric for Text Segmentation », *Comput. Linguist.*, vol. 28, n° 1, p. 19-36, March, 2002.

Plummer J., Clark G. T., « Obsecro Te », *Beyond Use : A Digital Database of Variant Readings In Late Medieval Books of Hours*, 2015.

Renton G., Soullard Y., Chatelain C., Adam S., Kermorvant C., Paquet T., « Fully convolutional network with dilated convolutions for handwritten text line segmentation », *International Journal on Document Analysis and Recognition*, vol. 21, p. 177-186, 2018.

Riedl M., Biemann C., « TopicTiling : A Text Segmentation Algorithm based on LDA », *Proceedings of ACL 2012 Student Research Workshop*, Association for Computational Linguistics, p. 37-42, 2012.

Rosenwein B. H., « Y avait-il un « moi » au haut Moyen Âge ? », *Revue historique*, vol. n 633, n° 1, p. 31-52, 2005.

Rézeau P., *Les prières aux saints en français à la fin du Moyen Age*, Publications romanes et françaises, Droz, Genève, 1983.

Rézeau P., *Répertoire d'incipit des prières françaises à la fin du Moyen âge : addenda et corrigenda aux répertoires de Sonet et Sinclair, nouveaux incipit*, Droz, Genève, 1986.

Sánchez J. A., Romero V., Toselli A. H., Vidal E., « ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset », *International Conference on Frontiers in Handwriting Recognition*, 2016.

Shi Z., Huang M., « A Deep Sequential Model for Discourse Parsing on Multi-Party Dialogues », *AAAI*, 2019.

Sinclair K. V., *Prières en ancien français : nouvelles références, renseignements complémentaires, indications bibliographiques, corrections et tables des articles du "Répertoire" de Sonet*, Archon books, Hamden, 1978.

Sinclair K. V., *French devotional texts of the Middle Ages : a bibliographic manuscript guide*, Westport, 1979.

Sinclair K. V., *French devotional texts of the Middle Ages : a bibliographic manuscript guide. First supplement*, Westport, 1982.

Sinclair K. V., *Prières en ancien français : additions et corrections aux articles 1-2374 du "Répertoire" de Sonet. Supplément*, James Cook Univ. of North Queensland, Townsville, 1987.

Sinclair K. V., *French devotional texts of the Middle Ages : a bibliographic manuscript guide. Second supplement*, New York, 1988.

Skorochod'ko E. F., « Adaptive Method of Automatic Abstracting and Indexing », *Information Processing*, 1972.

Sonet J., *Répertoire d'incipit de prières en ancien français*, n° 54 in *Société de publications romanes et françaises*, Droz, Genève, 1956.

Song Y., Mou L., Yan R., Yi L., Zhu Z., Hu X., Zhang M., « Dialogue Session Segmentation by Embedding-Enhanced TextTiling », *Interspeech*, p. 2706-2710, 09, 2016.

Strauß T., Leifert G., Labahn R., Hodel T., Mühlberger G., « ICFHR2018 Competition on Automated Text Recognition on a READ Dataset », *International Conference on Frontiers in Handwriting Recognition*, 2018.

Sánchez J. A., Romero V., Toselli A. H., Villegas M., Vidal E., « ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset », 2017.

Takanobu R., Huang M., Zhao Z., Li F., Chen H., Zhu X., Nie L., « A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning », *IJCAI-ECAI*, p. 4403-4410, 2018.

Utiyama M., Isahara H., « A Statistical Model for Domain-independent Text Segmentation », *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, p. 499-506, 2001.

Wieck R. S., Poos L. R., Reinburg V., Plummer J. H., Walters art museum, *Time sanctified : the Book of Hours in medieval art and life*, G. Braziller, New York, 1988.

Yaari Y., « Segmentation of Expository Texts by Hierarchical Agglomerative Clustering », *CoRR*, 1997.

Édition comparative intermédiale de séries traductives : exploiter les homologies pour créer des visualisations modulables

Karolina Suchecka* — **Nathalie Gasiglia**** — **Karl Zieger*****

* *ALITHILA (EA 1061), Université de Lille, karolina.suchecka@univ-lille.fr*
** *STL (UMR 8163), Université de Lille, nathalie.gasiglia@univ-lille.fr*
*** *ALITHILA (EA 1061), Université de Lille, karl.zieger@univ-lille.fr*

RÉSUMÉ. *En examinant une série traductive du dixième livre des* Métamorphoses *d'Ovide (mythe d'Orphée et Eurydice), cet article interroge l'apport des logiciels PHŒBUS et MEDITE pour le traitement de textes proches du point de vue du contenu, mais éloignés au niveau de la structure, du lexique et de l'état de la langue. Notre objectif est de prouver que, après quelques enrichissements du corpus et en faisant coopérer ces deux outils, on peut concevoir une visualisation modulable et lisible par tous. Ceci présuppose que PHŒBUS et MEDITE traitent des corpus XML en prenant en compte des éléments ou leurs attributs. En suggérant des pistes d'évolution pour ces logiciels, nous essayons d'engager un dialogue étroit entre les informaticiens qui développent les outils et les chercheurs littéraires qui les exploitent.*

ABSTRACT. *The analysis of Ovid's* Metamorphoses *tenth book (myth of Orpheus and Eurydice) leads us to examine the treatment of texts that are closely related in terms of content, but not on the level of structure, vocabulary and state of language. Our objective is to prove that, after some enrichements to the corpus and by adjusting the tools to cooperate, it is possible to create a modular and readable visualization adapted to all readers. It assumes that these tools treat the XML structured corpus by allowing researches to focus on elements or their attributes. By suggesting evolutionary paths for both softwares, we try to engage a close dialogue between computer scientists who develop the tools and literary researchers who exploit them.*

MOTS-CLÉS : *PHŒBUS, MEDITE, XML:TEI, série traductive, homologies, alignement, visualisation spatialisée, graphe linéaire, visualisation modulable.*

KEYWORDS: *PHŒBUS, MEDITE, XML:TEI, chain of translation, homologies, alignment, spatialized visualization, linear graph, modular visualization.*

1. Introduction

Cette analyse s'inscrit dans le cadre de l'édition comparative intermédiale[1] des réécritures du mythe d'Orphée et Eurydice accompagnée d'une ressource pour le repérage intertextuel, qui a pour objectif de combiner un héritage culturel spécifique avec les technologies les plus récentes afin de permettre une présentation novatrice des œuvres et de leur structure. Dans le cadre de cet article, un échantillon du corpus a été exploité à l'aide des logiciels[2] PHŒBUS[3] et MEDITE[4] afin d'interroger leurs apports et leurs limites pour traiter des textes de contenus relativement proches mais éloignés au niveau de la structure, du lexique et de l'état de la langue. Ainsi, nous examinons une série traductive du dixième livre des *Métamorphoses* d'Ovide qui ouvre le cycle d'Orphée, en analysant un corpus de six traductions[5] : (1) en ancien français, en vers [AFv], cf. Anonyme (1315-1325), et en prose [AFp], cf. Walleys (1493) ; (2) juxtalinéaire, en vers [JXv] et en prose [JXp], cf. de Parnajon (1880) ; (3) littéraire, moderne ou contemporaine (ci-après dites "moderne"), en vers [Mv], cf. Cosnay (2006), et en prose [Mp], cf. Villenave (2003).

Nous procédons d'abord à une recherche d'homologies (de correspondances intertextuelles) avec PHŒBUS afin de visualiser, sous forme de graphes de relations[6], les proximités repérées entre les textes. Nous déterminons ainsi la nature des éléments les plus récurrents afin de détecter les correspondances entre les traductions et d'en proposer une visualisation (linéaire ou spatialisée, qui colore, place et relie les points en fonction des relations repérées). Les lecteurs peuvent alors accéder soit à l'édition d'une œuvre précise, en cliquant sur son nœud, soit à la comparaison des deux textes, en cliquant sur l'arc qui relie leurs nœuds.

Prouvant, avec PHŒBUS, que la série traductive présente bien des homologies, nous souhaitons apprécier plus textuellement les proximités et nous mobilisons pour cela MEDITE, un outil dédié à la comparaison de deux états d'un texte. Il signale les remplacements, les suppressions, les insertions et les déplacements. Son utilité pour la recherche génétique est prouvée[7], mais il nous intéresse de voir ce qu'il propose pour des textes aussi différents que ceux de notre corpus, afin d'apprécier s'il présente des problèmes de détection et de lisibilité qui mériteraient d'être surmontés.

1. Les œuvres intermédiales combinent plusieurs médias (textes, sons, images, etc.), comme le font les bandes dessinées ou les représentations de spectacles vivants.

2. L'équipe ACASA du LIP6, sous la direction de J.-G. Ganascia, les développe en collaboration avec l'ITEM (Institut des textes et manuscrits modernes) et des chercheurs en humanités numériques littéraires du laboratoire d'excellence OBVIL (Observatoire de la vie littéraire).

3. Cf. `http://obvil-dev.paris-sorbonne.fr/phoebus/`, et Boukhaled *et al.* (2015) et Ganascia *et al.* (2014).

4. Cf. `http://obvil.lip6.fr/medite/`, et Ganascia et Bourdaillet (2006), Fenoglio et Ganascia (2008) et Ganascia (2011).

5. Les citations reprennent l'orthographe des textes cités ou des traitements opérés par les outils.

6. Nous mobilisons pour cela l'outil Gephi (§ 4.2.2 à 4.3), cf. `https://gephi.org/`.

7. Cf. projet ANR Phœbus : eBalzac (`https://ebalzac.com/genetique`).

Nous voulons montrer que, avec un corpus enrichi et en faisant coopérer PHŒBUS et MEDITE[8], il est possible de concevoir une visualisation modulable et lisible adaptée à tous (des spécialistes, des élèves francophones ou non, ou des lecteurs passionnés). En effet, nous considérons que les éditions numériques comparatives doivent impérativement apporter des réponses pertinentes aux besoins et aux attentes des différents types de lecteurs et graduer la complexité de ce qui est présenté. Il s'agit non seulement d'exploiter les possibilités offertes par les outils testés, mais aussi de réfléchir aux modalités d'adaptation des contenus enrichis pour une lecture numérique, qui diffère de la lecture papier notamment par sa non-linéarité.

Notre contribution se propose ainsi de démontrer la plus-value qu'aurait l'intégration de traitements, dans PHŒBUS et MEDITE, de corpus balisés en XML:TEI P5 (le standard le plus employé dans les humanités numériques littéraires). Elle ambitionne aussi de montrer l'intérêt d'un dialogue étroit entre les informaticiens qui conçoivent et développent les outils et les chercheurs littéraires (non informaticiens) qui les utilisent. Nous essayons d'engager ce dialogue en observant certaines limites des traitements actuels et en suggérant des pistes d'évolution pour les deux logiciels.

2. Présentation du corpus

Les *Métamorphoses* d'Ovide ne sont pas la première œuvre où apparaît la figure d'Orphée. Nous le trouvons dans les *Argonautiques* d'Apollonios de Rhodes (III^e^ siècle av. J.-C.). L'histoire de son amour avec Eurydice apparaît au livre IV des *Géorgiques* de Virgile (37-30 av. J.-C.), elle est brièvement évoquée dans la narration de l'invention de l'apiculture par Aristée. Enfin, un récit détaillé du mythe figure dans les livres X et XI des *Métamorphoses* d'Ovide. Ce long poème, datant probablement du début du I^er^ siècle, regroupe quinze chants sur le thème des métamorphoses issus des mythologies grecque et romaine. L'histoire d'Orphée débute au moment de son mariage avec Eurydice et finit à la mort du poète. Dans notre corpus, nous nous arrêtons à la fin du livre X, ce qui correspond à un fragment de 85 vers[9].

Dans cet extrait, Orphée convoque Hyménée, le dieu des noces, en Thrace pour célébrer son mariage avec Eurydice, mais le flambeau devant symboliser leur union refuse de s'allumer. Ce mauvais présage pour les futurs mariés se réalise lors d'une promenade : Eurydice meurt d'une morsure de serpent. Inconsolable, Orphée descend aux Enfers pour obtenir que les dieux lui rendent sa femme. Émues par son chant,

8. Nous rejoignons, entre autres, les réflexions de Gallet *et al.* (2016) dans le cadre du projet HyperApollinaire : « Peut-être pourrons-nous concevoir des outils capables [de] reconnaître des formes [de l'intertextualité] élargies, ou plus allusives. Si nous disposions sous forme numérisée de la bibliothèque d'Apollinaire, enrichie de ses lectures connues, aussi érudites qu'erratiques, dans des bibliothèques comme la Mazarine, les outils d'alignement révéleraient sans doute bien des réappropriations, et enrichiraient la liste d'exemples que la mémoire, le flair et la ténacité de grands chercheurs apollinariens ont pu faire émerger ».

9. L'extrait couvre les vers de « *Inde per immensum croceo uelatus amictu* » à « *Aetatis breue uer et primos carpere flores.* »

les âmes qui y subissent des peines éternelles oublient leurs châtiments et Orphée se voit accorder une dernière chance de récupérer sa bien-aimée : ils peuvent sortir des Enfers, à condition qu'Eurydice marche derrière lui et que celui-ci ne se retourne pas pour la regarder. Cependant, juste avant d'atteindre la surface, le poète regarde derrière lui pour s'assurer que sa femme le suit. La condition est ainsi brisée, et Eurydice meurt pour la deuxième fois et retourne aux Enfers. Stupéfait par cette perte, Orphée passe sept jours aux bords du Styx avant de retourner en Thrace. Très courtisé, il refuse ensuite de se lier à une autre femme, mais dirige ses passions vers de jeunes garçons.

Les six textes de notre corpus comportent des modifications d'importances variées. Dans les textes en ancien français, Eurydice meurt en fuyant le dieu champêtre Aristée[10] et l'évocation de deux amants changés en pierre est omise. L'amour homosexuel manque dans les deux traductions juxtalinéaires, qui se terminent au moment de la retraite d'Orphée sur le mont de Rhodope. Eurydice est accompagnée par les Naïades uniquement dans la traduction moderne en vers. En outre, beaucoup de reformulations, de substitutions synonymiques et de métaphorisations sont présentes et, parmi les trois traductions en vers, seule celle en ancien français est rimée.

3. Méthodologie de l'expérimentation

L'analyse simultanée de tous les textes du corpus est compliquée par les graphies des deux textes en ancien français, qui diffèrent des textes postérieurs, mais aussi entre elles : AFv est un peu plus ancien et porte des traces du dialecte normand, alors que la langue d'AFp est plus centralisée.

Au début de notre travail, nous faisons divers choix de transcription[11]. Nous gardons les graphies originales sans moderniser les textes, mais nous introduisons la ponctuation moderne et les majuscules. Les accents ne sont pas restitués, mais les abréviations sont développées. Un prétraitement des textes est effectué ensuite pour moderniser les graphies anciennes et pour lemmatiser chaque mot-occurrence.

Selon les usages actuels de l'édition numérique savante, nous avons initialement structuré notre corpus selon le standard XML:TEI P5, mais les logiciels utilisés imposent de ne pas conserver le balisage. Une première transformation XSL nous a donc permis de restituer les textes avec leurs graphies originales (nous parlons alors improprement de "textes bruts"), une deuxième nous a permis de remplacer les mots par les lemmes des graphies éventuellement modernisées. Enfin, comme la versification peut poser des problèmes d'alignement du fait des sauts de ligne et que nous considérons que sa préservation n'est pas indispensable pour comparer les contenus textuels, nous uniformisons la présentation de tous les textes en mettant une phrase par ligne.

10. Cet épisode est effectivement présent dans le mythe, mais chez Virgile. Pour venger Eurydice, les nymphes font perdre à Aristée toutes ses abeilles. Il ne les récupère que par des sacrifices expiatoires. Cf. Virgile, *Géorgiques*, l. IV, v. 315-558.

11. Nous remercions Matthieu Marchal pour son aide à la transcription des textes médiévaux. Pour les règles générales concernant l'édition des manuscrits, cf. Lepage (2001).

Nous traitons ensuite chaque couple de textes bruts puis lemmatisés. Par cette démarche, nous espérons démontrer la complémentarité du traitement des textes bruts et lemmatisés, le besoin de l'automatisation des traitements multitextes et, à terme, la possibilité de rendre PHŒBUS et MEDITE plus complémentaires et capables de traiter des corpus structurés en XML:TEI (textes et balisages inclus).

Dans la suite de cet article, nous présentons les résultats de notre expérimentation avec PHŒBUS et avec MEDITE, en détaillant leurs fonctionnalités. Notre choix de ces deux logiciels en particulier est motivé par le fait que ce sont des outils mis à disposition des chercheurs littéraires et exploitables par ces derniers, notamment grâce à l'interface graphique. Nous sommes pleinement conscients de l'existence d'outils et de techniques plus récents et, sans doute, plus performants, notamment en ce qui concerne l'alignement de traductions[12]. Toutefois, ils demandent souvent des compétences techniques et informatiques qui les rendent difficilement exploitables par les chercheurs littéraires non informaticiens[13].

4. Établissement des graphes de relations avec le logiciel PHŒBUS

Le logiciel PHŒBUS est conçu pour détecter des réutilisations textuelles (des plagiats aux reformulations), mais nous montrons ici que son exploitation s'avère très fructueuse aussi pour le traitement de séries traductives. Il peut en effet apporter un plus, par rapport à d'autres logiciels d'alignement, en permettant de déterminer la nature des éléments les plus récurrents, et donc de détecter des correspondances[14].

4.1. *Fonctionnement du logiciel*

PHŒBUS permet la comparaison simultanée de deux textes qu'il prétraite grâce à des techniques talistes. Il élimine notamment les mots faibles (*stop-words*), comme les articles, les auxiliaires ou les prépositions, et procède à la racinisation (*stemming*) des mots conservés, à l'aide de l'algorithme Snowball (Porter, 2001 ; Tomlinson, 2004), afin de ne garder que les racines des mots retenus (cf. note 3) ce qui fait, par exemple,

12. L'exploitation d'outils d'alignement des traductions pour traiter des séries traductives a toutefois été critiquée par exemple par Barzilay et McKeown (2001) : « [...] *parallel corpus is far from the clean parallel corpora used in MT. The rendition of a literary text into another language not only includes the translation, but also restructuring of the translation to fit the appropriate literary style* ».

13. Nous renvoyons notamment aux travaux appuyés sur deux logiciels de détection des réutilisations textuelles, TextPAIR (ARTFL Project, Université de Chicago), cf. Horton *et al.* (2010) et Abdul-Rahman *et al.* (2016), et Tracer (Marco Büchler, Georg-August-Universität de Göttingen), cf. Büchler *et al.* (2012) et Franzini *et al.* (2014), et aux études de Ho (2011) et Reboul (2017).

14. Concernant l'intertextualité et les outils numériques, cf. par exemple Coffee *et al.* (2012), Forstall *et al.* (2014) et Ferrero et Simac-Lejeune (2015).

que *aimer*, *aimaient*, *aimerai*, mais aussi *aimant*, sont réduits à la racine *aim*. Ensuite, le logiciel détecte les mots racinisés qui apparaissent dans les deux textes selon les paramètres définis par l'utilisateur. Nous utilisons les paramètres par défaut : le nombre maximal de "trous" autorisés entre deux occurrences et le nombre minimal de mots communs sont fixés à trois (sauf pour les textes en ancien français où nous admettons deux mots communs), et l'ordre des mots n'est pas pris en compte. Ainsi, les reprises et les réécritures moins directes ont les meilleures chances d'être repérées, puisque le logiciel détectera les correspondances même si les temps verbaux ou l'ordre des mots sont différents.

Dans le produit du traitement d'AFp et de Mv (fig. 1), les attributs de l'élément XML <phoebus> indiquent les paramètres du traitement (@gapSize : nombre maximal des trous ; @patternSize : nombre minimal de correspondances à trouver ; @respectWordOrder : respect, ou non, de l'ordre des mots). Chaque correspondance repérée est dans un élément <reuse> porteur d'un identifiant unique (@id). Son fils <pattern> fournit les mots racinisés (stemmes) qui ont permis de repérer la correspondance – ici : *plaindr* (*plaindre* dans les deux textes), *pein* (*peines* dans AFp vs *peine* dans Mv), *derni* (*dernier* dans les deux textes) et *aim (aimer* vs *aimée)*. Les éléments frères postposés permettent de localiser les extraits repérés et de calculer leurs longueurs en nombre de caractères (de <text_1_char_start_index> à <text_2_char_end_index>) et de mots (de <text_1_word_start_index> à <text_2_word_end_index>). Enfin, les deux segments de textes mis en correspondance sont reproduits au sein des éléments <text_1_reuse> et <text_2_reuse>.

```
<phoebus file_1_path="/var/www/obvil/phoebus/tmp/phoebus/TXToZGFsb" file_2_path="/var/
www/obvil/phoebus/tmp/phoebus/TXTI5LFOA" gapSize="3" patternSize="3" respectWordOrder=
"false">
  <reuse id="0">
    <pattern size="4">@plaindr@pein@derni@aim</pattern>
    <correctness>1</correctness><precision>H</precision>
    <text_1_char_start_index>4717</text_1_char_start_index>
    <text_1_char_end_index>4827</text_1_char_end_index>
    <text_2_word_start_index>840</text_2_word_start_index>
    <text_2_word_end_index>872</text_2_word_end_index>
    <text_1_reuse>seconde mort. Mais de lui ne se peut plaindre, fors de trop aimer. Le dernier
    salut luy rendit que a peines</text_1_reuse>
    <text_2_reuse>plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée ?), elle dit un
    dernier Adieu, qu'il peut à peine entendre<text_2_reuse>
  </reuse>
</phoebus>
```

Figure 1. *Extrait du produit (en XML) du traitement par PHŒBUS d'AFp et de Mv*

4.2. *Traitement du corpus brut*

Nous avons d'abord manipulé notre corpus sans le prétraiter (ni modernisation ni lemmatisation). Pour AFv, onze correspondances ont été trouvées avec AFp et une

avec un seul des textes postérieurs, JXv (« doloit pour sa double mort » / « stupéfait de la double mort »). Pour AFp, celles avec les textes juxtalinéaires et modernes sont un peu plus nombreuses, nous en recensons huit : une avec JXp, une avec JXv, deux avec Mv et quatre avec Mp. La comparaison des résultats montre la récurrence d'une correspondance, présentée dans le tableau 1, entre le texte de AFp[15], à gauche, et les autres textes, à droite. Chaque cooccurrence est délimitée par des crochets et suivie par l'identifiant du texte lié en indice. Les stemmes cooccurrents sont soulignés et suivis par les identifiants des textes liés en exposant.

[mains^{AFv} qui retenir la cuid ^{AFv}a. Mais riens ne print fors^{AFv} vent ^{AFv} et ainsi se part^{AFv}it Erudice de son amy et mourut de [[[second$^{AFv\ JXp\ Mp}$e mort^{AFv}. Mais de lui ne se peut plaindre$^{JXv\ JXp\ Mv\ Mp}$,]$_{Mp}$ fors de trop aim$^{JXv\ JXp\ Mv}$er. Le derni^{JXv}er$^{JXp\ Mv}$ salut^{AFv}] $_{JXp}$ luy rendi^{AFv}t que a peine$^{JXv\ Mv}$s l']$_{JXv\ Mv}$ entendi^{AFv}t Orpheus. Forment se plaignoit de la seconde mort de s'amie et voulut retourner pour trouver la morte mais la porte^{AFv} trouva^{AFv} ferm ^{AFv}ee. Et le portier^{AFv} qui la gardoit ^{AFv} lu]$_{AFv}$	**AFv :** « tent ses mains et prendre cuide, mes ne prent fors vent vain et vuide. Cele se part de son mari, qui de seconde mort mori. Mes ne se puet de lui blasmer se ne se plaint de trop amer. Le desrain salut li rendi, que cil a paines entendi. Orpheüs forment se doloit pour sa double mort et voloit retorner pour querre la morte, mes il trouva fermé la porte et le portier qui le gardoit, » **JXv :** « quoi en effet se plaindrait-elle sinon soi avoir été aimée ? Et elle dit pour la dernière fois un adieu, tel que celui-ci pût le recevoir à peine de » **JXp :** « meurt une seconde fois, mais sans se plaindre de son époux ; de quoi en effet se plaindrait-elle sinon d'être aimée ? Elle lui adresse un dernier adieu ») **Mv :** « plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée ?), elle dit un dernier « Adieu », qu'il peut à peine entendre »[16] **Mp :** « Eurydice meurt une seconde fois, mais sans se plaindre ; »[17]

Tableau 1. *Correspondance entre AFp et les autres textes*

La correspondance trouvée entre AFp et AFv est beaucoup plus longue que celles repérées avec les autres textes. Seize stemmes communs ont été trouvés entre eux, dont *second*, qui est aussi commun avec d'autres textes du corpus (JXp, Mp), mais c'est le seul, alors que nous en demandions deux pour qu'une correspondance soit établie.

15. « Orpheus tendit ses mains qui retenir la cuida. Mais riens ne print fors vent et ainsi se partit Erudice de son amy et mourut de seconde mort. Mais de lui ne se peut plaindre, fors de trop aimer. Le dernier salut luy rendit que a peines l'entendit Orpheus. Forment se plaignoit de la seconde mort de s'amie et voulut retourner pour trouver la mort mais la porte trouva fermee. Et le portier qui la gardoit lui retarda son chemin et si lui dist que jamais recouvrer ne la pourroit. »
16. L'autre occurrence est : « touchoit les cordes et de sa bouche se print a chanter telle chançon » / « pleuraient sur lui disant de tels chants, et touchant ses cordes selon ».
17. Les trois autres occurrences sont : « serpent tellement la blessa qu' » / « fleurie, un serpent la blesse au » ; « rive du fleuve infernal fut » / « assis sur la rive infernale, » ; « meschine, fuyant tout amour femenine » / « fuyait les femmes et l'amour ».

Au contraire, les autres correspondances repérées regroupent un nombre restreint de stemmes assez récurrents, qui peut être réduit à quatre lemmes au total : *plaindre, aimer, dernier* et *peine*.

Notons trois difficultés. La racinisation du mot *dernier* et de sa forme féminine *dernière* n'est pas la même (*dernier / derni*), ce qui est problématique, comme la graphie ancienne est un obstacle pour la mise en correspondance de *aimer* (*amer*), *peine* (*paines*) et *dernier* (*desrain*). Nous examinerons l'amélioration des résultats après la modernisation des textes et la lemmatisation (§ 4.3). Par ailleurs, les emplois de synonymes (*blasmer* dans AFv et *plaindre* dans les autres textes) bloquent les repérages de correspondances.

4.2.1. *Analyse des correspondances détectées*

Malgré ces difficultés, les correspondances croisées (où le même fragment d'un texte correspond, au moins partiellement, à d'autres entités du corpus) ont permis d'établir un graphe de visualisation qui donne un premier aperçu des relations repérées entre les textes. Dans ce cadre, nous avons créé deux tables des correspondances : celle des nœuds et celle des liens. La table de nœuds (tab. 2) recense les extraits trouvés les plus complets, leurs @id et les nombres totaux de correspondances trouvées. Les nœuds étant identiques ou plus longs que les segments mis en correspondance dans le tableau 1, nous ne reproduisons pas les extraits qui y sont déjà cités.

@ID	Extraits	Nombres
AFv9	[tab. 1]	17
AFp10	[tab. 1]	30
JXp14	dissipe. Déjà elle [tab. 1] qui parvient à peine à ses oreilles, et elle est de nouveau replongée dans le même gouffre. Orphée, qui voit la mort lui ravir une seconde fois son épouse,	34
JXv12	retirent. Et déjà mourant pour la seconde fois, elle se plaignit en quoi que ce soit de son époux : de [tab. 1] tel que celui-ci pût le recevoir à peine de ses oreilles ; et elle fut replongée de nouveau au même lieu. Orphée resta stupéfait de la double mort de son épouse,	36
Mv11	Mourant une deuxième fois, de son époux elle ne se [tab. 1] et elle roule au lieu où elle était avant. Devant la mort double de sa femme, Orphée resta immobile	24
Mp18	[tab. 1]	7
Mp19	crime de l'avoir trop aimée ! Adieu, lui dit-elle d'une voix faible qui fut à peine entendue	6

Tableau 2. *Table des nœuds*

Les nœuds du tableau 2 correspondent à la totalité des résultats trouvés pour l'épisode de la deuxième mort d'Eurydice dans tous les textes. Remarquons que le nœud AFv9, pour lequel la correspondance de quinze patterns a été trouvée avec AFp10, regroupe également « doloit pour sa double mort », la seule correspondance de AFv avec le texte en français moderne JXv (JXv12, deux patterns) : la taille de nœud est donc la

somme des patterns pour ces deux correspondances (dix-sept). Les correspondances de la plus petite taille ont été trouvées pour Mp avec deux extraits pour un total de vingt-six mots :

> Le malheureux Orphée lui tend les bras, il veut se jeter dans les siens : il n'embrasse qu'une vapeur légère. [Eurydice meurt une seconde fois, mais sans se plaindre ;]$_{Mp18}$ et quelle plainte eût-elle pu former ? Était-ce pour Orphée un [crime de l'avoir trop aimée ! Adieu, lui dit-elle d'une voix faible qui fut à peine entendue]$_{Mp19}$; et elle rentre dans les abîmes.

La métaphorisation et la synonymisation, présentes dans cet extrait, empêchent le repérage de correspondances plus étendues. Les patterns repérés sont les mêmes que dans les autres textes (*mourir, seconde, plaindre, fois, mais* pour Mp18 et *adieu, aimée, peine* pour Mp19), mais les réécritures poétisées (« les abîmes du trépas ») et les reformulations synonymiques (« deux fois ravie » au lieu de « double mort », « seconde mort », etc.), empêchent la réunion des deux correspondances courtes en une longue, du moins en continuant à limiter la taille des "trous" à trois mots.

Nous remarquons également que la correspondance trouvée entre AFp et Mv (« plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée ?), elle dit un dernier « Adieu », qu'il peut à peine entendre ») fait partie d'un nœud plus large, Mv11, pour lequel vingt-quatre patterns ont été trouvés au total. Ces relations spécifiques sont décrites dans la table des liens (tab. 3), où sont renseignés les nœuds sources et cibles, les lemmes des mots[18] qui ont permis d'identifier les correspondances et leurs nombres.

Nœuds sources	**Nœuds cibles**	**Lemmes des mots qui ont permis d'identifier les correspondances**	**Nombres**
AFv9	AFp10	mains, cuider, fors, vent, second, mort, plaindre, salut, rendre, entendre, porte, trouver, fermer, portier, garder	15
AFp10	Mp18	second, mais, plaindre	3
JXp14	Mp18	mourir, seconde, fois, plaindre	4
JXv12	Mv11	époux, quoi, plaindre, dernier, adieu, peine, lieu, Orphée, rester	9
JXv12	JXp14	déjà, seconde, époux, quoi, effet, plaindre, sinon, aimée, dernier, adieu, peine, oreille, replonger, nouveau, même, Orphée, mort, épouse	18
AFv9	JXv12	double, mort	2
AFp10	JXp14	seconde, plaindre, aimer, dernier	4
Mv11	Mp19	aimer, adieu, peine	3
AFp10	JXv12	plaindre, peine, dernier, aimer	4
AFp10	Mv11	plaindre, peine, dernier, aimer	4
JXp14	Mv11	époux, plaindre, peine, dernier, aimer, quoi, adieu, être	8
JXv12	Mp19	adieu, aimer, peine	3

Tableau 3. *Table des liens*

18. Nous reconstituons les stemmes après la racinisation et les lemmatisons afin d'améliorer la lisibilité des résultats.

Nous pouvons constater que quatre correspondances sont recensées pour le nœud Mv11 : avec JXv12 (neuf mots communs), Mp19 (trois), AFp10 (quatre) et JXp14 (huit). Plus généralement, pour les sept nœuds présentés dans le tableau 2, nous trouvons douze liens au total, dont le plus grand compte dix-huit stemmes (JXv12 / JXp14). Parmi les trente-huit lemmes recensés, nous relevons six occurrences de *plaindre* et d'*aimer*, cinq de *dernier*, *second* et *peine* (qui fait partie de la locution adverbiale *à peine*), quatre de *adieu* (contre une seulement de *salut*), trois de *mort*, etc. L'entité nommée *Orphée* n'a été détectée que pour deux comparaisons (JXv12 / Mv11 et JXv12 / JXp14). Simultanément, elle est absente de la correspondance JXp14 / Mv11, de taille assez importante (huit patterns), pour laquelle la correspondance a été établie sur *époux*. Nous nous pencherons sur la question des entités nommées pour l'alignement et la comparaison de notre corpus au § 6.

4.2.2. *Visualisation spatialisée*

Les tableaux 2 et 3, fournis à un logiciel de visualisation de graphes comme Gephi, permettent de générer différents schémas. Nous proposons d'abord une représentation spatialisée, particulièrement utile pour des analyses détaillées de correspondances croisées. Elle montre notamment les groupes des nœuds les plus larges, que nous plaçons aux périphéries du graphe (fig. 2) afin d'optimiser la visualisation et de faciliter l'analyse détaillée de chaque groupe. Une couleur spécifique est attribuée à chaque texte afin de permettre un aperçu général de la structure des correspondances détectées. Beaucoup de relations ont été repérées uniquement entre deux textes : il s'agit majoritairement des correspondances AFv (bleu) / AFp (vert), mais nous en recensons également deux Mv (rose) / JXp (orange), une JXv (noir) / Mp (violet) et une AFv / JXv. L'épaisseur des liens est, quant à elle, dépendante du nombre de patterns impliqués dans la mise en correspondance. Six graphes ont une taille supérieure à quatre nœuds, mais celui qui concerne l'épisode de la deuxième mort d'Eurydice (fig. 3) est le seul qui regroupe tous les textes du corpus. Chaque nœud établit entre deux (Mp18, Mp19, AFv9) et cinq (AFp10) liens. Curieusement, pour ceux de Mp, les deux nœuds présents ne correspondent pas aux mêmes textes : Mp18 établit des liens avec AFp et JXp, et Mp19 avec JXv et Mv. AFv9 est lié uniquement à JXv et AFp, mais ce dernier lien est d'une taille très importante (quinze patterns communs).

Le graphe le plus large compte onze nœuds : il s'agit du chant d'Orphée implorant les dieux des Enfers de lui rendre Eurydice (fig. 4). Parmi ces nœuds, quatre appartiennent à Mp, trois à JXp et deux à JXv et Mv, aucune correspondance n'a donc permis d'alignement avec les textes en ancien français. Quatorze liens au total sont établis entre les nœuds : moins que pour le graphe de l'épilogue (fig. 5), qui en compte quinze (mais seulement neuf nœuds) et qui intègre également deux extraits de AFp. Le troisième graphe le plus large correspond à la réaction des âmes résidant aux Enfers au chant d'Orphée (fig. 6), il compte neuf nœuds et quatorze liens.

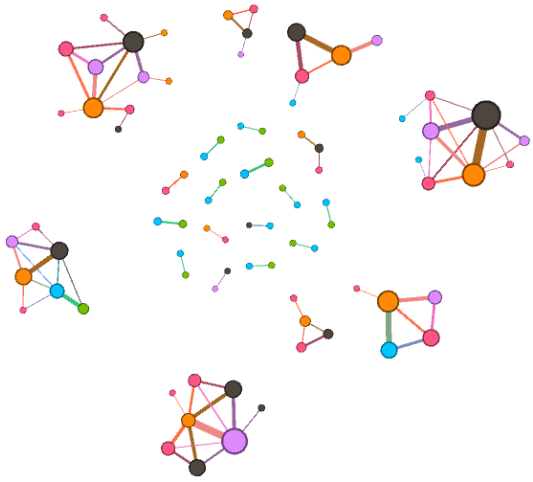

Figure 2. *Visualisation spatialisée de la totalité des correspondances (AFv : vert ; AFp : bleu ; JXv : noir ; JXp : orange ; Mv : violet ; Mp : rose)*

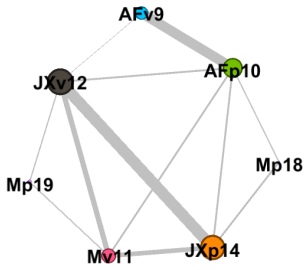

Figure 3. *Graphe de l'épisode de la deuxième mort d'Eurydice*

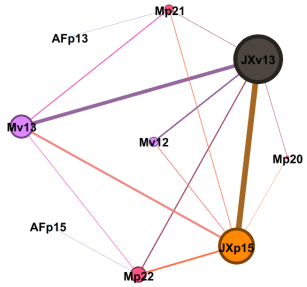

Figure 5. *Graphe de l'épilogue*

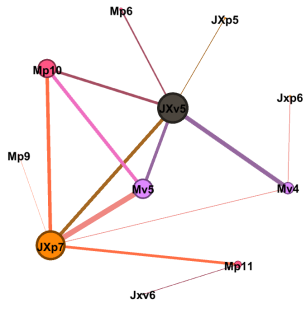

Figure 4. *Graphe du chant d'Orphée*

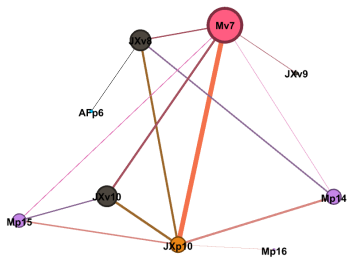

Figure 6. *Graphe de la réaction des âmes des Enfers au chant d'Orphée*

La visualisation spatialisée nous semble propice à l'analyse orientée vers un épisode ou un champ lexical en particulier, en permettant l'observation de représentations des relations détectées. Une édition numérique proposant ce type de visualisation permet de restreindre les recherches aux lemmes spécifiques ou à la taille des graphes et de se focaliser sur un texte précis. En pointant le curseur sur un nœud, le texte correspondant s'affiche dans une infobulle et, en pointant sur le lien, on accède à la liste des correspondances. Un clic sur un nœud ou sur un lien permet de visualiser la comparaison détaillée entre deux correspondances, ce que nous présentons au § 7.

4.2.3. *Visualisation linéaire*

Une autre visualisation, linéaire, nous semble plus adaptée à l'analyse généralisée de l'évolution des correspondances au fil du texte. Sa lisibilité (fig. 7) n'est pas aussi optimale que celle de la visualisation spatialisée, notamment là où les relations deviennent complexes. Cependant, elle offre une vue globale des relations entre les textes : le lecteur peut constater plus aisément que très peu de relations ont été détectées entre les textes en ancien français et les textes modernes ou que l'alignement de toute une partie du chant d'Orphée n'a pas été fait entre AFv et AFp (repère [B]).

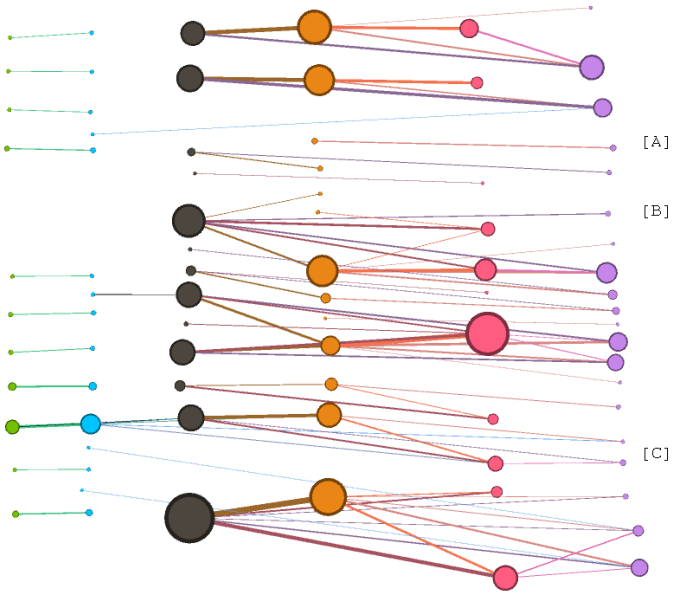

Figure 7. *Graphe linéaire*

Très peu de correspondances ont également été détectées entre le mariage d'Orphée et Eurydice et le début du chant ([A]). Si elles sont là, elles concernent plutôt des couples de textes, une relation plus complexe n'est pas observable (comme elle l'est

par exemple au niveau du [C]). Notons également qu'il y a extrêmement peu de problèmes d'alignement : les résultats proposés par PHŒBUS renvoient bien aux mêmes épisodes.

Nous reviendrons sur l'analyse de ces visualisations après avoir vu ce que fournit le logiciel MEDITE (§ 5), mais avant, confrontons les résultats de PHŒBUS pour le corpus brut avec ceux obtenus en traitant le corpus modernisé, pour les textes en ancien français, et lemmatisé.

4.3. *Traitement du corpus modernisé et lemmatisé*

Sans nous attarder sur la modernisation, notons que la lemmatisation des mots[19] a été effectuée avec TreeTagger[20] et revue manuellement. PHŒBUS a ensuite opéré avec les mêmes paramètres que précédemment (fig. 1). Comme attendu, les repérages des correspondances se sont nettement améliorés pour les textes en ancien français, tant en ce qui concerne le nombre des nœuds, que leur taille et le nombre des liens avec les textes plus modernes. Sur quatorze nœuds détectés pour AFv, seulement cinq liens l'ont été uniquement avec AFp, sans aucune possibilité d'alignement avec les textes postérieurs. Pour AFp (dix nœuds), une relation binaire a été détectée en plus avec Mp (« moult de lui se plaindre. Orphée être celui qui premier apprendre celui » / « soupirer ; tout se plaindre de son refus. mais ce être lui qui, par son exemple, apprendre au » [lem.]). Cette seule occurrence est d'ailleurs très significative, puisqu'elle ne concerne pas uniquement les questions de réécriture, mais également celles de l'adaptation du texte : il s'agit de la mention finale de l'amour homosexuel, absente des traductions juxtalinéaires. Mv l'aborde de manière très euphémique (« Chez les peuples thraces, il fut l'auteur de ceci : transférer l'amour sur les tendres garçons et cueillir l'avant de la jeunesse, le printemps bref, les premières fleurs… »). Dans AFv, nous observons une formulation plus descriptive (« Ce fu cil qui premierement aprist ceulz de Trace a retraire d'amour femeline et a faire des joennes malles lor deduit, dont or sont cil de Trace tuit. »).

En appliquant exactement la même démarche que dans le traitement du texte brut pour la constitution des nœuds généraux de la totalité des correspondances trouvées, nous arrivons à un nombre de nœuds très limité, mais de taille extrêmement importante. Quarante-trois nœuds au total ont été constitués (quatre-vingt-cinq pour le corpus brut) : quatorze pour AFv, dix pour AFp, trois pour JXv et JXp, deux seulement pour Mv et dix pour Mp. Quasiment la totalité des textes est recensée et le nœud le plus grand (JXv2) compte 318 patterns communs (249 pour JXp3 et 245 pour Mv2). La visualisation spatialisée, contrairement à celle du texte brut, permet de constater que

19. « ils appellent Eurydice. Elle se tenait parmi les ombres nouvellement arrivées » → « il appeler Eurydice. elle se tenir parmi le ombre nouvellement arriver » [lem.]. Par « [lem.] », nous signalons que le texte cité a été lemmatisé.
20. Cf. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

nous obtenons uniquement une grande galaxie de relations extrêmement complexes (fig. 8).

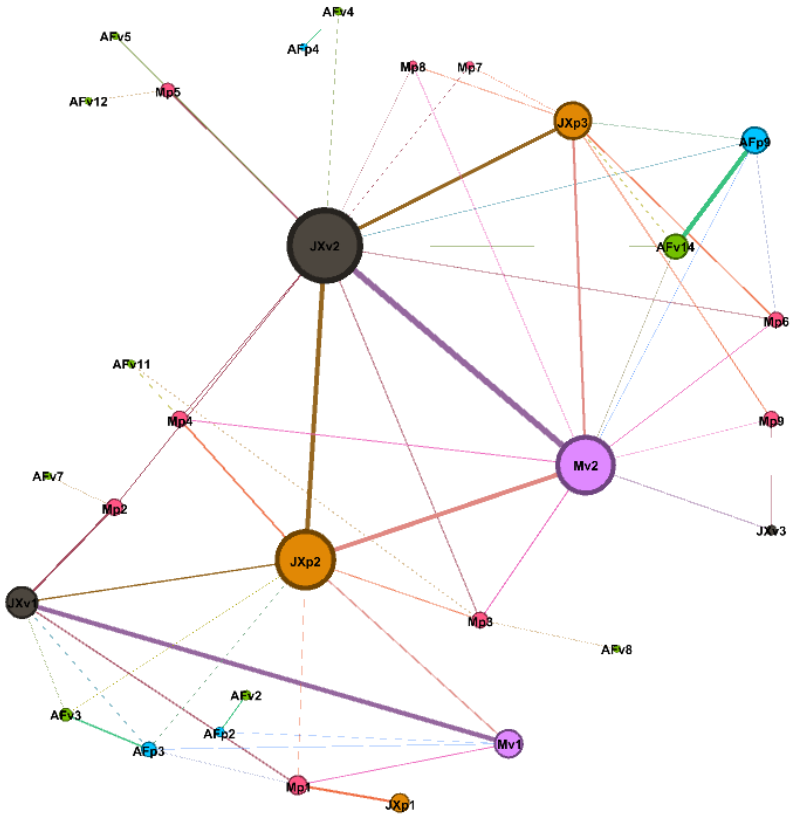

Figure 8. *Visualisation spatialisée du corpus modernisé et lemmatisé*

Cela prouve que le traitement des textes lemmatisés améliore la détection des correspondances et pas seulement pour les textes en ancien français. Toutefois, avec des nœuds devenus parfois très larges, le traitement détaillé des correspondances est moins lisible : nous voyons que JXv2 est lié avec JXp2, ce qui est très pertinent, mais il l'est aussi avec JXp3, ce qui l'est moins. Le moment exact où la correspondance a été trouvée et le nombre d'extraits détectés ne sont pas visibles, puisque la taille du nœud correspond à une autre relation très importante qui a été détectée entre JXv2 et Mv2 et qui regroupe plus des deux tiers du texte (du commencement du chant d'Orphée jusqu'à la seconde mort d'Eurydice). L'analyse semble être plus aisée pour les textes plus fragmentés, comme AFv, AFp et Mp. Simultanément, le nombre de résultats qui ne correspondent pas aux mêmes fragments du récit, qui sont donc de fausses correspon-

dances, augmente de manière significative, surtout pour les petites correspondances de trois patterns. Ainsi, pour la comparaison AFv / JXp, seulement trois résultats sur six ont été pertinents pour notre étude[21].

Au contraire, la visualisation linéaire (fig. 9) semble gagner en lisibilité et fournir davantage d'informations que lors du traitement du texte brut.

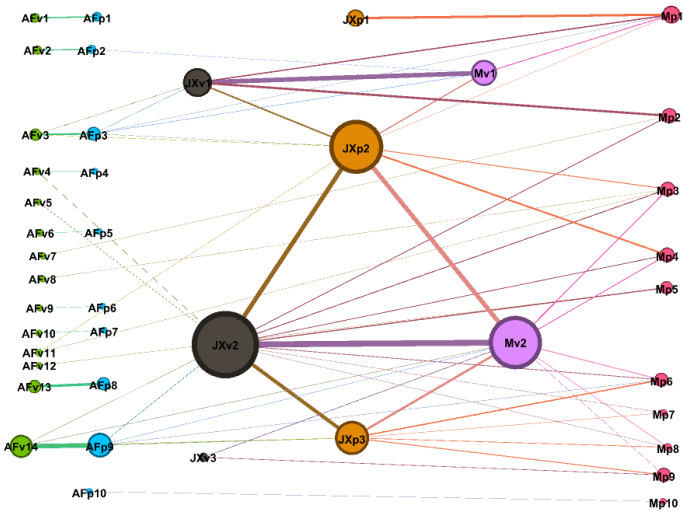

Figure 9. *Visualisation linéaire du corpus modernisé et lemmatisé*

Si des relations binaires restent présentes entre AFv et AFp, elles deviennent beaucoup plus significatives. Ainsi, nous constatons que le texte correspondant à AFv1 et AFp1 est absent des autres traductions, puisqu'il correspond au résumé de la métamorphose précédente (« dessus avoir oïr le fable comment Yphis fille fils devenir et prendre femme. » / « dessus avoir oïr comment Yphis devenir un beau jouvenceau et comment il épouser Yenta son ami. » [lem.]). AFv n'est plus presque uniquement en relation avec sa version en prose, mais il établit également des liens assez fréquents avec les textes modernes, notamment Mp, qui permet ensuite son alignement avec les textes plus éloignés (principalement Mv, pour lequel deux correspondances ont été trouvées, et dont une seulement a été retenue). Mv est le texte qui compte le moins de nœuds : alors que nous y trouvons la quasi-totalité du texte, la taille de Mv2 (245) et le nombre des liens qu'il établit avec d'autres nœuds (onze) sont inférieurs notamment à ceux de JXv2, qui regroupe 318 patterns repartis entre douze liens.

21. Aucune correspondance n'a été trouvée en comparant les textes bruts.

4.4. *Conclusions relatives à l'emploi de PHŒBUS*

Malgré la croissance de la fréquence des détections erronées, les différences de résultats du traitement de notre corpus, soit brut racinisé soit lemmatisé, prouvent que la possibilité de moduler le type de données prises en compte (tokens racinisés ou lemmes de formes éventuellement antérieurement modernisées) s'avère utile pour les utilisateurs, tant pour la visualisation que pour l'analyse des correspondances, puisque ce qui est révélé est complémentaire. Un traitement prenant en compte les différents types de données pourrait aboutir à la création de graphes multidimensionnels qui permettraient, par exemple, de basculer d'une vue générale (traitement du texte lemmatisé) à une vue plus détaillée (traitement du texte brut). Notons aussi que ceci aurait une validité comparable pour la détection de liens non avérés entre des textes sans relation évidente, ce qui est la destination première de PHŒBUS.

Cette observation nous conduit à considérer que, si nous posons qu'il faudrait pouvoir traiter plusieurs états des textes alternatifs, il nous faut les faire coexister dans le corpus manipulé plutôt que de décliner le corpus, comme nous l'avons fait ici, en une version brute et une version modernisée et lemmatisée. Cette coexistence est usuelle dans les corpus XML:TEI au sein desquels les mots sont balisés (<w>) et porteurs d'attributs pour les graphies modernisées (@ana) et les lemmes (@lemma).

Ce traitement alternatif pourrait également permettre de lier la détection des homologies avec la comparaison des versions en rendant les logiciels PHŒBUS et MEDITE complémentaires et en ouvrant de nouvelles possibilités d'affichage pour l'édition numérique savante. Mais, pour ce faire, voyons ce que nous permet le logiciel MEDITE.

5. Comparaison modulable avec le logiciel MEDITE

MEDITE est un outil de comparaison des versions d'une œuvre qui puise entre autres dans l'algorithme d'alignement par fragments grâce à la détection des homologies, une méthode de détection des séquences (utilisée initialement pour l'alignement de macromolécules – ADN ou protéines) afin de faire ressortir leurs régions homologues (cf. note 4). Actuellement, il propose une comparaison de deux textes bruts (cf. *supra*) : les blocs communs sont analysés et les différentes variantes sont signalées grâce à des codes de couleur. Les remplacements sont marqués en bleu, les insertions en vert, les suppressions en rouge et les déplacements en gris. Les possibilités de modulation des paramètres sont un peu plus larges que dans PHŒBUS : par défaut, le traitement est sensible à la casse, aux séparateurs et aux signes diacritiques. La longueur minimale des blocs communs est de cinq caractères. Pour que deux variantes soient considérées comme un remplacement, le ratio de la longueur des deux chaînes repérées doit être supérieur ou égal à cinquante pour cent (« abcd » est remplacé par « efgh », mais « a » est supprimé et « efgh » est inséré). Enfin, dans le cas de fortes densités des blocs communs et des variantes, les premiers sont insérés dans la variante si la différence de leur longueur par rapport à la longueur des variantes est supérieure ou égale à cinquante pour cent. Ici, nous modifions uniquement le ratio des rempla-

cements à un pour cent, en considérant que les suppressions et les insertions sont des blocs qui n'ont pas leurs homologues dans l'autre texte. Les autres paramètres gardent les valeurs par défaut.

Très pertinent pour la recherche génétique et l'alignement des différentes versions du même texte et se différenciant des autres outils de ce type notamment par sa capacité à détecter des déplacements, MEDITE est largement exploité dans de nombreux projets de recherche aux objectifs très variés[22].

5.1. *Traitement des textes dits bruts*

Un exemple des résultats de la comparaison pour deux textes détectés par PHŒBUS comme étant très proches, JXv et Mv, permet de constater que, actuellement, la complémentarité des résultats de PHŒBUS et de MEDITE (fig. 10) existe, mais qu'elle n'est pas pleinement perceptible :

élevée à travers son corps ; et non autrement qu'Olénus, qui attira sur lui le crime, et voulut paraître être coupable ; et que toi, malheureuse Léthéa, ayant eu confiance dans ta beauté, cœurs autrefois très unis, maintenant pierres, que l'humide Ida supporte.
Le nocher avait repoussé lui priant, et voulant en vain traverser de nouveau.
Il resta assis cependant sur la rive durant sept jours, sale, sans don de Cérès.
Le souci, et la douleur de son cœur, et ses larmes furent ses aliments.
S'étant plaint les dieux de l'Érèbe être cruels, il se retire sur le haut Rhodope et sur l'Hémus battu par les aquilons.

Et Olenos, qui a pris sur lui un crime et a voulu sembler cruel, et toi, oh si confiante en ta figure, pauvre Léthéia, cœurs autrefois tout unis, maintenant pierres, que l'Ida humide porte.
Orphée supplie en vain, il veut passer encore une fois, le batelier l'écarte.
Pendant sept jours il reste assis sur la rive, sans don de Céres.

L'amour, la douleur de l'âme, les larmes le nourrissent.
Il se plaint que les dieux de l'Érèbe sont cruels,
Il se retrouve en haut du mont Rhodope, et sur l'Hémus battu des vents.
Pour la troisième fois le Titan avait fini l'année, fermée par les Poissons des Eaux et Orphée fuyait Vénus et toute femme, soit parce que les choses choses avaient mal tourné pour lui, soit parce qu'il avait donné sa foi.
Beaucoup avaient l'ardeur de s'unir au poète.
Beaucoup souffrirent d'être repoussées.
Chez les peuples thraces, il fut l'auteur de ceci : transférer l'amour sur les tendres garçons et cueillir l'avant de la jeunesse, le printemps bref, les premières fleurs…

Figure 10. *Comparaison en texte brut de JXv et Mv*

La plupart des variantes détectées ont été analysées comme étant des remplacements. Parmi les quatre suppressions, une seulement (« S'étant ») n'est pas la suite d'un remplacement interrompu par la détection d'un déplacement. Les insertions sont plus nombreuses pour Mv, notamment à cause de l'absence de l'épisode de l'amour

22. Outre les projets de génétique littéraire, comme l'ANR Phœbus : eBalzac ou HyperApollinaire, cf. notes 7 et 8, MEDITE est exploité pour l'enseignement, par exemple dans le cadre du projet ANR JCJC Compétences et difficultés des élèves en matière d'écriture à l'entrée du collège (ECRICOL), Lafont-Terranova *et al.* (2017). Guerry (2018), lui, diversifie ainsi les attendus : à partir des résultats de MEDITE, il élabore une liste des mots-clés présents dans les variantes détectées entre les textes et constate, par exemple, que « [l]'intérêt de MEDITE ne consiste pas tant dans sa capacité de mettre à jour une homologie isolée, qu'à fournir une sorte d'index exhaustif et facilement lisible de toutes les récurrences lexicales entre deux textes ».

homosexuel dans les traductions juxtalinéaires. Toutefois, certains déplacements détectés ailleurs dans JXv viennent interrompre cette insertion, ce qui produit même une erreur de reconstitution du texte en redoublant le mot « choses » (« que les choses choses avaient mal tourné »). Nous n'allons pas nous attarder sur la question des déplacements, dont la détection peut être modulable.

Pour l'extrait présenté à la figure 10, l'alignement des blocs communs ne pose pas de problèmes majeurs (même si l'on peut noter le bloc commun « en vain », qui ne correspond pas au même contexte, « voulant en vain traverser de nouveau » / « supplie en vain »). Il peut y avoir plus de bruits (fausses détections) ou de silences (absences de détections), notamment pour des extraits très différents (fig. 11) :

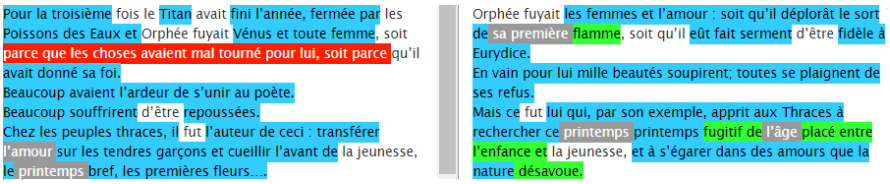

Figure 11. *Comparaison en texte brut Mv / Mp*

Dans la comparaison de Mv et Mp, le bloc commun « d'être » ne correspond pas au même contexte (« il avait donné sa foi » / « il eût fait serment d'être fidèle à Eurydice » et « Beaucoup souffrirent d'être repoussées » / « toutes se plaignent de ses refus »). La répétition de la conjonction « soit » n'a pas été repérée, alors qu'elle permettrait de faire correspondre « parce que les choses avaient mal tourné pour lui » et « qu'il déplorât le sort de sa première flamme ». Enfin, l'accumulation de signalements de remplacements rend les résultats de la comparaison trop généraux.

5.2. *Traitement du corpus modernisé et lemmatisé*

Si MEDITE était capable de prendre en compte des formes lemmatisées, cela permettrait, au choix, de ne pas repérer les différences de flexion ou de conjugaison, ou de n'en sélectionner que certaines (fig. 12) :

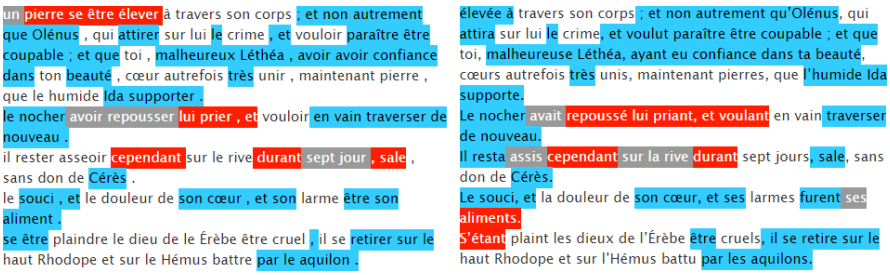

Figure 12. *Résultats pour JXv en format lemmatisé (à gauche) et brut (à droite)*

Pour cet extrait, cinq occurrences de ce type de variantes sont observables : deux occurrences de « vouloir » (« voulut » / « a voulu » et « voulant » / « veut »), une de « rester » (« resta » / « reste »), une de « repousser » (« repoussé » / « repoussées ») et une de « être » (« être » / « sont »). Parmi elles, la première, la troisième et la quatrième pourraient être détectées sans lemmatisation puisque MEDITE peut faire une recherche sur des sous-chaînes de caractères incluses dans les mots plutôt que sur les mots entiers. La deuxième permet d'éviter le seul problème d'alignement pour cet extrait (« en vain »), en séparant les deux parties de la phrase (« le nocher avait repoussé lui priant, et voulant » et « traverser de nouveau »). La cinquième permet d'augmenter la taille des blocs communs repérés (« plaindre le dieu de le Érèbe être cruel » [lem.]). En revanche, « repousser » engendre l'apparition d'un déplacement qui n'est pas avéré (« Le nocher avoir repousser » / « beaucoup souffrir de être repousser » [lem.]).

Pour certaines recherches, les formes actualisées restent significatives et méritent d'être conservées, tandis que, pour d'autres, effectuer la comparaison sur les lemmes serait préférable. La prise en compte des valeurs d'attributs @lemma dans la structure XML permettrait de moduler aisément cet affichage, en traitant soit tous les lemmes soit une partie d'entre eux. En effet, la prise en compte des codes catégoriels et flexionnels (@pos et @msd) rendrait possible une spécification encore plus poussée (comme « remplacement flexionnel passé simple / passé composé »). Enfin, même si la comparaison était effectuée sur les lemmes, leur enregistrement en tant qu'attributs permettrait d'afficher le texte original afin d'améliorer la lisibilité des résultats.

6. Extension des traitements aux entités nommées et autres objets textuels balisables

Pour des textes comparables, mais très différents en surface, comme ceux de notre corpus, il importe de traiter le contenu mais aussi d'autres indices qui peuvent être fournis par un balisage XML idoine et semi-automatisable grâce à des outils de TAL. Pris en compte par PHŒBUS et MEDITE pour leurs traitements, les indices inclus dans le XML permettraient d'intégrer dans la comparaison les entités nommées et leurs périphrases, mais aussi les interprétations du récit source qui affectent tant le contenu que les protagonistes invoqués (comme l'absence de la mention de l'amour homosexuel), la synonymie (« elle fut replongée de nouveau au même lieu » / « elle roule au lieu où elle était avant »), l'hyperonymie (« il fut présent à la vérité » / « il est là »), la métaphorisation (« elle ne put s'animer bien qu'il l'agite » / « le dieu qui l'agite ne peut ranimer ses mourantes clartés »), etc.

Faute de pouvoir tout exposer, observons concrètement un cas d'entité nommée (<persName>) avec deux désignations d'Hyménée (« Hymen » et « diex de noçoiement ») pour lesquelles nous proposons (fig. 13) un balisage XML:TEI qui fournit, en tant qu'attributs, l'identifiant de l'entité nommée (@corresp) et, pour les mots (<w>), les codes grammaticaux (@pos) et flexionnels (@msd), les formes lemmatisées (@lemma), les modernisations (@ana), les champs lexicaux (@corresp) et les synonymes (@sameAs) :

```
<persName corresp="#Hyménée">
  <w pos="NAM" msd="sing" lemma="Hymen" ana="Hymenée" corresp="#déité">Hymen</w>
</persName>
<persName corresp="#Hyménée">
  <w pos="NOM" msd="sing−CS" lemma="dieu" ana="dieu" sameAs="déité divinité
  providence">diex</w>
  <w pos="PRP" lemma="de" ana="de">de</w>
  <w pos="NOM" msd="sing" lemma="noçoiement" ana="noce" corresp="#mariage" sameAs=
  "épousailles mariage">noçoiement</w>
</persName>
```

Figure 13. *Balisage de deux <persName> désignant Hyménée : « Hymen » et « diex de noçoiement »*

Dans l'extrait traité au § 5.2, seuls « Dieux de l'Érèbe », « Rhodope », « Hémus » et quelques pronoms personnels désignant Orphée ont pu être détectés comme blocs communs. Ces correspondances sont encore plus faibles pour les textes en ancien français.

La possibilité offerte par MEDITE d'ignorer les signes diacritiques est suffisante pour traiter certaines variations des désignations d'entités, comme pour le cas de figure « Cérès » / « Céres ». Mais cette fonctionnalité s'avère risquée pour le traitement des textes littéraires, en ignorant, par exemple, les différences entre présent et participe passé ou les lettres logogrammiques, souvent importantes pour l'analyse. Et PHŒBUS ne propose pas de tels traitements. La structuration XML que nous proposons (fig. 14 pour certaines entités de JXv et de Mv) est certes riche, mais en grande partie automatisable en modifiant les sorties de TreeTagger, par exemple. Elle offre l'intérêt de multiplier les indices et donc de permettre la détection des différentes dénominations d'entités.

Ainsi, pour la variante « humide Ida » et « Ida humide », qui présente un simple déplacement de l'adjectif, la proximité des blocs communs et la taille de l'entité font que ce déplacement n'est pas reconnu par MEDITE. Même si, dans ce cas précis, il s'agit effectivement d'un déplacement, généraliser ce traitement produirait des erreurs indésirables, comme pour cette variante entre Mp et Mv : « il ne porte ni visage serein, ni présage heureux » / « il n'apporte ni parole rituelle ni visage heureux » où l'on voudrait plutôt voir un remplacement synonymique (« visage serein / visage heureux »), une variante dénominative (« présage heureux / parole rituelle ») et le déplacement général de tout le syntagme nominal. Alors que ce dernier exemple est très difficilement automatisable, pour « humide Ida », sa présence au sein de l'élément <persName> offrirait un indice supplémentaire autant sur l'intégralité du nom et de son épithète que de la correspondance des formes « humide Ida » et « Ida humide ».

```
<!--Ida humide-->
<persName corresp="#Ida">
  <w pos="NAM" msd="sing" lemma="Ida">
  Ida</w>
  <w pos="ADJ" msd="sing" lemma="humide"
  sameAs="embué mouillé moite">
  humide</w>
</persName>
```

```
<!--humide Ida-->
<persName corresp="#Ida">
  <w pos="ADJ" msd="sing" lemma="humide"
  sameAs="embué mouillé moite">
  humide</w>
  <w pos="NAM" msd="sing" lemma="Ida">
  Ida</w>
</persName>
```

```
<!--Dieux de l'Érèbe-->
<persName corresp="#Charon #Hadès #Perséphone">
  <w pos="NOM" msd="pl" lemma="Dieu" sameAs="déité divinité providence">Dieux</w>
  <w pos="PRP" lemma="de">de</w>
  <w pos="DET:ART" msd="sing" lemma="le">l'</w>
  <persName corresp="#Érèbe">
    <w pos="NAM" msd="sing" lemma="Érèbe">Érèbe</w>
  </persName>
</persName>
```

```
<!--nocher-->
<persName corresp="#Charon">
  <w pos="DET:ART" msd="sing" lemma=
  "le">le</w>
  <w pos="NOM" msd="sing" lemma="nocher"
   sameAs="navigateur nautonier pilote">
  nocher</w>
</persName>
```

```
<!--batelier-->
<persName corresp="#Charon">
  <w pos="DET:ART" msd="sing"  lemma=
  "le">le</w>
  <w pos="NOM" msd="sing" lemma=
  "batelier" sameAs="nautonier passeur
  pilote">batelier</w>
</persName>
```

```
<!--malheureuse Léthéa-->
<persName corresp="#Léthéa">
  <w pos="ADJ" msd="sing fem" lemma=
  "malheureux" sameAs="pauvre funeste
  déplorable">malheureuse</w>
  <w pos="NAM" msd="sing" lemma=
  "Léthéa">Léthéa</w>
</persName>
```

```
<!--pauvre Léthéia-->
<persName corresp="#Léthéa">
  <w pos="ADJ" msd="sing" lemma="pauvre"
  sameAs="misérable piteux
  malheureux">pauvre</w>
  <w pos="NAM" msd="sing" lemma=
  "Léthéia">Léthéia</w>
</persName>
```

Figure 14. *Table d'entités nommées, de JXv à gauche et de Mv à droite, et leur structuration en XML:TEI*

La syntaxe TEI permet de couvrir tous les cas de figure, comme la possibilité d'imbriquer des entités (noms de lieux dans des noms de personnes, par exemple) et de cumuler des identifiants renseignés au sein de l'attribut @corresp : cette nécessité est visible dans l'occurrence « Dieux de l'Érèbe », qui désigne plusieurs entités (Perséphone, Hadès et Charon), tout en renvoyant à l'entité Érèbe. Le balisage des entités nommées permet également de repérer les périphrases et de les lier à l'entité désignée, comme c'est le cas, par exemple, de « rive », qui, tant dans JXv que dans Mv, désigne le Styx, ou « nocher » et « batelier » désignant Charon. Cela permet de lier ces occurrences à celles plus directes présentes dans notre corpus (JXp : « Orphée essaie de fléchir Charon ; vainement il veut traverser de nouveau le Styx »), mais également entre elles, et de spécifier leur nature grâce aux synonymes renseignés (@sameAs). Même

si la relation synonymique n'est pas aussi directe que pour « malheureuse Léthéa » et « pauvre Léthéia », nous retrouvons la correspondance au niveau des synonymes communs (« nautonier » et « pilote »). Simultanément, l'indice supplémentaire fourni par l'encadrement du texte par des éléments <persName> ou <placeName> et l'identification uniformisée grâce à @corresp permettent un affichage modulable, par exemple concentré uniquement sur les entités nommées. En signalant exclusivement les variantes en leur sein, nous pouvons nous permettre une analyse et un étiquetage plus détaillés et effectuer un alignement partiel en faisant abstraction des autres éléments du texte.

Si nous insistions ici sur la plus-value à attendre pour le traitement par MEDITE, il serait tout à fait équivalent (pour des raisons comparables) pour celui par PHŒBUS.

7. Coopération de PHŒBUS et de MEDITE

Bien évidemment, l'intégration de toutes les modalités de traitement évoquées *supra* et considérées comme étant pertinentes pour les recherches avec PHŒBUS et MEDITE serait coûteuse. Toutefois, nous considérons que le traitement de corpus XML, en permettant aux chercheurs de focaliser leur attention sur des sous-arbres et de demander que des calculs soient opérés en leur sein et intégrés aux résultats d'analyse affichés, marquerait un réel progrès.

Une modulation plus poussée du traitement nous semble pouvoir aboutir à des résultats très probants et susceptibles de faire émerger des besoins plus spécifiques, comme la recherche focalisée sur un champ lexical (amour, souffrance, etc.[23]) et la création de graphes multidimensionnels qui combinent plusieurs types de traitements (fig. 15).

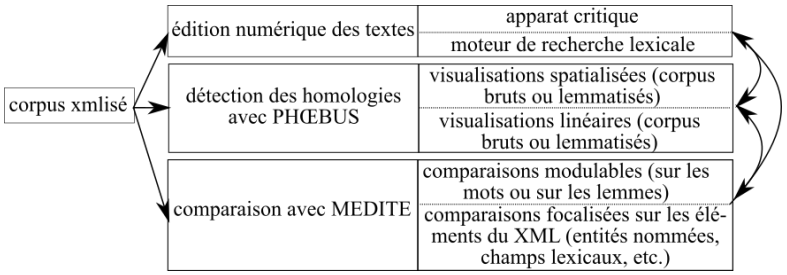

Figure 15. *Schéma éditorial qui prend en compte la coopération entre PHŒBUS et de MEDITE : possibilité d'obtenir plusieurs sorties à partir d'un seul corpus xmlisé.*

Ainsi, la visualisation linéaire des homologies détectées avec PHŒBUS permettrait de cibler les recherches dans les comparaisons textuelles proposées par MEDITE.

23. Cf. fig. 13, où l'appartenance du mot « noçoiement » au champ lexical du mariage est signalée au sein de l'@corresp.

Simultanément, la possibilité de s'appuyer sur une visualisation linéaire pour le travail de comparaison permettrait de localiser plus facilement les extraits qui intéressent le plus le lecteur et de naviguer aisément entre les couples de comparaisons. Par exemple, un clic sur un nœud précis le sélectionnerait comme premier texte et ouvrirait un menu permettant de choisir le texte à lui comparer, puis transporterait le lecteur vers chaque partie intéressante de la comparaison (tout en conservant la possibilité de naviguer dans la totalité des textes). Les visualisations spatialisées (fig. 2 à 6) permettraient, quant à elles, de focaliser l'attention sur un fragment précis pour lequel des homologies ont été détectées entre plusieurs textes. Elles seraient également plus adaptées pour le traitement des corpus sans relation évidente. Dans ce cas de figure, un clic sur l'arête qui lie deux nœuds textuels afficherait leur comparaison avec MEDITE (fig. 16).

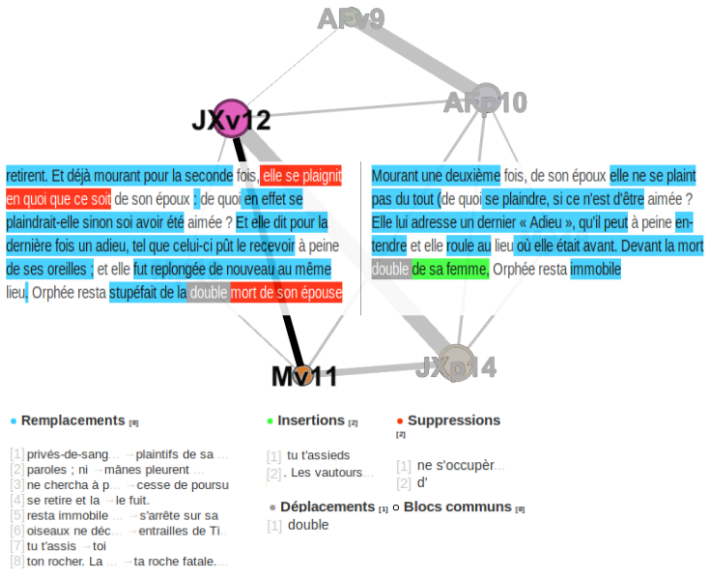

Figure 16. *Exemple d'une coopération possible entre PHŒBUS et MEDITE : le clic sur l'arête qui lie deux nœuds textuels affiche leur comparaison avec MEDITE.*

Une préparation du corpus au préalable par une structuration XML:TEI adaptée à ces besoins spécifiques permettrait d'aboutir à cette liberté de traitement en fonction des contenus de chaque corpus et des objectifs des chercheurs. Elle donnerait la possibilité de lier le traitement de PHŒBUS avec celui de MEDITE et de focaliser la comparaison sur les facteurs communs balisés dans le XML (§ 4 à 6) et des sous-arbres d'éléments (certains syntagmes, les entités nommées, etc.). Pour lier les différents extraits dans les deux types de traitements, chaque mot (<w>) se verrait attribuer un identifiant unique (@id).

8. Conclusion

Notre contribution a cherché à montrer la plus-value de l'évolution proposée des modalités de traitement par PHŒBUS et MEDITE, que ce soit pour la série traductive intermédiale des réécritures du mythe d'Orphée et Eurydice qui nous occupe ou, plus largement, pour l'épanouissement d'autres projets littéraires numériques diversifiés et innovants. Pendant la phase préparatoire de cet article, la participation d'un des coauteurs (durant quelques mois) aux travaux de l'équipe ACASA, a permis d'observer la prise en compte de certaines des observations que nous avons formulées[24]. En remarquant quelques limites des traitements actuels et en suggérant des pistes d'évolution pour les deux logiciels, nous espérons poursuivre ce dialogue étroit entre les chercheurs et les ingénieurs en informatique qui développent les outils et les chercheurs littéraires qui les exploitent.

9. Bibliographie

Abdul-Rahman A., Roe G., Olsen M., Gladstone C., Morrissey R., Cronk N., Chen M., « Constructive visual analytics for text similarity detection », *Computer graphics forum*, vol. 36, n° 1, p. 237-248, février, 2016.

Anonyme, *Ovide moralisé*, Rouen, 1315-1325. Transcrit à partir de Ms. O.4 (fol. 246v-248r). [Attribué de manière incertaine à de Vitry P. et à Legouais Ch.].

Barzilay R., McKeown K. R., « Extracting paraphrases from a parallel corpus », *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Toulouse, France, p. 50-57, 2001.

Büchler M., Crane G., Moritz M., Babeu A., *Increasing recall for text re-use in historical documents to support research in humanities*, Springer Berlin Heidelberg, Berlin, p. 95-100, 2012.

Boukhaled M.-A., Sellami Z., Ganascia J.-G., « Phoebus : un logiciel d'extraction de réutilisations dans des textes littéraires », *22ème conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France, p. 391-396, 2015.

Coffee N., Koenig J.-P., Shakti P., Ossewaarde R., Forstall C., Jacobson S., « Intertextuality in the digital age », *Transactions of the American Philological Association*, vol. 142, n° 2, p. 383-419, septembre, 2012.

Cosnay M., « Ovide, Les Métamorphoses, X », *Musagora*, 2006.

de Parnajon F., « Ovide, choix de Métamorphoses », *Les auteurs latins expliqués d'après une méthode nouvelle par deux traductions françaises*, Hatier, Paris, p. 418-427, 1880.

Del Lungo A., Suchecka K., « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intertextuelles », *DHNord 2019 "Corpus et archives numériques"*, MESH, Lille, octobre, 2019.

Fenoglio I., Ganascia J.-G., « Le logiciel MEDITE : approche comparative de documents de genèse », *L'édition du manuscrit – De l'archive de création au scriptorium électronique*, vol. 10, p. 209-228, 2008.

24. Cf. Ganascia (2019), Del Lungo et Suchecka (2019).

Ferrero J., Simac-Lejeune A., « Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat », *Actes de la 15ème conférence internationale sur l'extraction et la gestion des connaissances*, Luxembourg, p. 287-298, 2015.

Forstall C., Coffee N., Buck T., Roache K., Jacobson S., « Modeling the scholars : Detecting intertextuality through enhanced word-level n-gram matching », *Literary and linguistic computing*, vol. 30, n° 4, p. 503-515, mai, 2014.

Franzini G., Franzini E., Büchler M., Mueller M., Burns P., *Towards a historical text re-use detection*, Springer International Publishing, Suisse, p. 221-238, décembre, 2014.

Gallet O., Michel L., Murat M., Pradeau C., « Apollinaire numérique », *Revue d'histoire littéraire de la France*, vol. 116, n° 3, p. 533-546, 2016.

Ganascia J.-G., « MEDITE – A unilingual text aligner for Humanities. Application to textual genetics and to the edition of text variants », *Supporting Digital Humanities (SDH 2011)*, Copenhague, 2011.

Ganascia J.-G., « Graphes et intertextualité », *Humanités numériques*, Centre Universitaire Méditerrannéen, Nice, septembre, 2019.

Ganascia J.-G., Bourdaillet J., « Alignements unilingues avec MEDITE », *Actes des huitièmes journées internationales d'analyse statistique des données textuelles*, Paris, France, p. 427-437, 2006.

Ganascia J.-G., Glaudes P., Del Lungo A., « Automatic detection of reuses and citations in literary texts », *Digital scholarship in the Humanities*, vol. 29, n° 3, p. 412-421, juin, 2014.

Guerry F.-X., « Góngora et ses premiers biographes : une analyse comparative moyennant des outils numériques », *e-Spania*, 2018.

Ho Y., *Corpus stylistics in principles and practice : A stylistic exploration of John Fowles' The Magus*, Advances in stylistics, Bloomsbury Publishing, 2011.

Horton R., Olsen M., Roe G., « Something borrowed : Sequence alignment and the identification of similar passages in large text collections », *Digital Studies / Le Champ numérique*, 2010.

Lafont-Terranova J., Badin F., Niwese M., Comte E., Chevrot G., Colin D., « Modéliser le processus d'écriture d'un scripteur de haut niveau : intérêt et limites du repérage automatique des opérations de réécriture à l'aide du logiciel MEDITE », MSH Val de la Loire, 2017.

Lepage Y., *Guide de l'édition de textes en ancien français*, Champion, Paris, 2001.

Porter M. F., « Snowball : A language for stemming algorithms », *Retrieved March*, 2001.

Reboul M., *Comparaison semi-automatique des traductions en langue française de l'Odyssée d'Homère (1547-1955)*, 2017. Thèse de doctorat en Littérature comparée, Masson, J.-Y. (dir.), Université Paris IV.

Tomlinson S., « Lexical and algorithmic stemming compared for 9 european languages with Hummingbird SearchServer at CLEF 2003 », *in* P. C., G. J., B. M, K. M. (eds), *Comparative evaluation of multilingual information access systems*, p. 286-300, 2004.

Villenave G. T., « Ovide, Les Métamorphoses, X », *Bibliotheca Classica Selecta*, 2003.

Walleys T., *La Bible des poëtes. Métamorphose d'Ovide moralisée par Thomas Walleys et traduite par Colard Mansion*, Paris, 1493. Transcrit à partir de A. Vérard (fol. 107v-109v).

Vector space models of Ancient Greek word meaning, and a case study on Homer

Martina Astrid Rodda[*,**] — **Philomen Probert**[*] — **Barbara McGillivray**[**,***]

[*] *University of Oxford*
(martinaastrid.rodda@jesus.ox.ac.uk; philomen.probert@wolfson.ox.ac.uk)
[**] *The Alan Turing Institute*
[***] *University of Cambridge (bm517@cam.ac.uk)*

ABSTRACT. *Our paper describes the creation and evaluation of a Distributional Semantics model of ancient Greek. We developed a vector space model where every word is represented by a vector which encodes information about its linguistic context(s). We validate different vector space models by testing their output against benchmarks obtained from scholarship from the ancient world, modern lexicography, and an NLP resource. Finally, to show how the model can be applied to a research task, we provide the example of a small-scale study of semantic variation in epic formulae, recurring units with limited linguistic flexibility.*

RÉSUMÉ. *Notre article démontre à la fois la création et l'évaluation d'un modèle de sémantique distributionnelle du grec ancien. Tout d'abord nous avons développé un modèle d'espace vectoriel où chaque mot est représenté par un vecteur qui codifie les informations qui concernent ses contextes linguistiques. Ensuite nous avons validé différents modèles d'espace vectoriel en testant leur output par rapport à des références obtenues à partir de trois sources: un savant de l'Antiquité, la lexicographie moderne et la ressource WordNet. Enfin, en vue de démontrer comment le modèle peut être appliqué à une activité de recherche, nous fournissons une étude de cas, à petite échelle, de la variation sémantique dans les formules épiques, à savoir les unités recurrentes qui ont une flexibilité linguistique limitée.*

KEYWORDS: *Distributional Semantic Models, Diorisis Ancient Greek corpus, vector-space models, evaluation of distributional resources, ancient Greek epic poetry, formulaic language, semantic variation.*

MOTS-CLÉS : *Modèles de sémantique distributionnelle, Diorisis Ancient Greek corpus, espaces vectoriels, évaluation des ressources distributionelles, poésie épique du grec ancien, formules linguistiques, variation sémantique.*

TAL. Volume 60 – n°3/2019, pages 63 à 87

1. Introduction: the broader research question

This paper presents the creation and evaluation of a computational model that was developed in the context of a broader study of linguistic variation in ancient Greek epic formulae. As the requirements of this research question shape the entirety of the study presented here, we shall begin by outlining some details about the archaic Greek epic tradition and its language. Archaic Greek epic is the product of an oral tradition of which the Homeric poems are the main remnant. Poems about heroes and gods were composed orally and performed by skilled singers in front of an audience. A singer's abilities lay equally in their knowledge of the heroic myths and in their ability to use traditional language according to and beyond the expectations of their listeners.[1]

Throughout this tradition, devices were developed for easier composition and understanding; the most important of these is formulaic language. Formulae are recurring linguistic units which allow for limited flexibility in structure and meaning. They range from noun-epithet pairs (*swift-footed Achilles*, *golden Aphrodite*) to more complex phrases, e.g. speech introductions (*and to him/her spoke...*). The latter often include open slots that need to be filled with additional material: a speech introduction will almost inevitably contain a reference to the name of the speaker (*and to him/her spoke golden Aphrodite*) and/or to some attendant circumstances (*and to him/her Aphrodite spoke in reply*).

While these open slots do not come in a fully pre-defined shape, as opposed to the less flexible parts of the formula itself, restrictions on their flexibility still apply: for instance, in the formulaic structure discussed above, it is possible to address someone *in reply* or *in anger* or *with a smile*, but not *with winged words*, a famous phrase that can only be used in combination with other types of speech introduction formulae.[2]

Formulaic behaviour has been compared to that of idioms and other linguistic constructions characterised by limited syntactic flexibility (Kiparsky 1976; Bozzone 2014; Antović and Cánovas 2016): both formulae and multi-word expressions in everyday language are restricted in their patterns of variation and change. Meaning change and flexibility, in particular, play an important role in the evolution of language usage. Research has highlighted how semantic flexibility promotes the productivity of constructions with open slots in modern languages: the range of different meanings a construction can accommodate in its open slots has an effect on whether it survives and spreads through time (Barðdal, 2008; Perek, 2016). As the behaviour of formulae is similar to that of linguistic constructions, we can expect the semantic openness of a formula to also influence the vitality of its usage through time; however, this mechanism has never been studied in early Greek epic until now.

1. Cf. e.g. Foley (1999). For a discussion of formulaic language in a usage-based linguistic perspective, see Kahane (2018).

2. The metre of archaic Greek epic (hexameters) also plays an important role in these restrictions, a circumstance that this paper will mostly ignore due to its focus on semantics.

Our paper lays the groundwork for an approach to ancient Greek epic formulae that can take into account the role of semantic flexibility in their behaviour and evolution. To do so, we present the first computational model which uses Distributional Semantics to assess the scope of linguistic variation in formulaic language. As this paper's main focus is on the technical requirements and optimisation of the model, we will only show a simple example of its application, by sketching out an analysis of the semantic flexibility of a small group of transitive-verb formulae. Further work will explore the links between semantic flexibility and diachronic variation.

1.1. *Motivation and relationship to previous work*

Distributional Semantics offers a quantitative approach to the study of semantic flexibility and represents an especially promising tool in Historical Linguistics, where no native speaker input can be sought. The meaning of a word is defined in a distributional perspective as a function of its collocates in a corpus: words that share a linguistic context are also related in meaning (Harris, 1954; Fabre and Lenci, 2015). Distributional Semantics naturally presents itself as a valuable method to assess the range of meaning flexibility in formulaic language: it provides a way to combine and process a large quantity of information and to make judgements that do not rely on the intuition of the researcher, as well as being quantifiable to a very fine-grained level.[3] In applying Distributional Semantics to the study of ancient languages, we build on previous work on computational semantics in ancient Greek, which applied similar methods to questions such as semantic drift, the semantics of verbs of motion, and polysemy (Rodda *et al.*, 2017; Grewcock, 2018; McGillivray *et al.*, 2019). Our main interest lies in creating a computational method that will carry useful information for scholars in the humanities, beyond the purposes of the current project.

For these reasons, we focus at length on the validation of the model and the fine-tuning of parameters relative to the pre-processing of the input (see below, 2.2). We will also briefly address how the configuration of parameters relates to existing literature on the evaluation of Distributional Semantics Models (DSMs) (Bullinaria and Levy, 2007; Bullinaria and Levy, 2012; Tanguy *et al.*, 2015). However, our priority is not achieving high absolute rates of accuracy compared to the benchmark sets, but establishing which combination of parameters is most accurate and useful for the purposes of this Digital Humanities study. While the fine-tuning of these hyperparameters is not fully portable across corpora, our article also shows a step-by-step breakdown of the evaluation process, which can be applied to different datasets. Moreover, due to the size of the corpus used (the largest freely available lemmatised corpus of ancient Greek literature), the evaluation provided in this article is significant for research on ancient Greek in general.

3. On these requirements in general, see Jenset and McGillivray (2017). On the potential of digital humanities in addressing traditional philological tasks, see Boschetti (2018).

2. Method

2.1. *Corpus*

Data for this study was gathered from the Diorisis Ancient Greek corpus (Vatri and McGillivray, 2018) (`https://www.doi.org/10.6084/m9.figshare.6187256`). This corpus comprises 820 literary Greek texts, spanning chronologically from Homer (8th century BC?) to the 5th century CE, for a total of over 10 million words. The texts are mainly sourced from the Perseus Canonical Greek Literature repository (along with The Little Sailing and Bibliotheca Augustana digital libraries); each text is fully lemmatised using a custom-built dictionary, and PoS-tagged with TreeTagger (Schmid, 1994) trained on the Ancient Greek Dependency Treebank (Celano, 2014) and the Ancient Greek portion of the PROIEL treebank (Haug and Jøhndal, 2008).

The Diorisis corpus is fully lemmatised; all vector space models discussed here were built on the lemmas, not inflected words. We decided to include the entirety of the material available, without chronological filtering. This decision was made both to compensate for the relatively small size of the corpus itself (in comparison with modern language corpora), and on account of the nature of the "gold-standard" datasets that will be introduced in section 3.1.

2.2. *Corpus processing*

Data was extracted through a Python script written for this purpose, available at `https://zenodo.org/badge/latestdoi/174973156`. The script extracts the collocates for each word in the corpus, i.e. the words that occur together with the target word, within a window of co-occurrence defined by the user; for our study we used windows of 1, 5, and 10 words on both sides of the target word (not including the target itself). Context windows are sensitive to sentence boundaries: we only included words in the same sentence as the target word. Sentence boundaries are already encoded in the Diorisis corpus, and are defined according to ancient Greek punctuation rules: a full stop, semicolon (used as a question mark), or middle dot (equivalent to a colon or semicolon) all end a sentence.

A frequency threshold, applied over the whole corpus, can also be set by the user; we built our semantic spaces respectively on words that occur at least 100 times in the corpus, at least 50 times, at least 20 times, and finally with no frequency threshold at all (including everything down to *hapax legomena*). All spaces were filtered for stop-words, i.e. words that perform a primarily syntactic function, co-occurring with other words independently of their semantic properties, and are therefore considered irrelevant to semantic modelling. The stop-words are not used as either targets or contexts.[4] On the basis of their study of large English-language corpora, Bullinaria and

4. The list of stop-words, compiled by Alessandro Vatri based on the Perseus Hopper source, is available at `https://figshare.com/articles/Ancient_Greek_stop_words/9724613`.

Levy (2012) argue that stop-word filtering does not significantly improve or reduce the quality of the results, but it does reduce the dimensionality of the spaces, which is desirable for computational reasons.

The resulting combinations of parameters leads to 12 different semantic spaces to be assessed (3 window sizes x 4 frequency thresholds). It will be interesting, albeit tangential to the scope of the present study, to note how the behaviour of each of these spaces compares to literature on English-language corpora: in particular, Bullinaria and Levy (2007; 2012) argue that the best results are achieved with the smallest possible window size, i.e. 1, and without filtering for frequency. The authors, however, are working on minimally processed corpora (stripped of sentence boundaries and unlemmatised); this is a very different situation from the richly annotated Diorisis corpus.

2.3. *Semantic spaces*

The vector space models used in this article were built using the DISSECT toolkit (Dinu *et al.*, 2013). To use the terminology of Baroni *et al.* (2014), DISSECT is a count model, not a predictive model (neural network model). While predictive models such as word2vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014) perform at least as well or better than count models on larger corpora (Baroni *et al.*, 2014), the very limited size of our ancient Greek corpus (10 million tokens vs. 2.8 billion tokens in Baroni *et al.* [2014], 6 billion tokens in Mikolov *et al.* [2013], 1 to 42 billion tokens in Pennington *et al.* [2014]) would create a problem for a predictive model, as an even smaller training corpus would need to be extracted, dramatically reducing the size of the available material.[5] Comparison with results obtained via word embeddings could be an interesting avenue for future work, but is not within the aims of the present study.

We used Positive Pointwise Mutual Information (PPMI) as our association measure (Evert, 2008), and applied Singular Value Decomposition (SVD) to reduce the resulting matrices to 300 latent dimensions. Dimensionality reduction is widely recognised to improve the accuracy of vector space models by reducing noise and highlighting the contribution of the most significant dimensions (Landauer and Dumais, 1997). While Bullinaria and Levy (2012) show that models with several thousands of dimensions appear to perform better on large, unprocessed corpora, there is no reason to think this should be the case for a small, information-rich corpus such as Diorisis.

One vector space model was generated for each combination of parameters as detailed in 2.2: `w1_t1`, `w1_t20`, `w1_t50`, `w1_t100`, `w5_t1`, `w5_t20`, `w5_t50`, `w5_t100`, `w10_t1`, `w10_t20`, `w10_t50`, `w10_t100`, with `w` referring to the size of the context window on each side of the target word and `t` to the frequency threshold.

5. Pennington *et al.* (2014) provide data on how GloVe performance scales up with corpus size.

3. Evaluation

As opposed to previous contributions on distributional semantics applied to ancient Greek (Rodda *et al.*, 2017; McGillivray *et al.*, 2019), we decided to assess the performance of the vector space models not by manually screening a sample of the results, but by comparing them to lists of synonyms obtained from three separate sources:[6] ancient scholarship (the *Onomasticon* by Julius Pollux, a Greek scholar of the 2nd century CE), modern lexicography (a 19th-century etymological dictionary of Greek), and a Natural Language Processing resource, the Open Ancient Greek WordNet (Boschetti *et al.*, 2016). Each resource is described in detail below.

Using pre-existing resources instead of individuals' *post-hoc* judgements allows us to provide a robust assessment based on independent data. The characteristics of each source limit the absolute rate of accuracy (see 3.2), but they still allow us to compare the relative accuracy of different Distributional Semantics Models. While work on English routinely uses independent benchmarks, such as the TOEFL performance data (Landauer and Dumais, 1997; Bullinaria and Levy, 2007; Bullinaria and Levy, 2012), handcraft thesaurus (Curran, 2004), and various other benchmark resources (Baroni *et al.*, 2014), to show how the ability of a DSM to identify synonyms compares to human performance on the same task, to the best of our knowledge this is the first time that such a comparison with pre-existing data has been attempted for ancient Greek. In particular, it is tricky to come up with existing lists of synonym pairs, as no ready-made resource of the kind used for English is available. Moreover, the lexicographical resources available for ancient Greek are particularly idiosyncratic, due to the lack of direct input from native speakers. This lack can be partially bridged by resorting to lexicographical works from the ancient world; one of these, Pollux' *Onomasticon*, was used in this article. Due to its unusual characteristics for a resource used in computational linguistics, it deserves more space than the other two resources used in this article, and will be discussed at length in section 3.1.1.

All of our sources for comparison have some level of diachronic depth. This provides another reason, in addition to size considerations, to avoid dividing the Diorisis corpus into periods (see above, 2.1). Pollux' *Onomasticon*, composed in the late second century CE, gives us a wealth of information on Greek vocabulary, with particular but by no means exclusive emphasis on words used by classical authors (see further section 3.1). Somewhat similarly, Schmidt's dictionary of synonyms (Schmidt, 1876-1886) privileges classical and Homeric usage but also takes Hellenistic sources into account, while the Open Ancient Greek WordNet includes data from dictionaries based on texts from a wide chronological span. While the level of diachronic depth varies from source to source, none of them provides a synchronic snapshot for any specific period, and there is no reason to try to achieve this in the corpus.

6. While synonymy and semantic relatedness are only partially overlapping concepts, as semantic relatedness has a broader scope (Levy and Goldberg, 2014), our work on lists of synonyms extracted from independent sources matches standard practice for English-language corpora.

3.1. *Gold-standard sets*

3.1.1. *Ancient lexicography*

For any semantic model to be worthwhile, it ought (as suggested already) to make some predictions that can be tested against something other than itself. For example, it might be expected to predict—to some degree of accuracy better than random guessing—whether native speakers will say that a particular pair of words is semantically similar or not. For a modern language, one way to evaluate a semantic space model is therefore to ask native speakers for some input: how well do the model's predictions stand up to testing against their intuitions? As mentioned in section 1.1, we cannot directly involve native speakers in a study of an ancient language. Yet numerous lexicographical works survive from the ancient Greek-speaking world, and it is worth considering what we can learn from these.

In an ideal world we would identify an ancient lexicographer whose goals were somewhat similar to ours: someone interested in telling us which words are semantically very close to one another, which words lie just slightly further away, and so on. A remarkable ancient work that in some respects does just this is the *Onomasticon* of Julius Pollux, who held the Chair of Rhetoric at Athens in the late second century CE (see Bethe [1917]; Dickey [2007, 96]; Vessella [2018, 24–5]).

Pollux announces in the Preface to Book 1 (*Onomasticon* 1.2) that his work will reveal which words are *sunṓnuma* (usually translated "synonyms"), and then adds by way of clarification that *sunṓnuma* are words that can be substituted for one another—a remarkably distributional way of thinking. While the work does much more than tell the reader which words are *sunṓnuma* (it is a work designed to be read for pleasure as well as instruction, full of learned discussions), it is indeed built around lists of *sunṓnuma*, arranged by topic. The lists show that words counting as *sunṓnuma* for Pollux do not necessarily denote the same thing as each other. While some of his lists collect expressions that do denote roughly the same thing (e.g. being bald, 2.25), others collect expressions denoting members of some category (e.g. parts of the body, 2.22–3). Once again Pollux' concept of semantic similarity is broader than traditional concepts of "synonymy" in our own culture. His concept is arguably closer to the one underlying Distributional Semantics, insofar as words denoting different members of some category might be expected to occur in similar contexts.

In the Preface to Book 1 (1.2), Pollux also announces the structure of the work: after starting with words for gods, he says, he will move on to other topics in the order in which they occur to him. The *Onomasticon* is a more carefully planned work than Pollux lets on here,[7] but it is indeed constructed so that one topic keeps leading to another via a chain of associations. Near the beginning of Book 2 (2.8–16), for example, Pollux gives a long list of expressions for male humans, starting with newborn babies and working up to old men. There follows a list of expressions for female

7. See König (2016, 301), and for an overview of the main themes dealt with in each of the ten books, see Bethe (1917, 776–7).

humans, working from young to old again and including some cross references back to the previous list—the expressions for "baby", for example, all turn out to be gender neutral (2.17). From here Pollux moves on to expressions for giving birth (2.19), for arriving at various ages (2.20), and for parts of the body (2.22–3); then words morphologically related to *thríks* "hair" (2.24), expressions for having various kinds of hair (2.25), expressions for being bald (2.25), and so on.

By arranging the work in this way, Pollux seems to suggest that Greek words (and phrases) are semantically linked via a web of closer and more distant associations, so that to follow them up in a linear order we have to choose an arbitrary path through this web. This thought too anticipates some of our own assumptions. At the same time, we should be wary of reading all our assumptions and goals too hastily into Pollux' work. Most importantly, it is not remotely his goal to sample the Greek lexicon in a way that avoids bias on his part. On the contrary, Pollux' work is designed to reflect his own value judgements, both in the topics and in the vocabulary he prioritises.

As regards vocabulary,[8] which is of particular interest to us, in the Preface to Book 4 Pollux explicitly tells his addressee (the emperor or future emperor Commodus) not to be too surprised if he notices that some word has been omitted. "For I might have omitted it even though I knew about it", Pollux says, "because I did not approve of it" (4.2; cf. Mauduit and Moretti [2010, 524]). Pollux was active at the height of the "atticistic movement"—an extraordinary obsession with reviving the Greek of classical Athens ("classical Attic Greek") for elegant writing and speech. A number of overtly prescriptive lexica from this period ("atticistic lexica"), listing approved and unapproved words, survive in their entirety or in part (see e.g. Dickey [2007, 9, 77, 96–9]; Vessella [2018, 12–26]). Pollux' work is often called an "atticistic lexicon" too, and it belongs in this cultural context, although it is more descriptive in its presentation than the others we have (see e.g. Mauduit and Moretti [2010, 523–4]; Tosi [2013, 144–5]; König [2016, 298–9]; Vessella [2018, 24–5]). Pollux generally lists words without explicitly attaching any value judgements, but occasionally he gives a list of words and then comments that certain further words are inelegant, or are suspected of being inelegant. For example, after giving various words for "delay" or "slowness", he proceeds to comment that the further word *hupérthesis* is "suspected of being cheap", and that *straggeíā* is "very bad" (9.137). By implication, the words he includes without any such comment are approved for use. The usage of classical authors of the fifth and fourth centuries BCE, writing in Attic Greek, looms large in his approved vocabulary, although he takes in vocabulary from works composed in other literary varieties of Greek too, and he is also more open to postclassical vocabulary than most atticists (see e.g. Bethe [1917, 778–9]; König [2016, 299, 304, 307–8]).

If we take Pollux' *Onomasticon* as a proxy for native speaker intuitions, then, we do so with a pinch of salt: Pollux lived five or six centuries after the authors whose vocabulary he tended to prioritise, and he drew on earlier lexicographical works as well as his own reading (see e.g. Bethe [1917, 777–8]; Mauduit and Moretti [2010,

8. On the topics Pollux prioritises, see especially König (2016).

532–6]). In addition, his work does not come down to us in its original form: the work we have is in essence an abbreviated version, although it shows signs of later additions as well as abbreviation (see Bethe [1900-1937, vol. 1, v-vii]; Bethe [1917, 776]).

All these caveats must be kept in mind, then, but the *Onomasticon* remains by far the most substantial surviving work of ancient Greek lexicography to be arranged by topic (the version that survives is about 120,000 words in length, and provides information on the semantics and/or morphology of about 16,000 words);[9] as such, we have chosen to explore its use as a source of independent judgements against which to test semantic space models.

In particular, we chose to find out what nouns (if any) Pollux considers most closely related to 32 target nouns, listed (as "headwords") at `https://github.com/alan-turing-institute/ancient-greek-semantic-space/blob/ancient-greek/pollux_lexicon.txt`. We first of all looked up each of these nouns in the physical *Index Glossarum* to Pollux (i.e. an index of words listed or mentioned, usually in lists of *sunónuma*) prepared by Gunnar Andersen and published in the third volume of Bethe's physical edition of the *Onomasticon* (Bethe, 1900–1937, iii. 14–128). From these instances of our words (including inflected forms) we discarded those in which one of our words occurs as part of a multi-word expression; an example is the occurrence of the accusative plural of our word *ómma* "eye" in the expression *epéskhe tà ómmata* "raised its eyes", as part of a list of expressions for what lightning does when it appears (1.117). We also discarded instances in which one of our words is listed together with other words with which it shares a feature of morphology rather than meaning; an example is an instance of our word *ómma* "eye" in a long list of derived nouns with the suffix *-ma* (6.181).[10]

Once we had arrived at a collection of instances of Pollux listing one of our words along with other semantically similar expressions, we then decided where exactly to draw the line between one of Pollux' lists and the next. In some instances this was a straightforward task, but in others some judgement was needed. For example, we

9. See e.g. Dickey (2007, 96); Tosi (2007, 3-6); König (2016, 298, 301-3). The total word count is a rounded version of the one given by the *Thesaurus Linguae Graecae* ("*TLG*", `http://www.tlg.uci.edu`). The estimated number of words forming the object of linguistic analysis is an estimate of the number of entries in the *Index Glossarum* to Bethe's edition of the *Onomasticon* (Bethe, 1900–1937, iii. 14–128); on this index see further below. Many words feature as the object of linguistic analysis more than once, in different connections; our figure of c. 16,000 counts such words only once each.

10. As a check on this method of data collection we re-collected our data from Pollux using simple word searches of the electronic version of Bethe's text of the *Onomasticon* in the *TLG*, for all case- and number-forms of our 32 target words. This process was far more cumbersome than the one based on the *Index Glossarum*, because *TLG* searches turned up many instances of Pollux simply using one of our target words rather than making it the object of his linguistic analysis. There was therefore a much larger number of irrelevant hits to be discarded through careful reading of the passages involved. The results of the two data collection methods were almost identical, but the *TLG* searches added one data point (an instance of the target word *ánemos* "wind", being linked semantically to *pnoé* "blowing", at 2.77).

decided to count the list of expressions for male humans (2.8–16) as one list rather than either (a) dividing it into sub-lists for male humans of various age categories or (b) considering it part of a longer list that also includes the expressions for female humans that follow. While this decision can be justified to some extent from Pollux' presentation (he begins the list of female humans by referring back to the beginning of the list of male humans, as if these two lists are parallel structural units) no hard and fast rules can be given, and different decisions could have been made. Here and in many other places Pollux gives us lists which can be divided up in more than one way, perhaps because he saw that reality too can be divided up in multiple ways.

After making working lists of all words (including words Pollux suggests are inelegant) that Pollux includes in the same immediate list (or lists) as one of our 32 target words, we discarded from these working lists all expressions consisting of more than one word, and all one-word expressions consisting of something other than a noun (for the rationale behind this decision, see section 3.1.2). For these purposes we defined "nouns" as words with an entry in Liddell *et al.* (1996) ("LSJ") treating them as basically nouns. Substantivised adjectives (including substantivised participles) were thus discarded, unless they are given an LSJ entry separate from that of the adjective or (in the case of participles) the verb. A substantivised adjective was defined as having its own LSJ entry if LSJ gives the noun its own lemma in bold type. We also turned words into the form in which they are in fact listed in LSJ; this was normally the nominative singular, but occasionally the nominative plural.

The resulting word lists are pale reflections of Pollux' work, and should not be confused with his work itself. What they give us is some judgements that various nouns are closely related in meaning to particular nouns among the 32 that we took as a starting point. These judgements are derived from Pollux' work, and are independent of the semantic space models that we would like to evaluate.

3.1.2. *Modern lexicography*

The modern lexicographical resource used for comparison is J.H. Schmidt's three-volume *Synonymik der griechischen Sprache* (Schmidt, 1876-1886). This dictionary contains lists of ancient Greek synonyms, organised into 150 lexical areas according to the editor's judgement. The material included, compiled with traditional philological methods, mostly reflects the usage of classical Greek authors (from the 5th and 4th century BCE) and of the Homeric poems. Each list of synonyms is followed by a detailed discussion, highlighting differences in usage and nuance.

Schmidt's sections are not organised by part of speech, nor do they exclusively contain synonyms in the stricter sense of words that can be used interchangeably in the same sentence context, but rather words that are closely related in meaning and/or derive from the same root. So, for instance, section 1 broadly gathers words for "speaking", ranging from verbs meaning "to speak" (with different nuances, from *légein* "to say" to *laleîn* "to chatter") to nouns for "word" and "voice". This organisation provides a fairly good match for the synsets included in Ancient Greek WordNet (see below).

Resource	Date	Lemmas	Creation	Synonymy defined as
Pollux	2nd c. CE	ca. 400	manual	substitution in context
Schmidt	1876–1886	ca. 1250	manual	similar meaning/etymology
AGWN	2016	ca. 22400	automatic	Princeton WordNet synset

Table 1. *Summary of the characteristics of the benchmark resources. For Pollux and Schmidt, the size reported is that of the sample used. The datasets are available at* `https://zenodo.org/badge/latestdoi/174973156`.

As was the case for Pollux's *Onomasticon*, the assessment that follows only considers nouns in each section of Schmidt's dictionary—not adjectives or verbs. This is to obtain a better match for the example analysis in section 4, where only nouns that act as fillers in Homeric formulae will be considered. Due to the purposes of this analysis, we are interested first and foremost in how accurately the DSM will be able to match existing resources in assessing the semantic similarity between nouns, not other parts of speech.

3.1.3. *Ancient Greek WordNet*

The third resource used for the benchmark comparison is Ancient Greek WordNet (Boschetti *et al.*, 2016) (`http://hdl.handle.net/20.500.11752/ILC-56/`), a lexico-semantic resource developed in collaboration between the Institute of Computational Linguistics "Antonio Zampolli" in Pisa, the Perseus Project in Boston, the Open Philology Project in Leipzig and the Alpheios Project in New York. AGWN is a WordNet built by automatically extracting Greek-English pairs from existing Greek-English dictionaries (LSJ, Middle Liddell, Autenrieth) and then linking the English word to its corresponding synset in the Princeton WordNet (Bizzoni *et al.*, 2014).

English, therefore, acts as an intermediate step in the construction of AGWN, which introduces a potentially significant amount of errors, including but not limited to erroneously grouping words that are translated by English homonyms under the same synset, misinterpreting the part of speech of a word (e.g. by considering expressions including the adjective "joint" as synonyms of the noun "joint" in its various possible meanings), etc. While some level of manual correction was performed by the developers to remove the most obvious mismatches (like the ones arising from the introduction of modern semantic areas such as aviation or telecommunications), AGWN still contains a high amount of noise (Bizzoni *et al.*, 2015).

The characteristics of the three gold-standard resources are summarised in table 1.

3.2. *Evaluation method*

In order to assess the way in which different parameters such as size of the context window and frequency threshold for the lemmas included in the semantic spaces

affected the spaces themselves, and in order to gain insights into the linguistic properties of the three lexical resources we relied on for this study, we compared different parameter configurations against the gold-standard sets.

The semantic spaces are defined based on corpus co-occurrence frequency counts, and therefore they offer a direct way to conceptualise geometric distances between lemmas in terms of their distributional features. For example, in the 300-dimensional semantic space obtained with the SVD-based dimensionality reduction on the corpus co-occurrence matrix defined by a context window of size 5 and a frequency threshold of 50, the top 10 neighbours of the lemma *hiketeía* "supplication", excluding *hiketeía* itself, are: *déēsis* "entreaty" (cosine similarity with *hiketeía*: 0.38), *oîktos* "pity" (0.43, *hiketeúō* "to beg" (0.44), *hikesía* "supplication" (0.44), *epikláō* "to bend, move to pity" (0.47), *epḗkoos* "listening" (0.48), *hupereîdon* "looked over" (past tense) (0.48), *aítēsis* "request" (0.49), *mneía* "mention" (0.49), and *liparḗs* "persisting" (0.50). On the other hand, as we have seen in section 3.1, the three lexical resources we considered as gold-standard differ to some extent in the way they group lemmas into groups of semantically related words. In the case of AGWN, these groups contain synonyms – and are therefore called synsets – which are defined based on linguistic and world knowledge rather than directly on corpus data. For example, AGWN records *hiketeía* in the synset 07187638-n glossed as "a humble request for help from someone in authority". This synset contains the following other lemmas, which are all synonyms of "prayer", "plea", and "supplication": *skêpsis*, *paraítēsis*, *próphasis*, *lipárēsis*, *goúnasma*, and *hikesía*. The aim of this evaluation was to find which semantic space model(s) most closely preserve(s) the semantic features displayed in the gold-standard resources, thus highlighting any linguistically-relevant differences.

3.2.1. *Precision and recall*

With the aim of measuring to what extent the corpus-driven distributional definition of a lemma's neighbours matched the definition of synonymy or semantic relatedness from the gold-standard resources, we focussed on the lemmas that appear in the resources and are also listed as top 10 corpus neighbours ("neighboursets") in the semantic spaces; we called these lemmas "shared lemmas". For each shared lemma, we compared its top 10 corpus neighbours[11] with its resource's synonyms,[12] and calculated precision and recall. Precision of a lemma l is defined in terms of the number of corpus neighbours for l which are also in the synset for l, divided by 10 (the number of corpus neighbours at our disposal); precision measures the proportion of

11. The selection of the top 10 corpus neighbours was driven by feasibility considerations. DISSECT offers the option of displaying the top x neighbours for each lemma. Given the considerable size of the output data returned by DISSECT and the high number of parameter combinations, we had to make the decision on x upfront. Choosing different values of x and measuring the effect of this variation could be the focus of further research, and beyond the scope of this study.

12. In the rest of this article, we will use the term "synonym" to broadly refer to any semantically related word that we included in the gold-standard lists, and the term "synset" for the list of such related words associated to a given lemma in the gold-standard resources.

corpus neighbours which are considered "correct" in the gold-standard. On the other hand, recall is the same number divided by the number of elements in the synset of l, and measures how many of the expected synonyms (according to the gold-standard) appear as corpus neighbours.

$$P(l) = \frac{|synset(l) \cap neighbourset(l)|}{|neighbourset(l)|} \quad [1]$$

$$R(l) = \frac{|synset(l) \cap neighbourset(l)|}{|synset(l)|} \quad [2]$$

Together, these complementary measures give us a complete picture of the extent to which the two sets of resources overlap in terms of their content, thus allowing us to gain linguistic insights into their similarities and differences.[13] In the example for *hiketeía* illustrated above, the two sets both contain the lemma *hikesía*; therefore precision is 1/10=0.1 and recall is 1/7 = 0.14.

The fourth and fifth columns of table 2 report the mean precision and mean recall across all shared lemmas by combination of parameters for the semantic spaces, namely size of context window, frequency threshold, and gold-standard resource.[14] The sixth column of table 2 shows the range (minimum and maximum) of values corresponding to the number of overlapping lemmas between corpus neighbour-sets and resources' synsets. The last column shows the number of shared lemmas between the semantic spaces and each of the resources.

For the majority of lemmas, there is no overlap between the corpus neighbour-sets and the resources' synsets. Consequently, mean precision and recall have low values, ranging between 0.02 and 0.10, and between 0.02 and 0.09, respectively. We interpret this as being due to the fundamental difference between corpus neighbours, which tend to be lemmas that behave in a distributionally similar way, and the synonyms recorded in the resources, which reflect the author's knowledge and intuition about the lemmas' semantic properties and usage.

The gold-standard resource whose synonyms most closely match the corpus neighbours according to our definitions of mean precision is Pollux (with the combination of context window 5 and frequency threshold 20), which achieves the highest value for mean precision of 0.10. The highest mean recall value, 0.09, is reached by Schmidt (context window 1 and frequency threshold 100). In general, we can say that the semantic spaces perform better when compared with both the resource from ancient lex-

13. These measures, however, only take into account whether synonyms appear in the top 10 neighbour lists rather than on their numeric distances in the semantic spaces. The latter will be the focus of the rank-based measures described in section 3.2.2.

14. Because the computational cost of performing the evaluation on AGWN for the semantic space `w1_t1` (context window 1 and frequency threshold 1) was excessively high due to the high number of AGWN synonym pairs, results for this space are not reported and are left to future research.

Res.	W	Freq	Avg P	Avg R	Range	Coverage
AGWN	1	20	0.02	0.02	[0,5]	6864
	1	50	0.03	0.02	[0,5]	4666
	1	100	0.03	0.02	[0,4]	3329
	5	20	0.03	0.02	[0,5]	6865
	5	50	0.03	0.03	[0,5]	4666
	5	100	0.03	0.02	[0,4]	3329
	10	20	0.03	0.02	[0,6]	6865
	10	50	0.03	0.02	[0,5]	4666
	10	100	0.03	0.02	[0,4]	3329
POLLUX	1	1	0.04	0.02	[0,3]	309
	1	20	0.07	0.04	[0,5]	236
	1	50	0.07	0.05	[0,4]	177
	1	100	0.08	0.05	[0,4]	146
	5	1	0.08	0.04	[0,6]	313
	5	20	**0.10**	0.05	[0,6]	236
	5	50	0.09	0.05	[0,4]	177
	5	100	0.09	0.05	[0,4]	146
	10	1	0.07	0.03	[0,6]	313
	10	20	0.08	0.04	[0,5]	236
	10	50	0.08	0.04	[0,5]	177
	10	100	0.08	0.04	[0,4]	146
SCHMIDT	1	1	0.03	0.03	[0,3]	1029
	1	20	0.06	0.07	[0,4]	701
	1	50	0.08	0.08	[0,4]	531
	1	100	0.08	**0.09**	[0,4]	423
	5	1	0.05	0.04	[0,4]	1046
	5	20	0.07	0.07	[0,4]	701
	5	50	0.08	0.08	[0,5]	531
	5	100	0.07	0.08	[0,4]	423
	10	1	0.04	0.04	[0,5]	1046
	10	20	0.06	0.06	[0,5]	701
	10	50	0.07	0.07	[0,5]	531
	10	100	0.06	0.06	[0,4]	423

Table 2. *Values of three evaluation metrics calculated between the semantic spaces with different parameter combinations and the three gold-standard resources. Column 1 contains the name of the resource under consideration: Ancient Greek WordNet (AGWN), Schmidt, and Pollux. Columns 2 and 3 show the size of the context window and the frequency threshold used to define the semantic spaces, respectively. Columns 4 and 5 show mean precision and mean recall across all shared lemmas, calculated based on the overlapping lemmas between corpus neighbour-sets and resources' synsets. Column 6 shows the range of values corresponding to the number of overlapping lemmas between corpus neighbour-sets and resources' synsets, and column 7 shows the number of shared lemmas between the semantic spaces and the resources. See text for a detailed explanation.*

icography (Pollux) and the one from modern lexicography (Schmidt) than when compared with the computational resource (AGWN). Moreover, to the extent that there is a difference, given a fixed context window, frequency threshold levels of 20 or 50 tend to yield higher mean precision and recall values than lower and higher levels.

3.2.1.1. Precision and recall by frequency and polysemy class

When selecting the content to analyse in Pollux, we focussed on 32 lemmas shared by the semantic space with a frequency threshold of 50, AGWN, and Schmidt. The full list of lemmas was provided in section 3.1. Three of these words are not present in the `t_100` semantic space. We categorised these 32 words into two frequency classes ("frequent" and "infrequent") and two polysemy classes ("polysemous" and "monosemous"). The former categorisation was based on the Diorisis corpus frequency counts by setting a threshold corresponding to the average frequency among words that occur at least 50 times, i.e. 548.86.[15] The latter categorisation was based on the number of sense labels in the TLG's online version of LSJ (`http://stephanus.tlg.uci.edu/lsj/`): words with more than 3 senses (as indicated by Roman numerals) and words with more than one sense marked with a Latin letter or with two separate dictionary entries (homonyms) were all marked as polysemous. The 32 words were chosen so that each came from a different synset in Schmidt. We extracted synsets with a random number generator, then chose one word per synset with the required combination of traits. This categorisation allowed us to analyse the possible effects of frequency and polysemy on the precision and recall metrics across the gold-standard resources and according to different semantic space parameters.

We collected a dataset containing, for each of the 32 categorised lemmas, its frequency and polysemy class, its precision and recall measures for each of the combination of parameters available for the semantic spaces (context window and frequency threshold) and for each of the three gold-standard resources. We then fitted a linear regression model that predicts recall based on the best subset of all possible predictors. We selected the best model according to the lowest Akaike Information Criterion (Akaike, 1974) using the Stepwise Algorithm (Hastie and Pregibon, 1992). We obtained the model with the following predictors:

– precision;

– polysemy: values "TRUE" or "FALSE";

– resource: values "AGWN", "Pollux", or "Schmidt".

The model's diagnostic checks make us confident that it fits the data reasonably well. The model's R^2, i.e. proportion of variation in the response by it, is 0.60 and its adjusted R^2 (i.e. corrected for the number of predictors included in the model) is also 0.60. The scatter plot of standardised predicted values versus standardised residuals

15. We used the average frequency rather than the median (141) as the latter, being very low, would have lead us to select words that do not appear in Schmidt.

Predictor	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.006976	0.004405	-1.584	0.1137 *
precision	0.843005	0.027876	30.242	< 2e-16***
resourcePOLLUX	-0.048343	0.010777	-4.486	8.52e-06 ***
resourceSCHMIDT	0.038958	0.004509	8.641	< 2e-16 ***
polysemousTRUE	-0.011308	0.004486	-2.521	0.0119 *

Table 3. *Summary of linear regression model predicting recall of a lemma based on its precision, its frequency and polysemy class, and the gold-standard resource. Significance codes: "***" means significant at the 0.001 level, "**" at the 0.01 level, and "*" significant at the 0.05 level.*

shows that the dataset meets the assumptions of homogeneity of variance and linearity, and the residuals are approximately normally distributed.

Table 3 reports the summary of the model in terms of the estimated coefficient for each predictor's value, the standard error of this estimate, the t value, and the associated t-statistic and p-values. The predictors that are significant at the 0.05 level are indicated by the presence of asterisks in the table. Higher precision values lead to higher recall values: we can interpret the coefficient 0.84 as the average effect on recall of a one-unit *increase* in precision, holding all other predictors fixed. Although this result is not completely surprising, because precision and recall are calculated based on the same numerator (i.e. the number of overlapping synonyms), precision was included in the list of predictors because this led to a model that fitted the data well, and therefore could be interpreted in a meaningful way. Moreover, this allows us to ascertain the relative effect that linguistically-interesting predictors such as "polysemy" and "resource", compared to precision, have on recall.

Regarding the effect of the gold-standard resources, compared to the reference level (AGWN),[16] comparing the corpus-driven semantic spaces to Schmidt *increases* recall by 0.04, while using Pollux has the effect of *decreasing* recall by 0.05. Coming to the features of the lemmas, we can see that the coefficient relative to polysemy is negative, which means that if a lemma is categorised as polysemous, this leads to a *decrease* in recall of approximately 0.01, which is expected given that polysemous words present more challenges in such semantic similarity tasks.

3.2.2. *Rank-based evaluation*

In addition to calculating precision and recall metrics, which compare synsets and neighbour-sets in terms of their content as discrete groups, we devised two other evaluation methods, which take into account a graded measure of semantic relatedness. The aim in this case is to assess to what extent the relationship of semantic related-

16. Because the reference level for the "resource" variable is AGWN, this does not appear in the list of table 3.

ness recorded in the gold-standard resources is reflected in the semantic spaces by considering the ranking of synonyms in the different resources.

For each resource we defined a square co-occurrence matrix based on the distribution of synonyms across the resource's synsets. The rows and columns of the matrix correspond to the lemmas shared between the resource and the semantic spaces, in each of their parameter configurations. For example, in Schmidt (1876-1886) the lemmas *epiméleia* "care, concern" and *mérimna* "care, thought" occur together in one synonym set, the 86th one. Therefore, the entry for *epiméleia* in the co-occurrence matrix for Schmidt has value 1 in the cell corresponding to the column for *mérimna*. Similarly, in AGWN these two lemmas occur together in the synset 00267522-n and therefore in the co-occurrence matrix for AGWN the entry for *epiméleia* has value 1 in the cell corresponding to the column for *mérimna*, too. Following this approach, we can define a vector for each shared lemma and therefore calculate cosine similarity measures between pairs of shared lemmas in the resources' spaces. The co-occurrence matrix defined this way allows us to measure the distance (which is 1 minus the similarity score) between two lemmas according to a metric based on such resource-specific semantic relatedness. For example, the cosine similarity between the vector for *epiméleia* and the vector for *mérimna* in the space defined by the semantic relatedness from Schmidt's resource is 1, as these two lemmas co-occur in exactly the same synsets.

For each pair of synonyms in the lexicons, we compared their resource-based distance to their corpus-based distance. In the example for *epiméleia* and *mérimna*, their cosine similarity in the semantic space with context window 5 and frequency threshold 50 is 1-0.67, which is a consequence of the fact that these lemmas co-occur with different sets of lemmas in the corpus, and therefore their distributions are not identical, as in the case of the space defined from Schmidt's resource.

For each parameter configuration for the semantic spaces and for each of the three lexical resources, we calculated the Average Inverse Rank (InvR), an information-retrieval measure which has been used in other relevant studies, such as Henestroza Anguiano and Denis (2011). For each lemma l, we considered its top 10 corpus neighbours[17] and ranked them by decreasing corpus-based distance. All neighbours appearing in a synset of l in the lexical resource were considered as relevant. InvR is then the average, among all shared lemmas, of the sum of inverse ranks of relevant neighbours:

$$InvR = \sum_{n} \frac{1}{rank(n)} \quad [3]$$

For example, for the lemma *stráteuma* "expedition; army", two corpus neighbours are also considered semantically related to it by Pollux: *stratópedon* "camp; army" and *stratiőtēs* "soldier". Their positions in the ranking based on decreasing corpus

17. As explained previously, the choice of 10 as the number of corpus neighbours considered had the aim of keeping the calculations computationally manageable.

distance are 2 and 4, respectively, which leads to the inverse ranks of 0.5 and 0.25, respectively. Therefore, for *stráteuma*, the sum of inverse ranks is 0.50+0.25 = 0.75.

The fourth column of table 4 shows the average InvR for each configuration. This ranges from 0.36 and 0.54, which are comparable to the results on similar analyses for a modern language like French (Henestroza Anguiano and Denis, 2011). The maximum values are reached for Pollux (context window 1 and frequency threshold 100) and Schmidt (context 5 and frequency threshold 50).

We also ran a Spearman's correlation test on the following two distributions: the distribution of resource-based cosine similarity between pairs of shared synonyms and the distribution of corpus-based cosine similarity between the same pairs. The fifth column in table 4 reports the values of the coefficients of Spearman's correlation tests which returned significant results (α = 0.05). The coefficients indicate a very weak or weak positive relationship between the two distributions, ranging between 0.05 and 0.23. This is not completely unexpected, as this second rank-based measure imposes a stricter condition compared to InvR and the corpus-based distance is defined in a qualitatively different way from the resource-based distances. In spite of the weakness of this relationship, however, the Spearman's correlation coefficients are useful to identify the relative differences between the resources. The highest value is reached by the comparison between Pollux and the semantic space with context window 5 and frequency thresholds 50 and 100. As in the case of precision and recall measures, we notice that the semantic spaces perform better when compared with non-computational resources than when compared with AGWN according to both the rank-based measures.

4. Sketch of an application: the flexibility of Homeric formulae

This section will show the research potential of the distributional approach by applying it to a small-scale study of semantic flexibility in early Greek epic formulae. We will explore the behaviour of two verb phrase formulae denoting, respectively, holding and thinking: formulae of the type *(en/metà) khersìn ékhein*, "to hold [x] in one's hands" vs. formulae of the type *eû eidénai*, "to have [x] in mind". These two formulae were chosen for two reasons: (1) as they both contain a transitive verb, they involve an open slot that is essentially always filled; (2) they are the two most frequent formulae with characteristic (1) based on a formula search on the Chicago Homer (`http://homer.library.northwestern.edu/`).[18]

For both formulae, we considered their object fillers (the [x] variable in the above translation); we aim to highlight how the semantic coverage of possible objects of these formulae influences their behaviour. This will show the usefulness of the DSM to classicists, as well as provide further insights on its performance. While a sample size of 2 formulae obviously will not allow us to reach meaningful conclusions about

18. Excluding speech introduction formulae.

Resource	Window	Freq	Average IR	Spearman's corr.
AGWN	1	20	0.46	0.05
	1	50	0.48	0.06
	1	100	0.49	0.07
	5	20	0.45	0.06
	5	50	0.36	0.07
	5	100	0.46	0.08
	10	20	0.45	0.06
	10	50	0.46	0.07
	10	100	0.45	0.07
POLLUX	1	1	0.44	0.08
	1	20	0.43	0.14
	1	50	0.50	0.18
	1	100	**0.54**	0.18
	5	1	0.48	0.17
	5	20	0.49	0.20
	5	50	0.49	**0.23**
	5	100	0.46	**0.23**
	10	1	0.52	0.17
	10	20	0.48	0.19
	10	50	0.47	0.22
	10	100	0.44	0.21
SCHMIDT	1	1	0.49	0.05
	1	20	0.50	0.08
	1	50	0.52	0.09
	1	100	0.51	0.10
	5	1	0.44	0.07
	5	20	0.49	0.09
	5	50	**0.54**	0.09
	5	100	0.50	0.10
	10	1	0.42	0.07
	10	20	0.47	0.08
	10	50	0.47	0.09
	10	100	0.47	0.09

Table 4. *Values of rank-based evaluation metrics calculated between the semantic spaces with different parameter combinations and the three gold-standard resources. Column 1 contains the name of the resource under consideration: Ancient Greek WordNet (AGWN), Schmidt, and Pollux. Columns 2 and 3 show the size of the context window and the frequency threshold used to defined the semantic spaces, respectively. Column 4 shows the average inverse rank calculated based on the ranks of corpus neighbours which appear in the corresponding resource's synsets. Column 5 shows the correlation coefficient when a Spearman's correlation test run on the distributions of distances between pairs of shared synonyms (calculated in the spaces defined by the resources) and the distances between the same pairs of synonyms (calculated in the corpus semantic spaces) returned a significant result (significance threshold: 0.05).*

formulae in general, this is not the goal of the present use case; we merely aim to show how our distributional model can be applied in a "pocket-sized" version of what will be its ultimate application in further work.

The two formulae chosen for this study differ not only in their meaning (while both denoting a form of having/holding, literal or metaphorical), but also in their syntactic behaviour and usage. The formula for physically holding, "having in one's hands", only takes a direct object in the accusative, as opposed to the one for thinking/mentally holding, "having in mind", which can be construed with either the genitive, mostly used with nouns, or the accusative, mostly with pronouns such as *hóde* or *hoûtos*. (Cases in which the object is represented by a clause were excluded.) The latter rule, however, admits a few exceptions, where content words are construed in the accusative; this happens twice in the *Iliad* and four times in the *Odyssey* (*Il.* 13.665, 20.213; *Od.* 9.215, 11.442, 11.445, 14.365), a pattern that may hint at the accusative with content words being a more recent development. While pronouns like *hóde* or *hoûtos* have been filtered as stopwords, the above-mentioned cases of accusative with nouns and two cases with *pâs* and *hápas* "all, everything" are included. Therefore syntactic flexibility can be considered a relevant parameter.

Other notable differences concern frequency and attestation: the hands-formula occurs 59 times with an object, in both the Homeric poems and in poems from other branches of the tradition (3 times in Hesiod's *Theogony* and 6 in the *Homeric Hymns*), while the know-formula occurs 42 times and is limited to the Homeric epics. If we consider Hesiod and the Hymns to be later than the Homeric epics (Andersen and Haug, 2012), this may be a sign of chronological development; even if, however, we reject this chronological hypothesis,[19] we still know that the formula is productively used outside of Homer, i.e. in other branches of the epic tradition. The fact that the frequency of the two formulae is very similar in the Homeric epics (50 vs. 42 times) makes this unlikely to be due to sparsity of attestation of the know-formula.

Based on these characteristics, we could expect a correlation between semantic flexibility and either syntactic flexibility (hypothesis A, according to which the know-formula should be the most flexible of the two) or productivity of the formulae across traditions (hypothesis B, according to which we would expect higher flexibility in the hand-formula). While these are merely two possible explanatory factors, they serve the purpose of our example case, i.e. showing how quantitative information from the DSM can be used to assess hypotheses about semantic development. The experiment below allows us to test these hypotheses. This cannot, of course, lead to a claim of causality in a study conducted on a sample of two formulae; it can, however, provide a simple example of how the DSM model can be used to analyse formulaic behaviour.

19. For a detailed formulaic analysis that comes to the conclusion that the Hesiodic poems pre-date the Homeric ones, see Pavese and Venti (2000), Pavese and Boschetti (2003).

	Hand-construction	Know-construction
Min.	0.2640	0.3288
Median	0.4515	0.4085
Mean	0.4547	0.4044
Max.	0.6839	0.5404

Table 5. *Summary of the distribution of distances from the fillers of the hand- and know-constructions to their respective centroids in the* `w5_t50` *semantic space.*

4.1. *Choosing a DSM*

In light of the results of the benchmark comparison, the DSM chosen for our experiment is the `w5_t50` one (context window of 5 words on each side and frequency threshold of 50). The `w5` spaces consistently showed the best performance for both precision and recall among the three context windows that were tested (if we exclude one isolated result from the `w1_t100` space compared to Schmidt); as for frequency, while higher thresholds appear to improve accuracy, they also mean potentially being unable to draw conclusions for some of the fillers in our target constructions, as some of these words occur relatively rarely in the corpus. As it stands, out of 40 filler types for the hand-construction and 23 for the know-construction, respectively 10 and 5 are not present in the `t50` space, because of either stop-word filtering or the frequency threshold.

4.2. *Methodology and results*

Fillers for both constructions were extracted by hand through a lemma search on the Perseus Project's Scaife viewer (`https://scaife.perseus.org/`, last accessed March 2019). After extracting the filler words, a Python script (available at `https://zenodo.org/badge/latestdoi/174973156`) was used to compute the centroid of the coordinates of the fillers for each construction, and then measure the cosine similarity of each filler to this centroid. We can then use the radius of the distribution of the fillers as a proxy to its density, giving us a measure of how closely clustered (semantically similar) the fillers themselves are in the semantic space.

Finally, the distances for both constructions were loaded and analysed in R (R Core Team, 2019). Table 5 contains the summary of the two distributions, which were then compared using the Kolmogorov-Smirnoff test. Both the mean and the extremes of the distribution for the hand-construction are higher than for the know-construction; the difference is statistically significant (P = 0.003).

4.3. *Discussion*

As table 5 shows, the hand-formula shows higher semantic flexibility in its open slot. The objects of formulae of the type *(en/metà) khersìn ékhein*, "to hold [x] in one's hands", are more varied in their meaning than the objects of *eû eidénai*, "to have [x] in mind". If we return to our initial hypotheses as outlined before the experiment, we can now conclude that while the higher syntactic flexibility of the know-construction does not correlate with higher semantic flexibility, contrary to hypothesis A, the construction with higher semantic flexibility, i.e. the hand-construction, is the one that is productive outside of Homer, in accordance with hypothesis B.

This experiment, of course, needs to be extended to a much wider network of formulae in order to allow for significant conclusions to be drawn on the behaviour of Homeric formulae; however, this preliminary example shows how the DSM can be used to address the issue of formulaic behaviour from a quantitative standpoint.

5. Conclusions

As stated in section 1, our main objective for this study was to discuss how to build a resource that is actually useful for a DH approach, despite the inevitable limitations. These depend in part on the characteristics of the corpus, which is small in size and, by nature, cannot be expanded: this affects the accuracy of computational tools. They also have to do with the wider context of its use: we cannot gain direct access to native speakers' semantic judgements for ancient Greek, which means having to refer to comparison sources that are each in turn limited and idiosyncratic in their characteristics. The comparison between an ancient lexicographical work and the outcomes of modern computational semantic analysis is a topic that would warrant further study to understand how notions of synonymy and meaning in antiquity compare to modern distributional definitions. The gold-standard resources still allow for a reliable comparison between different DSMs, assessing their relative accuracy. Finally, the use case in section 4 showcases both the limitations and the strengths of the DSM. On the one hand, the proposed use of the semantic model also dictates its characteristics: frequency filtering improves the overall accuracy of the DSM but limits its applicability when dealing with rare words. On the other hand, our pocket-sized experiment shows how the distance measures extracted from the DSM can be used to assess hypotheses about semantic flexibility and formulaic usage from a quantitative perspective, something that has thus far never been attempted for ancient Greek epic.

Finally, this study wishes to highlight the strengths and the potential applications of Distributional Semantics as a resource for research in ancient studies. The most important advantage of this approach, and of computational models in general, is that they allow us to analyse semantic behaviour on a quantitative level, to make detailed and objective comparisons, and to advance quantifiable claims that can in turn be tested. These are all fundamental standards in quantitative historical linguistics (Jenset and McGillivray, 2017), and can all be reached with careful application and evaluation

of computational methods on ancient sources. Further developments in the accuracy of these methods can lead to valuable results for experimental research, an example of which we have outlined by discussing ancient Greek formulae. A more detailed study of this topic is in preparation, and we anticipate that more research on the subject will further strengthen the case for the use of distributional models in philological research.

Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. MAR's stay at The Alan Turing Institute was supported by the 2019/20 Enrichment scheme. All authors designed the study. MAR built the semantic spaces, compiled the lexicon from Schmidt (1876-1886), designed and carried out the analysis for the case study, and wrote all of the sections that are not otherwise credited. PP collected the data from Pollux and wrote section 3.1.1. BMcG designed and implemented the evaluation approach, advised on the creation of the semantic spaces, and wrote section 3.2. All authors gave final approval for publication.

6. References

Akaike H., "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, vol. 19, p. 716-723, 1974.

Andersen Ø., Haug D. T. (eds), *Relative chronology in early Greek epic poetry*, Cambridge University Press, Cambridge, 2012.

Baroni M., Dinu G., Denis P., "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 238-247, 2014.

Barðdal J., *Productivity: Evidence from case and argument structure in Icelandic*, John Benjamins, Amsterdam; Philadelphia, 2008.

Bethe E., *Pollucis Onomasticon*, Teubner, Stuttgart, 1900–1937.

Bethe E., "Iulius (398) Pollux", *in* A. Pauly, G. Wissowa, W. Kroll (eds), *Real-Encyclopädie der classischen Altertumswissenschaft, x.i*, Metzler, p. 773-779, 1917.

Bizzoni Y., Boschetti F., Diakoff H., Del Gratta R., Monachini M., Crane G., "The making of Ancient Greek WordNet", *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, European Language Resources Association (ELRA), Reykjavik, p. 1140-1147, 2014.

Bizzoni Y., Del Gratta R., Boschetti F., Reboul M., "Enhancing the accuracy of Ancient Greek WordNet by multilingual distributional semantics", *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, Accademia University Press, Trento, p. 47-50, 2015.

Boschetti F., *Copisti digitali e filologi computazionali*, CNR Edizioni, Roma, 2018.

Boschetti F., Del Gratta R., Diakoff H., "Open Ancient Greek WordNet 0.5", 2016.

Bullinaria J. A., Levy J. P., "Extracting semantic representations from word co-occurrence statistics: A computational study", *Behavior Research Methods*, vol. 39, p. 510-526, 2007.

Bullinaria J. A., Levy J. P., "Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD", *Behavior Research Methods*, vol. 44, p. 890-907, 2012.

Celano G., "Guidelines for the annotation of the Ancient Greek Dependency Treebank 2.0", 2014.

Curran J. R., *From Distributional to Semantic Similarity*, PhD Diss., Institute for Communicating and Collaborative Systems, School of Informatics, Edinburgh, 2004.

Dickey E., *Ancient Greek Scholarship: a guide to finding, reading, and understanding scholia, commentaries, lexica, and grammatical treatises, from their beginnings to the Byzantine period*, Oxford University Press, Oxford, 2007.

Dinu G., Pham N. T., Baroni M., "DISSECT - DIStributional SEmantics Composition Toolkit", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, p. 31-36, 2013.

Evert S., "Corpora and collocations", *in* A. Lüdeling, M. Kytö (eds), *Corpus linguistics. An international handbook*, Mouton de Gruyter, p. 1212-1248, 2008.

Fabre C., Lenci A., "Distributional Semantics today", *TAL – Traitement Automatique des Langues*, vol. 56, p. 7-20, 2015.

Foley J. M., *Homer's traditional art*, Pennsylvania State University Press, University Park (PA), 1999.

Grewcock R., *Computational semantics and the syntax of motion in Ancient Greek*, MPhil Thesis, University of Cambridge, Cambridge, 2018.

Harris Z. S., "Distributional structure", *Word*, vol. 10, p. 146-162, 1954.

Hastie T. J., Pregibon D., "Generalized linear models", *in* J. M. Chambers, T. J. Hastie (eds), *Statistical Models in S*, Wadsworth & Brooks/Cole, chapter 6, 1992.

Haug D. T., Jøhndal M., "Creating a parallel treebank of the old Indo-European Bible translations", *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, p. 27-34, 2008.

Henestroza Anguiano E., Denis P., "FreDist: Automatic construction of distributional thesauri for French", *TALN - 18e conférence sur le traitement automatique des langues naturelles, Jun 2011*, Montpellier, France, p. 119-124, 2011.

Jenset G. B., McGillivray B., *Quantitative Historical Linguistics. A Corpus Framework*, Oxford University Press, Oxford, 2017.

Kahane A., "The complexity of epic diction", *Yearbook of Ancient Greek Epic Online*, vol. 2, p. 78-117, 2018.

König J., "Re-reading Pollux: encyclopaedic structure and athletic culture in *Onomasticon* Book 3", *Classical Quarterly*, vol. 66, p. 298-315, 2016.

Landauer T. K., Dumais S. T., "A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge", *Psychological Review*, vol. 104, p. 211-240, 1997.

Levy O., Goldberg Y., "Dependency-based word embeddings", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Baltimore, Maryland, p. 302-308, 2014.

Liddell H., Scott R., Jones H., *et al.*, *A Greek-English Lexicon,* 9th edn, 1940, with revised supplement, 1996, Clarendon Press, Oxford, 1996.

Mauduit C., Moretti J.-C., "Pollux, un lexicographe au théâtre", *Revue des études grecques*, vol. 123, p. 521-541, 2010.

McGillivray B., Hengchen S., L̈hteenoja V., Palma M., Vatri A., "A computational approach to lexical polysemy in Ancient Greek", *Digital Scholarship in the Humanities*, 2019.

Mikolov T., Chen K., Corrado G., Dean J., "Efficient estimation of word representations in vector space", 2013.

Pavese C. O., Boschetti F., *A Complete Formular Analysis of the Homeric Poems*, Hakkert, Amsterdam, 2003.

Pavese C. O., Venti P., *A Complete Formular Analysis of the Hesiodic Poems*, Hakkert, Amsterdam, 2000.

Pennington J., Socher R., Manning C., "GloVe: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, p. 1532-1543, 2014.

Perek F., "Using distributional semantics to study syntactic productivity in diachrony: A case study", *Linguistics*, vol. 54, p. 149-188, 2016.

R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. 2019.

Rodda M. A., Senaldi M. S., Lenci A., "Panta Rei: Tracking semantic change with Distributional Semantics in ancient Greek", *Italian Journal of Computational Linguistics*, vol. 3, p. 11-24, 2017.

Schmid H., "Probabilistic Part-of-Speech tagging using decision trees", *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Schmidt J. H., *Synonymik der griechischen Sprache*, Teubner, Leipzig, 1876-1886.

Tanguy L., Sajous F., Hathout N., "Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques", *TAL – Traitement Automatique des Langues*, vol. 56, p. 103-127, 2015.

Tosi R., "Polluce: struttura onomastica e tradizione lessicografica", *in* C. Bearzot, F. Landucci, G. Zecchini (eds), *L'Onomasticon di Giulio Polluce: tra lessicografia e antiquaria*, Vita e Pensiero, p. 3-16, 2007.

Tosi R., "Onomastique et lexicographie: Pollux et Phrynichos", *in* C. Mauduit (ed.), *L'Onomasticon de Pollux: aspects culturels, rhétoriques et lexicographiques*, De Boccard, p. 141-146, 2013.

Vatri A., McGillivray B., "The Diorisis Ancient Greek Corpus", *Research Data Journal for the Humanities and Social Sciences*, 2018.

Vessella C., *Sophisticated speakers: Atticistic pronunciation in the Atticistic lexica*, De Gruyter, Berlin, 2018.

Chronique d'un échec : identification des métaphores dans les écrits des géographes

Suzanne Mpouli*

* *HTL, Université de Paris, CNRS, F-75013 Paris, France*
suzanne.mpouli@u-paris.fr

RÉSUMÉ. *La métaphore présente un intérêt indiscutable pour étudier en diachronie l'évolution des idées dans les textes scientifiques relevant des sciences humaines et sociales. Cependant, malgré les différentes méthodes proposées en pour détecter automatiquement les métaphores, très peu de travaux de recherche ont essayé de les appliquer à ce genre de textes. Dans cet article, nous présentons une tentative d'identification des métaphores conceptuelles dans des textes de géographie en français et en anglais qui utilise une méthode reposant sur LDA (Heintz* et al.*, 2013). Si la méthode testée s'avère, à l'issue de nos expérimentations, inadéquate pour notre objectif final, elle nous a cependant permis de cibler les difficultés inhérentes à ce type de projet ainsi que de futures perspectives de recherche.*

ABSTRACT. *Metaphors are often perceived as being essential to the diachronic study of ideas formulated in scientific texts pertaining to social sciences. However, very few research endeavours have tried to apply any existing NLP metaphor detection method to such texts. The present article describes an attempt to identify conceptual metaphors in geography texts written in English and in French using an LDA-based method (Heintz* et al.*, 2013). Although that particular method was ultimately unsuitable for our final goal, it enabled us to circumscribe the specific challenges inherent to this type of project as well as future research perspectives.*

MOTS-CLÉS : *métaphore, géographie, topic modelling (LDA), fouille de textes, humanités numériques.*

KEYWORDS: *metaphor, geography, topic modelling (LDA), text mining, digital humanities.*

TAL. Volume 60 – n°3/2020, pages 89 à 111

1. Introduction

Aucune figure de style n'a suscité tant de fascination ni fait couler autant d'encre que la métaphore, qualifiée à juste titre de « figure des figures » (Deguy, 1969). Plus qu'un simple ornement du langage, elle devient avec Lakoff et Johnson (1980), le fondement même de notre système de pensée : la métaphore nous permet de mieux décrire et appréhender les multiples phénomènes abstraits qui nous entourent à l'instar des idées, des mouvements et du temps. On peut ainsi, grâce à une métaphore, projeter de manière sélective certains traits d'un *domaine source*, typiquement assez concret sur un *domaine cible*, généralement plus abstrait.

Si la rhétorique étudie principalement la métaphore comme un ornement du langage, elle la différencie en premier lieu des autres figures de style par sa capacité à doter temporairement un terme d'une nouvelle signification. Ainsi, Aristote (1922) la définit comme une figure où il s'opère un « transfert par analogie » d'un mot à un autre tandis que Dumarsais (1818) précise qu'elle s'effectue « en vertu d'une comparaison qui est dans l'esprit ». La métaphore reposant sur une comparaison, elle est généralement, dans les ouvrages de rhétorique, mise en parallèle avec une autre figure de style reposant sur le même procédé : la comparaison figurative. Celle-ci, cependant, au contraire de la métaphore établit explicitement une comparaison entre des unités lexicales au moyen d'un terme de comparaison, traditionnellement « comme » en français. À titre d'illustration, Aristote (1922) oppose la métaphore « **Ce lion** s'élança » à la comparaison « [Achille] s'élança comme **un lion** » et conclut que cette dernière est moins agréable car plus longue. Du point de vue structurel, on a dans les deux phrases un comparé « Achille » et un comparant « lion » ; le comparé étant absent dans la métaphore, on parlera d'une métaphore *in absentia*. Le domaine cible mentionné étant « l'homme » et le domaine source « l'animal », la théorie de la métaphore conceptuelle (Lakoff et Johnson, 1980) classera ces deux phrases sous la métaphore conventionnelle UN HOMME EST UN ANIMAL.

Comme l'exemple de métaphore donné ci-dessus attribué à Homère, bon nombre d'exemples que proposent les rhétoriciens afin de discuter de l'esthétique du langage sont empruntés à des auteurs de textes littéraires, déjà reconnus pour leur maniement décrié ou admiré mais néanmoins toujours singulier de la langue. Commentant l'emploi des mots chez les poètes grecs anciens, Aristote (1922) considère l'usage juste des métaphores comme la qualité primordiale qu'un auteur doit posséder et la seule véritable marque d'un talent indéniable. De par leur pouvoir évocateur et le lien qu'elles entretiennent avec l'imaginaire, les métaphores constituent des outils de choix sur lesquels les auteurs peuvent non seulement innover du point de vue linguistique, marquer l'esprit de leurs lecteurs, mais aussi mettre l'accent sur des thèmes, des émotions ou des événements particuliers de leur récit. Si la présence des métaphores est attendue dans un texte littéraire, leur emploi dans d'autres types de discours, notamment le discours scientifique, a parfois été critiqué.

Cette condamnation de l'utilisation des métaphores dans le discours scientifique découle vraisemblablement à la fois de leur faculté d'ouvrir les portes de l'imaginaire

et de leur statut d'accessoire du langage. En effet, pour Bachelard (1967), par exemple, l'histoire de chaque science se caractérise par un dépouillement de la langue de toute métaphore et analogie, obstacles à l'accès à la vraie connaissance. Néanmoins, dans la pratique, les métaphores restent omniprésentes dans les textes scientifiques dans lesquels elles assurent non seulement un rôle didactique mais aussi créent des cadres de référence quasi universels (Molino, 1979). Ceci se vérifie clairement dans les trois types de fonctions qu'Ascher (2005) distingue dans les métaphores utilisées par les géographes :

– la fonction pédagogique, qui a surtout une valeur illustrative, se borne à constater une ressemblance et n'établit aucun développement analogique plus poussé.
Exemple : L'homme est **un loup** pour l'homme ;

– la fonction heuristique où l'analogie est plus exploitée et facilite l'analyse de phénomènes abstraits en mettant en exergue leurs similitudes avec d'autres phénomènes.
Exemple : Le concept de « modernité **liquide** » chez Bauman qui souligne les changements constants de l'ère actuelle dominée par des technologies à courte durée de vie et toujours en évolution ;

– la fonction modélisatrice où la métaphore pose les bases d'un modèle qui traduit les réflexions théoriques de l'auteur et qui étaye son argumentation.
Exemple : La métaphore textile du réseau social dans laquelle la société est assimilée à un filet constitué de liens.

Hormis les différents rôles qu'elles servent, les métaphores en géographie revêtent un intérêt majeur du point de vue épistémologique : en effet, depuis l'Antiquité, le courant qualitativiste copie l'écriture littéraire, produisant ainsi des textes dans lesquels la métaphore joue un rôle essentiel (Lévy, 2006). Cependant, les divers travaux qui se sont attelés à analyser et catégoriser les métaphores dans le discours des géographie se sont parfois limités à un type spécifique de métaphore, comme la métaphore organiciste (Bachimon, 1979 ; Berdoulay, 1982 ; Archer, 1993), ou alors ont dressé un panorama assez large des métaphores utilisées dans un sous-domaine de la géographie (Daniels et Cosgrove, 1993), souvent sans trop s'attacher à la dimension chronologique et sa signification. De fait, si l'on peut dire avec certitude que le vivant, le théâtre, le textile, la physique, les mathématiques, la mécanique ou encore l'écologie ont servi de domaines sources à des métaphores géographiques, il apparaît plus compliqué de recenser tous les auteurs qui y font référence et tous les termes qu'ils convoquent. Le projet GÉONUM se propose de combler ce vide, d'une part en s'appuyant sur la théorie de la métaphore conceptuelle ainsi que sur des méthodes de traitement automatique des langues et, d'autre part, en se focalisant sur des métaphores dans lesquelles la géographie est utilisée comme domaine cible en conjonction avec un domaine source prédéfini.

Au vu de la nécessité de ne pas se limiter à une structure syntaxique particulière de métaphores (par exemple, métaphores adjectivales ou verbales) et de pouvoir identifier les domaines sources et cible, la méthode de Heintz *et al.* (2013) nous est apparue comme celle qui correspondait le mieux aux besoins de ce projet. Dans la section

suivante, nous présentons en détail les particularités et les différentes étapes de cette méthode. Le code utilisé pour les expériences rapportées dans l'article de référence n'étant pas disponible, nous avons dû entièrement réimplémenter le système quasi à l'identique pour le tester sur des données en anglais. Dans la section 3, nous revenons sur les résultats de cette expérimentation et procédons à une analyse des principales erreurs que nous avons relevées. La section 4, quant à elle, porte sur les modifications apportées à la méthode initiale suite à nos premières expérimentations et sur les résultats obtenus sur des données en français. Enfin, dans la section 5, nous jetons un regard critique sur notre approche avant de conclure.

2. Description de la méthode de détection automatique des métaphores choisie

De manière générale, la métaphore du point de vue computationnel est assimilée à un écart sémantique créé par l'association de termes partageant peu, voire aucun sème. Elle peut donc se fonder sur la violation des préférences sémantiques (Fass, 1991 ; Kintsch, 2000 ; Krishnakumaran et Zhu, 2007), l'opposition concret abstrait (Turney *et al.*, 2011 ; Tsvetkov *et al.*, 2014) ou l'utilisation de termes appartenant à des domaines sémantiques distincts (Schulder et Hovy, 2014 ; Shutova *et al.*, 2017). C'est dans cette dernière famille de méthodes que rentre la méthode de Heintz *et al.* (2013) retenue pour nos expérimentations qui, en plus d'être applicable à plusieurs langues, a été conçue dans une optique sensiblement identique à la nôtre : détecter dans des textes bruts des métaphores ayant un domaine cible spécifique GOVERNANCE combiné à des domaines sources prédéfinis.

Contrairement à d'autres méthodes de détection de métaphores qui utilisent les thématiques extraites d'un corpus au moyen d'algorithmes de *topic modelling* comme un paramètre pour l'apprentissage automatique (Klebanov *et al.*, 2009 ; Bethard *et al.*, 2009 ; Klebanov *et al.*, 2014 ; Jang *et al.*, 2015), cette méthode s'en sert plutôt pour retrouver les mots appartenant à chaque domaine présélectionné. Plus concrètement, cette méthode présuppose qu'une phrase est métaphorique si un domaine source et un domaine cible prédéterminés font partie de ses thématiques prédominantes. Ainsi, dans une phrase telle que « *Moderates, we all hear, are an* ***endangered species*** », prévalent le domaine source, *Animaux* et le domaine cible, POLITIQUE auxquels sont respectivement rattachés les termes *endangered species* et *Moderates*.

Côté ressources, cette méthode non supervisée requiert une liste de domaines sources et cible, des articles de Wikipédia dans la langue de travail et une implémentation de *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003). Ce dernier point est sans aucun doute l'atout majeur de cette méthode. Très utilisés en recherche d'information et en fouille de textes, les modèles thématiques constituent une famille de modèles probabilistes génératifs qui permettent non seulement de faire émerger les mots appartenant aux thématiques présentes dans une collection de documents mais aussi de prédire quelles thématiques cette collection partage avec d'autres.

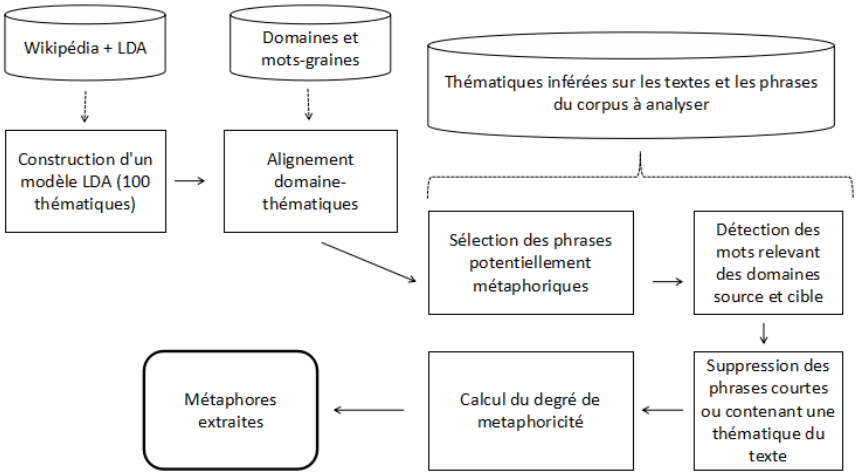

Figure 1. *Schématisation de la méthode de Heintz* et al. *(2013)*

Initialement, la méthode de Heintz *et al.* (2013) a été testée sur un ensemble de textes en anglais et en espagnol avec des résultats satisfaisants : les auteurs rapportent une F-mesure de 59 % pour les expérimentations en anglais. Dans la suite de la section, nous détaillerons les principales phases de cette méthode schématisées dans la figure 1.

2.1. *Phase manuelle autour des domaines cible et sources*

Il s'agit ici d'établir une liste des domaines cible et sources que l'on veut détecter, à raison d'un seul et unique domaine cible et de plus d'un domaine source. Heintz *et al.* (2013) ont défini 62 domaines sources ; nous en avons retenu 42 en concertation avec des spécialistes de l'épistémologie de l'écriture des géographes. Ensuite, pour chaque domaine, nous avons proposé des mots-graines, c'est-à-dire des termes qui, à notre sens, font typiquement partie du champ lexical de ce domaine. Heintz *et al.* (2013) proposent au maximum 4 mots-graines par domaine. Ainsi, un domaine comme *Art* devrait contenir des termes comme « musée », « art », « peinture » et « culture ». Cependant, pour prendre en compte la diversité des sous-disciplines existant en géographie, nous avons suggéré 9 mots-graines pour notre domaine cible. Le tableau 1 recense l'ensemble des domaines ainsi que les mots-graines choisis pour chacun d'eux dans les deux langues de travail ; nous avons indiqué entre parenthèses l'équivalent du mot-graine en anglais lorsque celui-ci diffère du français.

Domaines	Mots-graines
Art	peinture (*painting*), art, musée (*museum*), culture
Astronomie	astronomie (*astronomy*), étoiles (*stars*), planètes (*planet*), télescope (*telescope*)
Barrière	barrière (*barrier*), grillage (*fence*), clôture (*closing*), frontière (*frontier*)
Biologie	biologie (*biology*), espèces (*species*), nature, gène (*gene*)
Chimie	chimie (*chemistry*), acide (*acid*), enzyme, oxygène (*oxygen*)
Compétition	champion, gagner (*win*), course (*race*), médaille (*medal*)
Construction	bâtiment (*building*), pièce (*room*), mur (*wall*), édifice (*house*)
Corps	corps (*body*), tête (*head*), mains (*hands*), os (*bones*)
Créatures	titan, dragon, monstre (*monster*), sorcière (*witch*)
Désordre	fouillis (*jumble*), désordre (*disorder*), enchevêtrement (*tangle*), chaos
Dynamique	mouvement (*going*), bouger (*move*), avancer (*forward*), progrès (*progress*)
Économie	économie (*economy*), argent (*money*), financier (*financial*), banque (*bank*)
Éducation	éducation (*education*), école (*school*), étudiant (*student*), université (*university*)
Émotion	joie (*joy*), colère (*anger*), peur (*fear*), amour (*love*)
Expansion	fertilité (*fertility*), descendant (*offspring*), reproduire (*breed*), développer (*expand*)
GÉOGRAPHIE	territoire (*territory*), terre (*land*), environnement (*environment*), montagne (*mountain*), rivière (*river*), désert (*desert*), climat (*climate*), population, société (*society*)
Guerre	guerre (*war*), bataille (*battle*), armée (*army*), assaut (*attack*)
Langage	langage (*language*), verbe (*verb*), mot (*word*), alphabet
Livre	écrit (*writing*), livre (*book*), page, imprimer (*print*)
Maladie	cancer, guérison (*healing*), douleur (*pain*), maladie (*disease*)
Mathématiques	fonction (*function*), centre (*center*), matrice (*matrix*), cercle (*circle*)
Mécanique	rouage (*gear*), machine, chaîne (*chain*), panne (*breakdown*)
Météorologie	vent (*wind*), tempête (*tempest*), rafale (*blow*), pluie (*rain*)
Meuble	meuble (*furniture*), table, tapis (*carpet*), lit (*bed*)
Monarchie	roi (*king*), reine (*queen*), royaume (*kingdom*), empereur (*emperor*)
Mort	mourir (*die*), enterrement (*burial*), tombeau (*grave*), disparu (*extinct*)
Musique	rythme (*rhythm*), musique (*music*), concert, jazz
Nourriture	nourriture (*food*), viande (*meat*), manger (*eat*), boire (*drink*)
Organe	organe (*organ*), poumons (*lungs*), cœur (*heart*), cerveau (*brain*)
Physique	gravité (*gravity*), énergie (*energy*), nasa, satellite
Politique	gouvernement (*government*), président (*president*), élections (*elections*), politique (*politics*)
Prison	prison, crime, coupable (*culprit*), meurtre (*murder*)
Religion	paradis (*paradise*), saint, dieu (*god*), église (*church*)
Richesse	richesse (*wealth*), bijoux (*jewels*), trésor (*treasure*), riche (*rich*)
Sens	ouïe (*hearing*), vue (*seeing*), doux (*sweet*), amer (*bitter*)
Sport	tennis, football, athlétisme (*athletics*), champion
Statique	arrêté (*stopped*), figé (*rooted*), fixe (*fix*), rester (*stay*)
Technologie	technologie (*technology*), ordinateur (*computer*), internet, software
Théâtre	acteurs (*actors*), jouer (*play*), scène (*stage*), rôle (*role*)
Tribunal	tribunal (*court*), droit (*law*), avocat (*lawyer*), légal (*legal*)
Véhicule	véhicule (*vehicle*), essence (*gas*), voiture (*car*), train
Vêtement	veste (*coat*), plier (*fold*), porter (*wear*), déchirer (*tear*)
Voyage	voyage (*trip*), séjour (*journey*), voyager (*travel*), touristes (*tourists*)

Tableau 1. *Domaines et mots-graines sélectionnés pour nos expérimentations*

2.2. *Extraction des thématiques*

On commence par dresser une liste de mots vides, définis comme étant les 500 mots les plus fréquents du corpus d'articles issus de Wikipédia, afin de pouvoir retrancher ces mots dans la suite du processus. Puis, LDA est utilisée sur ce corpus pour extraire 1 000 thématiques en itérant 1 000 fois et en optimisant les hyperparamètres toutes les 100 itérations. Pour cette étape, la méthode d'origine utilise MALLET (McCallum, 2002)[1], librairie dont nous nous sommes également servis ensuite pour inférer les thématiques de chaque texte et de chaque phrase une fois le modèle LDA construit.

2.3. *Alignement des domaines aux thématiques extraites*

LDA identifiant les différentes thématiques par des nombres aléatoires, le but de cette étape est de pouvoir associer de manière pertinente une ou plusieurs thématiques à un unique domaine source ou cible. Pour ce faire, on considère comme les thématiques liées à un domaine, les n thématiques dans lesquelles la somme de probabilités LDA des mots-graines d'un domaine particulier est maximale pour une thématique et supérieure à la valeur d'un seuil prédéfini z_{align}. Plus formellement, soient t une thématique, c un domaine à aligner et $K(c)$ l'ensemble des mots-graines w associés à ce domaine, t et c sont alignés si : $\sum_{w \in K(c)} p(w|t) > z_{\text{align}}$

Dans l'article de Heintz *et al.* (2013), au maximum 3 thématiques sont alignées avec un domaine source et au maximum 5 avec le domaine cible tandis que le seuil z_{align} est fixé à 0,01. De plus, grâce à cet alignement, on peut également lier les thématiques identifiées précédemment dans chaque phrase et texte aux domaines sources et cible correspondants.

2.4. *Sélection des phrases potentiellement métaphoriques*

En accord avec l'hypothèse de départ, on cherche à déterminer ici les phrases qui parlent à la fois du domaine cible et de l'un des domaines sources prédéfinis. On va donc considérer les 10 thématiques avec la plus forte probabilité inférées par le modèle LDA construit à partir des articles de Wikipédia. Il est impératif d'une part qu'au moins une thématique alignée avec le domaine source et au moins une thématique alignée avec le domaine cible figurent parmi ces 10 premières thématiques inférées par le modèle LDA. D'autre part, on considère séparément pour les thématiques associées au domaine source et au domaine cible la plus forte probabilité en s'assurant que celle-ci est supérieure à un seul prédéfini, z_{relC} respectivement de 0,06 pour le domaine source et de 0,1 pour le domaine cible en anglais chez Heintz *et al.* (2013). Dans le cas où toutes ces conditions ne sont pas respectées, la phrase est ignorée.

1. `http://mallet.cs.umass.edu/`

Soient t une thématique, C un domaine aligné et $\Lambda(C)$ l'ensemble des thématiques dominantes de la phrase x à analyser, ce domaine est jugé être une thématique prédominante de cette phrase si : $\sum_{t \in \Lambda(C)} p(t|x) > z_{\mathrm{relC}}$.

2.5. *Identifications des mots utilisés dans la métaphore*

Dans le lot de phrases restantes, on va détecter précisément quels mots rattachés aux domaines cible et sources résultant de la phase précédente sont utilisés pour construire la métaphore. Cet ensemble de mots A'_C correspond aux mots qui appartiennent simultanément à un domaine prédominant C et à la phrase à analyser x tel que : $\sum_{t \in C} p(t|x) > z_{\mathrm{word}}$

Dans l'article de référence, z_{word} est fixé à 0,1. Par ailleurs, tout mot ne peut être associé qu'à un et un seul domaine. Ainsi, soient A_T, l'ensemble des mots du domaine cible et A_S, celui des mots du domaine source :

$$A_S = A'_S - A'_T$$

et inversement,

$$A_T = A'_T - A'_S$$

2.6. *Application des filtres*

Heintz *et al.* (2013) recensent 2 filtres qui doivent être mis en place pour diminuer le bruit et éliminer des phrases non pertinentes. Premièrement, l'article propose de ne garder que les phrases qui comptent au minimum 4 mots qui ne font pas partie de la liste des mots vides établis à partir des fréquences de mots dans le corpus Wikipédia. Deuxièmement, il suggère d'exclure toutes les phrases dont l'une des 10 thématiques prédominantes est aussi une thématique principale du texte car si le domaine source se retrouve dans tout le texte, il y a de fortes chances qu'il soit employé de manière littérale. À titre d'illustration, dans un texte qui parle de construction d'autoroutes, une phrase telle que « *Congress needs to pass a new highway bill* » serait ignorée pour la détection de métaphores liées au domaine GOVERNANCE même si elle combine bien ce domaine cible et un domaine source prédéfini : *highway* étant fréquemment utilisé dans le texte, son sens est forcément littéral.

2.7. *Calcul du score final*

Enfin, on classe une phrase qui a passé les deux filtres comme étant métaphorique si son degré de métaphoricité est supérieur à un seuil z_{final} manuellement fixé dans l'article à – 10 pour l'anglais. Définissons pour une thématique t, un domaine C, une phrase x et un mot w :

– $\lambda_C = \sum_{w \in K(C)} p(w|t)$

– $\rho_C(x) = \sum_{t \in \Lambda(C)} p(t|x)$
– $R_C(x) = \arg\max_{t \in \Lambda(C)} p(t|x)$
– $\omega_C(w) = \max_{w \in x} \sum_{t \in C} p(t|x)$

Toute phrase x contient une métaphore mettant en relation un domaine cible T et un domaine source S préalablement alignés à une thématique LDA si :

$$\ln\left(\lambda_S \times R_S(x) \times \rho_S(x) \times \omega_S(w) \times \lambda_T \times R_T(x) \times \rho_T(x) \times \omega_T(w)\right) > z_{\text{final}}$$

3. Galop d'essai : expérimentations sur des données en anglais

Cette expérimentation ayant déjà fait l'objet d'une première publication (Beligné *et al.*, 2017), nous allons décrire brièvement l'implémentation du système, le corpus utilisé et les résultats obtenus avant de passer à l'analyse des erreurs et ce que nous en avons retiré pour la suite du projet.

3.1. *Présentation du dispositif expérimental*

Pour cette première expérimentation, la méthode de Heintz *et al.* (2013) a été entièrement réimplémentée en Python. Si la librairie MALLET (McCallum, 2002) a été utilisée pour la création du modèle LDA ainsi que l'inférence des topiques, il est important de signaler que nous n'avons exploité que la moitié des articles de Wikipédia sélectionnés aléatoirement. Nous nous sommes servis de NLTK (Loper et Bird, 2002) pour segmenter les textes en phrases et effectuer la tokénisation. De plus, les seuils inhérents à la méthode étant fixés manuellement, après avoir testé différentes valeurs, nous en avons modifié certaines. Ainsi, nous avons fixé les valeurs respectives de z_{align}, z_{relT} (prédominance du domaine source dans la phrase) et z_{relS} (prédominance du domaine cible dans la phrase) à 0,0075, 0,007 et 0,3. Le corpus d'étude a été constitué à partir de 17 articles écrits entre 2012 et 2016 et comportant entre 30 et 256 phrases pour une taille totale de 41 620 mots. Le tableau 2 décrit plus en détail le corpus utilisé et donne plus d'informations sur l'origine de ces articles ainsi que sur les sujets dont ils traitent.

Afin d'évaluer la performance de la méthode choisie, ce corpus a été annoté par deux annotateurs suivant le protocole MIPVU (Steen *et al.*, 2010) qui consiste de manière très succincte, premièrement à lire le texte pour s'en imprégner puis à faire une deuxième lecture plus poussée en s'arrêtant sur chaque unité lexicale pour déterminer si elle est utilisée métaphoriquement ou non. Sur les 1 527 phrases que contenait au total le corpus, 365 ont été classées comme renfermant une métaphore portant sur un terme géographique. Pour cette tâche d'annotation, l'accord interannotateur calculé au moyen du Kappa de Cohen (Cohen, 1960) est de 0,63, ce qui est de loin supérieur au taux de 0,48 rapporté par Heintz *et al.* (2013) qui avaient également fait appel à deux annotateurs.

Support	Type	Nombre d'articles	Sujets
Atlantic Geology[2]	Revue scientifique	4	Roches, sédiments, marais salant
Nature Climate Change[3]	Revue scientifique	1	Niveau de l'eau
Hypergeo[4]	Encyclopédie en ligne	2	Érosion, océan
Geocurrents[5]	Forum d'échanges d'idées	2	Caucase, Australie
123helpme[6]	Base de devoirs en ligne	3	Hongrie, Papouasie, région arctique
The Guardian[7]	Journal	2	Population mondiale, région de Douro
Carbonbrief[8]	Site d'informations sur le changement climatique	2	Croissance des arbres, dégradation des forêts
American Geophysical Union [9]	Site d'informations sur la géophysique	1	El Niño

Tableau 2. *Description du corpus utilisé pour les expérimentations en anglais*

3.2. *Analyse des erreurs*

Dans l'ensemble, la performance du système s'est avérée plus que décevante. Seuls 27 des 43 domaines prédéfinis ont pu être automatiquement alignés avec une thématique LDA. En outre, comme le montre la matrice de confusion (tableau 3), les résultats finaux sont inexploitables pour une future automatisation : 58,9 % de précision et 18,9 % de rappel. Il faut préciser que ces résultats ne concernent que la classification d'une phrase comme étant métaphorique ou non. En regardant de plus près, au niveau de l'attribution des domaines sources, nous nous sommes rendus compte que sur les

2. `https://journals.lib.unb.ca/index.php/ag`
3. `https://www.nature.com/nclimate/`
4. `http://www.hypergeo.eu/`
5. `http://www.geocurrents.info/`
6. `https://www.123helpme.com/`
7. `https://www.theguardian.com`
8. `https://www.carbonbrief.org/`
9. `https://news.agu.org`

69 phrases correctement détectées, le bon domaine source n'a été trouvé que dans une seule phrase : « *The county of* ***hungry*** *is inhabited by roughly — 9,919,128 as of July 2014 [. . .]* ». Cependant, au vu du contexte d'utilisation, on s'aperçoit vite que cette phrase comporte deux coquilles : *county* au lieu de *country* et *hungry* à la place de « Hungary ». Cette première découverte soulève la question de la qualité non seulement de l'OCR des textes choisis, mais aussi de leur annotation. En effet, on pourrait supposer que des locuteurs natifs auraient remarqué de telles erreurs et les auraient corrigées.

		Méthode automatique		
		Métaphores	**Non-métaphores**	**Total**
Annotation manuelle	**Métaphores**	69	48	117
	Non-métaphores	296	1114	1410
	Total	365	1162	

Tableau 3. *Matrice de confusion pour les expérimentations en anglais*

En ce qui concerne les autres phrases correctement identifiées comme étant métaphoriques, nous avons décelé trois types d'erreurs liées à la détection des domaines sources :

– un domaine source aligné assigné à la place d'un autre domaine source aligné (26 phrases).
Exemple : *The Saint John River* ***bisects*** *the city at a location down river of the point of tidal influence that extends up-river to Mactaquac Da*m [*Politique* au lieu de *Mathématiques*] ;

– un domaine source aligné assigné à la place d'un domaine source non aligné (14 phrases).
Exemple : *To start we will look at the river that* ***cuts*** *through the grand city* [. . .] [*Théâtre* au lieu de *Mort*] ;

– un domaine aligné assigné à la place d'un domaine hors liste ou inexistant (28 phrases), ce qui relève à nouveau d'un problème d'annotation.
Exemple : *Freeze-thaw exploitation is dependent upon microclimate, which* ***controls*** *the number of cycles. . .* [attribué à *Tribunal* ; pas de domaine source précisé].

Pour mieux comprendre pourquoi l'assignation des domaines sources et l'identification des mots utilisés dans la métaphore fonctionnent aussi mal, nous sommes inté-

ressés dans un premier temps à l'alignement obtenu à partir du modèle LDA construit. Nous en avons retiré d'une part que contrairement aux substantifs, les verbes et les adjectifs ont tendance à avoir des probabilités significatives dans plusieurs domaines. Par exemple, *called* se retrouve à la fois dans les domaines *Mathématiques* et *Nourriture* tandis que *high* est présent dans les domaines GÉOGRAPHIE et *Éducation*. De plus, la probabilité attribuée à chaque mot est assez basse, la probabilité la plus haute dans un domaine aligné étant très souvent de l'ordre de 0,05. En classant ces probabilités par ordre décroissant et en considérant pour chacun des 27 domaines alignés les cent premières probabilités, on constate que seuls 81 mots ont une probabilité comprise entre 0,05 et 0,02 tandis que 204 mots ont une probabilité de l'ordre de 0,01. En comparant les scores d'alignement avec certains exemples donnés par Heintz *et al.* (2013), on note tout de suite une grande différence. Par exemple, pour le domaine *Véhicule*, leurs scores d'alignement sont de 0,035, 0,29 et 0,022 alors que nous obtenons pour le même domaine des valeurs deux fois ou plus inférieures : 0,011, 0,014 et 0,009. De plus, avec notre modèle LDA, en utilisant les mots-graines fournis par Heintz *et al.* (2013), nous ne parvenons pas à aligner avec succès comme eux le domaine *Animals*. L'une des hypothèses plausibles qu'on pourrait avancer pour expliquer cet échec serait que la partie de Wikipédia utilisée pour notre expérimentation couvre très mal ou peu certains domaines.

Ensuite, nous avons analysé la détection des thématiques prédominantes dans les textes et les phrases. Très souvent, le domaine inféré par le modèle LDA est erroné parce que le domaine source de la métaphore est souvent utilisé ponctuellement et n'est pas dominant dans toute la phrase.
Exemple : *The relief changes from the floodplain at an average of 10 m elevation to greater than 100 m in elevation at the top of the steep valley* ***wall*** *to the south to be related to the emplacement of the thrust slice, but is directed toward the west.*
Ici, le domaine source attendu, *Construction*, est surtout représenté par le mot *wall* même si on pourrait argumenter que *elevation* est utilisé dans le domaine de l'architecture et qu'*emplacement* peut désigner l'endroit où se situe un bâtiment. Dans d'autres cas, le domaine source est choisi par défaut car il figure parmi les 10 thématiques dominantes sans pour autant forcément être le domaine prédominant de la phrase.
Exemple : *The Artic is one of the few* ***unspoilt*** *wilderness areas in the world and must be conserved.*
Le domaine source attribué ici automatiquement *Éducation* est en fait la quatrième thématique mais devient par défaut la première thématique puisque les trois premières thématiques inférées ne sont pas alignées. Le même problème se rencontre lorsque l'on regarde de plus près la reconnaissance des domaines dans les faux positifs. À titre d'illustration, dans la phrase « *The typical aspect of the contact can be recognized as a gently westward-dipping plane* », deux raisons justifient que cette phrase soit considérée à tort comme contenant une métaphore liée au domaine *Musique* :

– la plupart des mots de la phrase se trouvent dans une thématique associée au domaine *Musique* ;

– *Musique* est la cinquième thématique prédominante dans la phrase avec une probabilité de 0,029 tandis que la première thématique a une probabilité de 0,18, et GÉOGRAPHIE, la troisième, une probabilité de 0,03.

Nous avons par la suite comparé, dans la mesure du possible, notre dispositif expérimental avec les informations que nous avons pu glaner de l'article de Heintz *et al.* (2013), les auteurs n'ayant pu nous fournir une copie du code et des données utilisés. Dans un premier temps, nous avons remarqué que presque tous les exemples donnés comme étant bien analysés par le système se rapportent à des phrases se composant d'une proposition indépendante, ce qui rend difficile toute interférence d'autres domaines sources. Il en est de même du seul exemple comptant deux propositions « *Moderates, we all hear, are an* ***endangered species*** ». Si on élimine tous les mots vides, on se retrouve avec 4 mots pleins qui limitent grandement le nombre de thématiques prédominantes possibles : *moderates*, *hear*, *endangered*, *species*. Il va sans dire que beaucoup de phrases de notre corpus sont bien plus complexes syntaxiquement ou comportent beaucoup plus de mots pleins. Pour y remédier, nous avons procédé à des tests avec des 4-grammes à la place de phrases entières mais n'avons pas relevé d'améliorations significatives.

Enfin, il est important de souligner que la méthode de Heintz *et al.* (2013) n'a jamais été évaluée sur l'ensemble de leur corpus sur lequel nous n'avons aucune précision hormis son origine (articles de journaux et de blogs). Les données avec lesquelles ils affirment avoir obtenu une F-mesure de 59 % sont composées de 600 phrases tirées pour une moitié, pour chaque domaine source et domaine cible, de phrases ayant les cinq plus hauts degrés de métaphoricité finaux auxquels s'ajoutent d'autres phrases jugées métaphoriques sélectionnées en fonction de leur degré de métaphoricité. L'autre moitié, quant à elle, contient 300 phrases classées par le système comme étant non métaphoriques et choisies au hasard. Dans la seconde évaluation dans laquelle il a été demandé à des utilisateurs d'AmazonTurk de juger de la métaphoricité des 250 phrases auxquelles leur méthode a attribué le plus haut degré de métaphoricité, la métaphoricité moyenne des métaphores conceptuelles annotées est de 0,39.

4. Nouveau protocole expérimental avec des données en français

À la suite de cette analyse d'erreurs qui a fait remonter les faiblesses du système implémenté, nous avons défini différentes modifications dans le protocole expérimental qui à notre sens devraient nous permettre d'obtenir des résultats plus pertinents.

Premièrement, nous avons résolu de guider l'extraction des thématiques LDA afin de mieux faire ressortir les domaines qui nous intéressent en sélectionnant automatiquement les articles de Wikipédia qui s'y rapportent. Par ailleurs, nous avons aussi décidé de faire varier le nombre de thématiques extraites afin de mesurer leur incidence sur l'alignement thématique domaine. Enfin, dans l'idée d'attribuer de plus fortes probabilités aux mots réellement caractéristiques d'une thématique, la probabilité qu'un mot appartienne à une thématique donnée a été calculée différemment.

Cette fois, nous avons travaillé sur des données en français, changement qui s'explique par deux principaux facteurs :

– le manque d'adéquation entre le corpus en anglais et notre objectif final : en effet, le corpus en anglais se compose en grande majorité de textes informatifs non scientifiques qui ne sont pas pour la plupart écrits par des géographes et ne correspondent donc pas au type de textes que nous souhaiterions analyser à terme ;

– les différents défauts de l'annotation déjà mentionnés plus haut : plutôt que de reprendre l'annotation sur des textes, somme toute, peu pertinents et avec des annotateurs qui maîtrisent approximativement l'anglais, travailler en français nous garantissait un accès plus facile à des locuteurs natifs formés en géographie.

Dans la suite de cette section, nous allons présenter les différentes modifications apportées à la méthode originale, puis nous nous intéresserons au corpus utilisé ainsi qu'à son annotation avant de discuter les résultats obtenus.

4.1. *Sélection sémantique des articles de Wikipédia*

L'idée de filtrer les articles en fonction de leur contenu nous a été inspirée par les travaux de Phan *et al.* (2008) qui utilisent des mots-clés pour choisir des pages Wikipédia parlant des thématiques qui les intéressent. Si l'extraction de thématiques LDA qui en résulte donne de bons résultats, malheureusement, aucune précision n'est donnée sur le choix de mots-clés ou le seuil minimal de mots en commun requis pour qu'un texte soit considéré comme pertinent. Plutôt que de nous limiter aux mots-graines de chaque domaine, afin de couvrir plus exhaustivement les domaines sources et cible, nous avons choisi d'utiliser le *Dictionnaire électronique des mots* (Dubois et Dubois-Charlier, 2010) qui associe chaque lemme d'une entrée à un domaine sémantique prototypique[10].

Au total, le *Dictionnaire électronique des mots* compte 32 des 43 domaines prédéfinis pour cette tâche, le présupposé étant que les 11 domaines absents (*Statique*, *Dynamique*, *Mort*, *Expansion*, *Sens*, *Désordre*, *Richesse*, *Prison*, *Monarchie*, *Compétition*, *Barrière*) sont soit inclus intégralement ou partiellement dans d'autres domaines, soit latents, dans les textes présélectionnés. Par exemple, le domaine *Monarchie* fait partie du domaine *Politique*, *Compétition* de celui de *Sports*, *Prison* de *Tribunal*, etc.

Une fois cette liste de mots-clés construite, nous avons utilisé Morphalou3[11] pour ajouter automatiquement les formes fléchies de chaque mot. Puis, nous avons gardé tout article de Wikipédia qui avait au moins 10 mots en commun avec l'un des domaines à aligner. Ce seuil a été choisi en tenant compte du fait que plusieurs des métadonnées des pages de Wikipédia telles que « Portail », « Catégorie », « Notes et références », « Articles connexes » sont susceptibles de contenir des termes qui peuvent appartenir au lexique de mots-clés. Il va donc de soi que les pages de Wikipédia sont

10. `http://rali.iro.umontreal.ca/DEM/domaines/index.html`
11. `https://repository.ortolang.fr/api/content/morphalou/2/LISEZ_MOI.html`

uniquement nettoyées *a posteriori*. Enfin, nous avons réduit le corpus aux articles les plus longs, c'est-à-dire, ceux qui dépassent 9,90 ko afin d'avoir suffisamment de contexte pour chaque mot.

4.2. *Augmentation du nombre de thématiques LDA extraites*

Le nombre de 100 thématiques fixé par Heintz *et al.* (2013) semble relativement modeste, surtout lorsque l'on pense d'une part, au nombre de thématiques que Wikipédia peut couvrir et d'autre part, qu'on aligne un domaine source au maximum avec 3 thématiques. Navarro Colorado et Tomás (2015) ayant montré qu'une sortie de 1 000 ou 2 500 thématiques sur l'ensemble de Wikipédia permet d'obtenir une meilleure granularité de thématiques et de mieux circonscrire dans une même thématique les mots appartenant au même champ lexical, nous avons également construit toujours à partir de notre corpus Wikipédia en français, des modèles LDA avec 1 000 et 2 500 thématiques.

4.3. *Identification des mots discriminants pour chaque thématique*

La métaphore se produisant surtout au niveau lexical, et prenant en compte la dimension probabiliste de LDA, il nous a semblé primordial de pouvoir définir si un mot est caractéristique d'une thématique particulière, en calculant différemment la probabilité qu'un mot relève d'une thématique donnée. Ainsi, soit z une thématique extraite, w un mot du corpus Wikipédia, et d un document de ce corpus :

$$p(z/w) = \frac{p(w/z) \times p(z)}{p(w)}$$

où $p(z) = \sum p(z/d) \times p(d)$ et p$(w) = \sum p(w/z) \times p(z)$

Une fois, ces nouvelles probabilités générées, l'entropie de Shannon a servi à filtrer les mots passe-partout qui ne sont discriminants pour aucune thématique, c'est-à-dire ceux dont aucune probabilité dans une thématique ne dépasse l'entropie de toutes leurs probabilités :

$$H(p(z/w)) = -\sum p(z/w) \log p(z/w)$$

Enfin, nous avons éliminé dans chaque phrase tous les mots non discriminants avant de procéder à l'inférence des thématiques.

4.4. *Présentation du corpus et de la méthode d'annotation*

Une dizaine d'articles parus entre 1972 et 1999 dans *L'Espace Géographique* a été sélectionnée au hasard pour constituer notre corpus d'évaluation. Les trois décennies

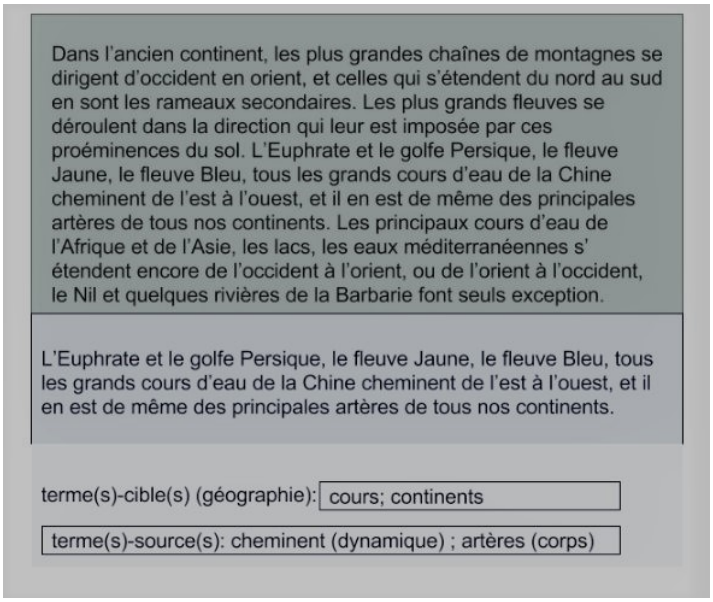

Figure 2. *Vue d'une phrase à annoter*

que recouvre ce corpus sont représentées (au moins 3 articles par décennie). Ce corpus rentre dans diverses sous-disciplines géographiques telles que la méthodologie de la recherche en géographie, la géographie économique ou encore la géographie humaine. Tous les textes sont issus du portail Persée[12].

Pour annoter le corpus, nous avons choisi de faire appel à des étudiants en master de géographie. À cet effet, nous avons mis en place une plate-forme en ligne dans laquelle chaque texte s'affiche phrase par phrase. Afin de la situer dans son contexte, chaque phrase est toujours affichée en dessous du paragraphe dont elle est tirée. Une fois la phrase lue, les annotateurs doivent indiquer si elle contient une métaphore ou non. Dans le premier cas, ils doivent alors rentrer les termes cibles et sources de la métaphore et préciser uniquement le ou les domaines auxquels le ou les termes sources appartiennent (figure 2). Avant le début de la phase d'annotation proprement dite, une séance d'explication a été organisée durant laquelle les grandes lignes du projet ont été exposées aux étudiants. Nous avons également conçu un guide d'annotation qui recensait tous les domaines sources, présentait des exemples d'annotation et décrivait les différentes étapes à suivre pour identifier les métaphores selon le protocole MIPVU (Steen *et al.*, 2010).

12. `https://www.persee.fr/`

Nous disposions au total de 9 annotateurs qui ont annoté partiellement 4 textes, soit 435 phrases dont 51 ont été considérées comme étant métaphoriques par plus de la moitié des annotateurs. On constate que le taux d'accord interannotateur est plutôt bas en ce qui concerne les phrases métaphoriques (0,30) par rapport à celui de l'étiquetage des domaines sources (0,52).

4.5. *Résultats et discussion*

Afin de mieux cibler les failles du système d'identification des métaphores, en plus de mesurer sa précision et son rappel pour la détection des phrases annotées comme étant métaphoriques, nous avons également évalué sa précision en ce qui concerne la détection des domaines sources.

Au terme de la phase de filtrage sémantique de Wikipédia, nous avons obtenu un corpus final de Wikipédia de 68 404 articles pour un total de 1 570 718 mots uniques[13]. Pour l'ensemble des tâches relevant de l'extraction des thématiques, nous avons utilisé Gensim (Řehůřek et Sojka, 2010) en itérant chaque fois 1 000 fois et en ne gardant pour chaque thématique que les 2 000 premiers mots avec la plus forte probabilité. Le tableau 4 montre qu'effectivement, déjà en réduisant le nombre de pages Wikipédia, on améliore légèrement l'alignement des domaines, cependant on reste dans le même ordre de résultats que ce que nous avions déjà obtenu pour l'expérimentation en anglais.

Corpus	Nombre de thématiques	Nombre de domaines alignés
Wikipédia - corpus entier	100	20
Wikipédia intermédiaire (873 807 articles)	100	22
Wikipédia final (68 404 articles)	100	23

Tableau 4. *Domaines alignés en fonction des corpus Wikipédia*

De même, comme on peut le voir dans le tableau 5, sur le modèle LDA construit à partir du corpus Wikipédia final, en ne considérant que les cinquante premiers mots et 2 500 thématiques, on améliore sensiblement l'alignement des domaines de sorte que 100 % des concepts sont alignés. Néanmoins, la comparaison des différentes performances du système listées dans le tableau 6 laisse à penser qu'un alignement optimal

13. La version de Wikipédia utilisée est celle de 20/07/2017, qui comporte environ 1 900 000 pages (articles, pages de discussion, pages de désambiguïsation…). Le fichier XML brut obtenu a été nettoyé de toutes les balises superflues avec WikiExtractor (`https://github.com/attardi/wikiextractor`).

n'a aucun impact majeur, au contraire. En effet, les meilleurs résultats sont obtenus avec 100 thématiques conformément à ce que proposent Heintz *et al.* (2013). En outre, plus on augmente le nombre de thématiques extraites, plus on fait baisser le rappel du système sans pour autant améliorer la détection des domaines sources.

Nombre de thématiques	Nombre de mots considérés avec la plus forte probabilité	Nombre de domaines alignés
100	2 000	23
1 000	20	42
2 5000	50	43

Tableau 5. *Impact du nombre de thématiques extraites sur l'alignement thématiques domaine*

Paradigmes	Précision système	Rappel système	Précision domaines sources
100 thématiques	16,60 %	49,00 %	0,04 %
100 thématiques + $p(z/w)$	16,20 %	43,10 %	0 %
100 thématiques + $p(z/w)$ + entropie	18,10 %	7,80 %	0 %
1 000 thématiques	18,40 %	23,50 %	0,10 %
1 000 thématiques + $p(z/w)$	15,20 %	25,40 %	0 %
2 500 thématiques	75,00 %	5,80 %	0 %
2 500 thématiques + $p(z/w)$	18,20 %	8,80 %	0 %

Tableau 6. *Performances du système*

Comme à l'étape précédente, nous avons procédé à une analyse des erreurs. D'abord, en ce qui concerne l'alignement des domaines, nous avons pu cibler quatre principaux problèmes :

– certains domaines sont ambigus, ce qui complique leur alignement. Le domaine *Barrière*, par exemple, est associé à des mots du type « sydney »,« bactérie »,« infectieuse »,« sauveteurs », et donc se réfère plutôt vraisemblablement à la Grande barrière de corail ;

– de même, certains des mots-graines proposés sont polysémiques et faussent l'alignement thématique domaine. C'est ainsi que la présence du mot « étoiles » lie le domaine *Astronomie* à une thématique qui fait plus penser au monde du cinéma : « passionné », « herbert », « julia », « marcello »« akkad », « réfuter »… ;

– certains domaines alignés ne sont pas suffisamment distincts et semblent contenir des mots se rapportant à d'autres domaines. Dans les thématiques alignées aux domaines *Art* et *Chimie*, on retrouve des noms d'autres domaines ; « religion » dans le premier domaine et dans le second, « biologie » et « astronomie ». Dans ce dernier exemple, il serait plus plausible que cette thématique relève des la science en général ou des sujets scolaires. De même, toujours dans une des thématiques associées au domaine *Art*, on retrouve plutôt des mots qui relèvent clairement de la géographie : (« peuples », « population », « régions », « rurales ») ;

– le grand nombre d'entités nommées surtout en ce qui concerne les lieux dans des domaines autres que GÉOGRAPHIE, par exemple « Paris » et « Tokyo » dans le domaine *Art*.

Au niveau de l'attribution des domaines sources, nous avons noté que la détection des mots impliqués dans la métaphore pose toujours problème. Par exemple, dans la phrase « La porte d'un parc est, à l'origine, un ensemble de réalisations, situées au fond d'une vallée, au plus près des limites de la zone centrale », le système conclut que la métaphore met en relation « origine » du domaine GÉOGRAPHIE et « ensemble » du domaine *Construction* respectivement à la place de « parc » et « porte ».

5. Bilan et nouvelles perspectives de travail

Au vu de ces nouveaux résultats négatifs, nous ne pouvons que conclure que la méthode testée est totalement inadaptée à notre objet de recherche ; avons-nous pour autant perdu notre temps ? Loin de là. D'une part, cet échec nous a permis de nous interroger sur le rôle que joue le type de textes dans la détection automatique des métaphores et d'autre part, nous a poussés à nous poser d'importantes questions méthodologiques touchant aussi bien la formulation de notre objet de recherche que notre protocole d'annotation.

La question de la portabilité de la méthode choisie rejoint une problématique récurrente en humanités numériques où les algorithmes de traitement automatique des langues disponibles ont souvent été uniquement testés sur un seul type de textes, typiquement des articles de journaux. En ce qui concerne la détection automatique des métaphores dans les textes, l'utilisation d'articles journalistiques s'explique d'autant plus qu'à cause du nombre limité, voire inexistant, de corpus pour plusieurs langues et de la perception subjective de ce que constitue une métaphore chez des annotateurs non experts, les recherches menées se sont surtout focalisées sur les types de textes les plus susceptibles de nous renseigner sur ce qui se passe au quotidien et comment nous le ressentons. Afin d'adopter la stratégie la plus appropriée face à un genre de textes jusque-là inexploré en détection automatique des métaphores, il nous semble opportun de nous demander dans un premier temps, quelle fonction joue la métaphore dans le discours des géographes et si toutes les métaphores sont équivalentes. En d'autres termes, des syntagmes du type « la porte du parc » ou encore « le flanc de la colline » traduisent-ils une idéologie particulière de l'auteur ou un simple choix stylistique ?

Exactement, à partir de quel moment les métaphores des géographes commencent-elles à devenir singulières et signifiantes ?

Pour répondre à ces questions, nous avons examiné des métaphores de géographes citées en exemple dans différents articles. On constate ainsi qu'en général, ces métaphores expriment une façon de penser et de voir le monde, mais aussi que très souvent, elles traversent plusieurs écrits du même auteur (par exemple Vidal de La Blache et Jean Brunhes pour la métaphore organiciste). Par conséquent, ces métaphores idéologiques pourraient se concevoir comme la surprésence dans un texte ou chez un auteur de termes appartenant à un ou plusieurs domaines distincts de la géographie. Comment donc déceler cette surprésence ?

À notre sens, cela passe en premier lieu par la construction d'un lexique exhaustif pour chacun des domaines prédéfinis. Cependant, en regardant de plus près notre liste initiale de domaines, leur hétérogénéité saute immédiatement aux yeux. En effet, si certains d'entre eux appartiennent à des champs d'étude bien définis (*Biologie*, *Art*...), d'autres au contraire se réfèrent à un concept bien précis (*Mort*, *Barrière*, *Expansion*...). Par conséquent, deux stratégies distinctes doivent être mises en place. En ce qui concerne le premier groupe de termes, pour les domaines qui y sont répertoriés, le plus simple nous paraît de tirer parti de la liste de termes par domaine extraite du *Dictionnaire électronique des mots* (Dubois et Dubois-Charlier, 2010) et des métadonnées des articles de Wikipédia liant un terme à un portail spécifique. Pour le second cas de figure, il serait plus approprié d'utiliser des dictionnaires de synonymes, des bases lexicales listant les mots de la même famille ou encore des modèles de plongements de mots préconstruits pour construire automatiquement le champ lexical de ces concepts. Une vérification manuelle pourrait être envisagée pour éliminer les termes les moins pertinents.

Ensuite, afin de pouvoir détecter la surutilisation des lexiques construits, il nous faudra un plus large corpus diachronique composé d'une part d'articles de géographie subdivisés en sous-disciplines et étiquetés par auteur et d'autre part, de textes plus génériques de la même période portant sur une toute autre discipline qui nous serviront de référence. Ainsi, nous pourrons inférer si un domaine est surreprésenté et régulièrement convoqué chez un auteur non seulement par rapport à sa présence relative dans les textes de référence, mais aussi dans les écrits des géographes contemporains à l'auteur en question. Il est important de souligner ici que certains domaines sont étroitement liés entre eux et ne devraient pas être impérativement considérés comme des domaines complètement séparés, par exemple *Organe*, *Mort* et *Maladie*.

Enfin, une fois les phrases intéressantes ainsi isolées, nous pourrons nous fonder sur différents scénarios syntaxiques de constructions métaphoriques tels que ceux définis par Tamine (1979), Krishnakumaran et Zhu (2007) et Dodge *et al.* (2015) pour identifier les termes relevant du domaine cible impliqués dans la métaphore. Mise à part la possibilité de détecter les métaphores *in absentia* ainsi que celles dans lesquelles le terme du domaine source est remplacé par un pronom personnel, cette nouvelle approche présente différents autres avantages. D'abord, elle nous évitera de construire une liste de termes pour le domaine GÉOGRAPHIE, tâche particulièrement

compliquée au vu des différents sous-domaines et thématiques que la géographie peut couvrir. Ensuite, l'annotation manuelle ne se concentrera, dans un premier temps, que sur les phrases présélectionnées. Enfin, elle nous permettra de construire de façon semi-automatique un corpus annoté qui pourra nous permettre ultérieurement d'utiliser des méthodes d'apprentissage supervisé pour découvrir les métaphores qui nous auraient échappé.

6. Conclusion

La métaphore joue un rôle primordial dans les textes de sciences humaines et sociales non seulement en tant qu'élément du langage, mais aussi en tant qu'agent de la circulation des idées. À cet effet, nous avons présenté dans cet article un projet en humanités numériques qui a pour but de détecter des métaphores dans les écrits des géographes et utilise principalement la méthode de Heintz *et al.* (2013) testée en anglais et en espagnol pour identifier des métaphores conceptuelles ayant pour domaine cible GOVERNANCE. Bien que nos expérimentations en français et en anglais soient loin d'avoir généré des résultats exploitables pour la suite, cette étude non seulement pose d'importantes questions sur l'adéquation de certaines méthodes de traitement automatique des langues pour des textes de sciences humaines et sociales, mais aussi nous a permis de radicalement redéfinir notre approche afin de mieux répondre à nos objectifs de départ.

Remerciements

Ce travail a été financé par l'Institut universitaire de France (IUF) et s'est effectué en collaboration avec Sabine Loudcher (ERIC, Université Lumière Lyon 2), Julien Velcin (ERIC, Université Lumière Lyon 2), Max Béligné (EVS, Université Lumière Lyon 2) et Isabelle Lefort (EVS, Université Lumière Lyon 2).

7. Bibliographie

Archer K., « Regions as social organisms : The Lamarckian characteristics of Vidal de la Blache's regional geography », *Annals of the Association of American Geographers*, vol. 83, n° 3, p. 498-514, 1993.

Aristote, *Poétique et rhétorique*, Garnier Frères, 1922.

Ascher F., « La métaphore est un transport. », *Cahiers internationaux de sociologie*, n° 1, p. 37-54, 2005.

Bachelard G., *La Formation de l'esprit scientifique : Contribution à une psychanalyse de la connaissance*, Vrin, 1967.

Bachimon P., « Physiologie d'un langage. L'organicisme aux débuts de la géographie humaine », *Espace Temps*, vol. 13, n° 1, p. 75-103, 1979.

Beligné M., Campar A., Chauchat J.-H., Lefeuvre M., Lefort I., Loudcher S., Velcin J., « Détection automatique de métaphores dans des textes de Géographie : une étude prospective », *Actes de la Conférence sur le traitement automatique des langues naturelles (TALN), 2017*, vol. 2, p. 86-93, 2017.

Berdoulay V., « La métaphore organiciste : Contribution à l'étude du langage des géographes », *Annales de géographie*, p. 573-586, 1982.

Bethard S., Lai V. T., Martin J. H., « Topic model analysis of metaphor frequency for psycholinguistic stimuli », *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, Association for Computational Linguistics, p. 9-16, 2009.

Blei D. M., Ng A. Y., Jordan M. I., « Latent dirichlet allocation », *Journal of machine Learning research*, vol. 3, , p. 993-1022, 2003.

Cohen J., « A coefficient of agreement for nominal scales », *Educational and psychological measurement*, vol. 20, n° 1, p. 37-46, 1960.

Daniels S., Cosgrove D., « Landscape metaphors in cultural geography », *Place/culture/representation*, 1993.

Deguy M., « Vers une théorie de la figure généralisée », *Critique*, vol. XXV, n° 269, p. 841-861, 1969.

Dodge E., Hong J., Stickles E., « MetaNet : Deep semantic automatic metaphor analysis », *Proceedings of the Third Workshop on Metaphor in NLP*, p. 40-49, 2015.

Dubois J., Dubois-Charlier F., « La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration », *Langages*, n° 3, p. 31-56, 2010.

Dumarsais C., *Les Tropes, éd*, Fontanier, Paris, 1818.

Fass D., « met* : A method for discriminating metonymy and metaphor by computer », *Computational Linguistics*, vol. 17, n° 1, p. 49-90, 1991.

Heintz I., Gabbard R., Srivastava M., Barner D., Black D., Friedman M., Weischedel R., « Automatic extraction of linguistic metaphors with lda topic modeling », *Proceedings of the First Workshop on Metaphor in NLP*, p. 58-66, 2013.

Jang H., Moon S., Jo Y., Rose C., « Metaphor detection in discourse », *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 384-392, 2015.

Kintsch W., « Metaphor comprehension : A computational theory », *Psychonomic bulletin & review*, vol. 7, n° 2, p. 257-266, 2000.

Klebanov B. B., Beigman E., Diermeier D., « Discourse topics and metaphors », *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, p. 1-8, 2009.

Klebanov B. B., Leong B., Heilman M., Flor M., « Different texts, same metaphors : Unigrams and beyond », *Proceedings of the Second Workshop on Metaphor in NLP*, p. 11-17, 2014.

Krishnakumaran S., Zhu X., « Hunting Elusive Metaphors Using Lexical Resources. », *Proceedings of the Workshop on Computational approaches to Figurative Language*, p. 13-20, 2007.

Lakoff G., Johnson M., *Metaphors we live by*, University of Chicago Press, 1980.

Lévy B., « Géographie et littérature. Une synthèse historique », *Le globe*, vol. 146, p. 25-52, 2006.

Loper E., Bird S., « NLTK : The Natural Language Toolkit », *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and*

Computational Linguistics - Volume 1, ETMTNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 63-70, 2002.

McCallum A. K., « MALLET : A Machine Learning for Language Toolkit », 2002, http ://mallet.cs.umass.edu.

Molino J., « Métaphores, modèles et analogies dans les sciences », *Langages*, n° 54, p. 83-102, 1979.

Navarro Colorado B., Tomás D., « A fully unsupervised Topic Modeling approach to metaphor identification », *Actas del XXXI Congreso de la Sociedad Espanola para el Procesamiento del Lenguaje Natural (SEPLN 2015). http ://www. dlsi. ua. es/˜ borja/NavarroTomas_PosterSEPLN2015. pdf*, 2015.

Phan X.-H., Nguyen L.-M., Horiguchi S., « Learning to classify short and sparse text & web with hidden topics from large-scale data collections », *Proceedings of the 17th international conference on World Wide Web*, ACM, p. 91-100, 2008.

Řehůřek R., Sojka P., « Software Framework for Topic Modelling with Large Corpora », *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45-50, 2010. `http://is.muni.cz/publication/884893/en`.

Schulder M., Hovy E., « Metaphor detection through term relevance », *Proceedings of the Second Workshop on Metaphor in NLP*, p. 18-26, 2014.

Shutova E., Sun L., Gutiérrez E. D., Lichtenstein P., Narayanan S., « Multilingual metaphor processing : Experiments with semi-supervised and unsupervised learning », *Computational Linguistics*, vol. 43, n° 1, p. 71-123, 2017.

Steen G. J., Dorst A. G., Herrmann J. B., Kaal A., Krennmayr T., Pasma T., *A Method for Linguistic Metaphor Identification : From MIP to MIPVU*, vol. 14, John Benjamins Publishing, 2010.

Tamine J., « Métaphore et syntaxe », *Langages*, n° 54, p. 65-81, 1979.

Tsvetkov Y., Boytsov L., Gershman A., Nyberg E., Dyer C., « Metaphor detection with cross-lingual model transfer », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, vol. 1, p. 248-258, 2014.

Turney P. D., Neuman Y., Assaf D., Cohen Y., « Literal and metaphorical sense identification through concrete and abstract context », *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 680-690, 2011.

The names of lighting artefacts: extraction and representation of Portuguese and Spanish terms in the archaeology of al-Andalus

Bruno Almeida* — **Rute Costa*** — **Christophe Roche****

* *NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal*
** *Condillac Group, LISTIC, Université Savoie Mont-Blanc, France*

ABSTRACT. *This paper is focussed on the Portuguese and Spanish terms for lighting artefacts, which were extracted from a corpus on the archaeology of al-Andalus. The purpose of the work described in this paper is the creation of an ontology-based multilingual terminological resource. Domain knowledge is represented through OntoAndalus, an OWL ontology which uses DOLCE+DnS Ultralite as a foundation. Language-specific information are modelled through Lemon, the Lexicon Model for Ontologies, which is currently in development by a community group within the W3C. Lemon allows for the representation of grammatical and semantic information, most notably lexicosemantic relations between terms and their reference to ontology elements in OntoAndalus.*

RÉSUMÉ. *Cet article se concentre sur les termes des artefacts d'éclairage dans un corpus sur l'archéologie d'al-Andalus. Le but de ce travail est la création d'une ressource terminologique multilingue basée sur une ontologie. La connaissance du domaine est représentée par OntoAndalus, une ontologie OWL qui repose sur DOLCE + DnS Ultralite. Les informations spécifiques à la langue sont représentées par le modèle Lemon, le Modèle lexical pour les ontologies, lequel est en cours de développement par le W3C. Lemon permet la représentation d'informations grammaticales et sémantiques, notamment les relations lexicosémantiques entre les termes et sa référence aux éléments ontologiques dans OntoAndalus.*

KEYWORDS: *Terminology, Lighting artefacts, Archaeology of al-Andalus, Corpus analysis, Ontologies, Lexicon Model for Ontologies (Lemon).*

MOTS-CLÉS : *terminologie, artefacts d'éclairage, archeologie de l'al-Andalus, analyse de corpus, ontologies, Modèle lexical pour les ontologies (Lemon).*

TAL. Volume 60 – n° 3/2019, pages 113 à 137

1. Introduction

Corpora have had multiple research and practical applications in the past decades, from linguistics and terminology to digital humanities, natural language processing and knowledge representation. Relevant applications include the extraction of terms, contexts and lexicosemantic information in specialised domains (Bowker and Pearson, 2002; Melby, 2012; Meyer, 2001), topic modelling and macroanalysis in the humanities (Jockers, 2013; Meeks and Weingart, 2012) and ontology development (Hitzler *et al.*, 2010; Sure *et al.*, 2009).

This paper describes terminology work carried out in the archaeology of al-Andalus with the purpose of creating an ontology-based multilingual terminological resource in the domain. [1] The constitution of a bilingual comparative corpus allowed to pursue a two-folded methodology based on the double dimension of terminology. [2]

In a first moment, the corpus supported the development of OntoAndalus, an ontology of artefacts in al-Andalusian archaeology, based on the interpretation of textual and visual information. OntoAndalus is being developed with the purpose of constituting a language-independent layer to which terms in several languages may refer to in the proposed terminological resource. The case of lighting artefact concepts is presented in this paper.

At a later time, the corpus allowed to identify and extract Portuguese and Spanish terms used by domain specialists. The case of simple and complex terms denoting lighting artefacts is presented in this paper. The former were extracted by means of the Sketch Engine corpus manager and text analysis tool, which further allowed to study the more frequent collocational patterns involving complex terms for domestic lamps. [3] The terminologies in each language were subsequently organised in lexical networks by means of taxonomy and synonymy relations, and several comparisons were drawn between each lexical network. In this regard, the present paper describes the different conceptual motivations for naming domestic lamps in Portuguese and Spanish. An overview of the terms in each language denoting lighting artefact concepts in OntoAndalus is also provided, in order to facilitate future terminological harmonisation in the domain.

1. The archaeology of al-Andalus is an important subdomain of medieval archaeology in Portugal and Spain (Carvajal López, 2014; Covaneiro *et al.*, 2013). "Al-Andalus" is the Arabic name given to the Iberian Peninsula under Islamic rule during the Middle Ages.

2. Terminology is understood in this paper as an interdisciplinary domain concerned with knowledge and its expression, with the purpose of compiling, studying and presenting terms and concepts in specialised fields (NF ISO 704, 2009). As such, terminology integrates a linguistic dimension and a conceptual dimension. These dimensions constitute independent levels of analysis in terminology work: language-specific (i.e. terms and other linguistic expressions) and language-independent (i.e. concepts and other units of knowledge) (Costa, 2013; Roche, 2015; Santos and Costa, 2015).

3. Available from `https://www.sketchengine.eu/`.

Finally, the relationship between linguistic and conceptual information was modelled through Lemon, the Lexicon Model for Ontologies, which is under development by the W3C Ontology-Lexicon Community Group (Cimiano *et al.*, 2016). Lemon, and in particular its core component Ontolex, has been proposed for a number of projects involving linguistic linked open data and the design of web-based terminological resources (Bosque-Gil *et al.*, 2015; Cimiano *et al.*, 2015; Almeida *et al.*, 2016).

2. Background and motivation

The work presented in this paper is motivated by terminological issues noted by Portuguese and Spanish specialists in the archaeology of al-Andalus. Islamic presence in the Iberian Peninsula covered a period of nearly eight centuries (from 711 to 1492 A.D.), having left behind a wide range of materials, such as pottery, architectural fragments, weaponry, jewellery and glassware. The comparison and study of related finds is made possible through the classification of artefacts. Terminology is closely associated with archaeological classification, since artefact categories require terms for identification and communication purposes. The lack of terminology harmonisation has been noted as a hurdle for scholarly communication, while the development of terminological studies has been recognised as a means to acquire and organise knowledge in the domain (Torres *et al.*, 2003). Terminology work in Portugal was inspired by pioneering studies carried out in Spain since the late 1970's, which were focussed on languages such as Spanish, Catalan and Arabic (Coll Conesa *et al.*, 1988; Rosselló-Bordoy, 1991; Rosselló-Bordoy, 1978).

In Portugal, the need to revitalise Islamic pottery studies led to the creation of the CIGA research group.[4] One of the original purposes of this group was to create a database of the more representative instances of pottery artefacts from the Gharb al-Andalus, i.e. the western province of the al-Andalus. In order to facilitate this purpose, the specialists published a Portuguese terminology and classification of entities such as artefact types, shapes, manufacturing and decorative techniques (Bugalhão *et al.*, 2010).

The CIGA group and its ties to terminological studies in Spain are evidence of a need that goes beyond that of harmonising Portuguese terms. It is therefore thought that an ontology-based multilingual terminological resource could help overcome the communication issues noted by the specialists, as well as help furthering the acquisition of knowledge across multiple communities of practice in Portugal and Spain.

4. CIGA is a Portuguese acronym for *Cerâmica Islâmica do Gharb al-Ândalus* (Islamic Pottery of the Gharb al-Andalus). Available from `http://www.camertola.pt/info/ciga`.

3. Related work

In the past, the development of machine-readable terminological resources relied on reading texts and manually extracting information from them. Terminology work has since evolved towards automatic or semi-automatic extraction of linguistic information from corpora (e.g. term candidates, knowledge patterns, contexts of usage) by means of NLP tools, such as concordancers, corpus managers and other text analysis software (Cabré and Palatresi, 2013; Meyer, 2001; Costa, 2001).

On the other hand, computational terminologies have evolved from simple termbases to full-fledged knowledge-based resources which, besides providing information about the terms used in a specialised domain, are informative of the underlying conceptual structure of the domain itself (Meyer *et al.*, 1992; Nazarenko and Hamon, 2002; Faber *et al.*, 2014; Condamines, 2018). In more recent years, the development of Semantic Web and Linked Data technologies, as well as further research in applied ontology, led to ontology-based approaches to the creation of terminological resources. In the latter, a fundamental distinction is placed between the domain ontology, through which language-independent knowledge is modelled, and the lexical network(s) representing language-specific information about terms and other linguistic units (Roche, 2012).

While corpora are today paramount for terminology work, textual approaches to terminology often ignore (or explicitly reject) the distinction between knowledge and language as distinct levels of analysis, which may lead to several misunderstandings in multilingual terminology work, such as conflating language-specific relations at the term level (e.g. hyponymy, meronymy) with relations drawn at the concept level (e.g. subsumption, part-whole) (L'Homme, 2004; Condamines, 2018). On the contrary, the approach described in this paper explicitly distinguishes between the domain ontology and the Portuguese and Spanish networks of terms, while relating both linguistic and conceptual dimensions of terminology work in an effort towards building an ontology-based terminological resource in the Semantic Web. Here, the role of NLP tools is firmly placed in the linguistic dimension of terminology work, where they excel in extracting language-specific information about terms and other linguistic units, which can then be related to the domain ontology by means of a specific model (in this case, Lemon).

This brings us to the matter of related work in archaeology. To the best of our knowledge, there are no ontology-based terminological resources in our domain of interest, nor in the wider field of archaeological typologies of artefacts. With regard to ontology development, the CIDOC-CRM has become relevant for documenting archaeological data following the ARIADNE project and the initial proposal of the CRMarchaeo extension (Doerr, 2014). However, these initiatives remain focussed on archaeological excavation. This motivated the development of OntoAndalus, a domain ontology focussed on artefacts in al-Andalusian archaeology, which will be briefly described in the following section.

4. OntoAndalus: an ontology of artefacts in OWL

OntoAndalus is an ontology of relevant artefacts in the archaeology of al-Andalus. It was developed as part of a PhD thesis (Almeida, 2019) and is presently made available under a Creative Commons License (CC-BY-4.0). [5] The development process was based on the interpretation of selected texts from a Portuguese and Spanish specialised corpus, as well as English reference works on archaeology. The more specialised texts consist of Portuguese and Spanish conference papers, journal articles, theses and monographs on the description, classification and terminology of Islamic artefacts (mostly pottery). OntoAndalus is based on the so-called "functional form" criterion of classification of the artefacts, which is followed in the artefact typologies of the CIGA group and Rosselló-Bordoy (Bugalhão *et al.*, 2010; Rosselló-Bordoy, 1991; Rosselló-Bordoy, 1978).

OntoAndalus was developed using the Protégé ontology-editor. [6] Protégé integrates a host of tools and plugins which are invaluable for the modelling process and for visualisation purposes (e.g. plugins for generating conceptual graphs). OWL was chosen as a modelling language due to its relative simplicity and status as a W3C recommendation (W3C OWL Working Group, 2012). DOLCE+DnS Ultralite (DUL) was chosen as the foundational ontology for the development of OntoAndalus. DOLCE was one of the first notable top-level ontologies following the development of applied ontology as a research field (Guarino and Musen, 2005; Munn and Smith, 2008). [7] DUL is a streamlined version of DOLCE-Lite, the original translation of DOLCE into OWL, based on ontology design patterns (Gangemi, 2016). The latter enable the reuse of smaller ontological components in order to more efficiently solve recurrent modelling problems (e.g. physical objects, events). Besides streamlining the original translation of DOLCE into OWL, DUL also integrates the Descriptions and Situations ontology (DnS) for modelling social and cognitive entities (e.g. information objects). Another advantage of DUL lies in its complete availability in OWL format, including all classes and binary relations. [8]

As of the writing of this paper, OntoAndalus consists of 161 classes, 30 object properties and 135 individuals (excluding the elements already defined in DUL). The ontology includes 72 artefact types, which are organised in the following categories: lighting artefacts, tableware, kitchenware, domestic artefacts, recreational artefacts, ritual artefacts, agricultural artefacts, construction artefacts, artisanal artefacts, storage artefacts, transportation artefacts and artefact components.

5. OntoAndalus is the topic of a forthcoming paper (Almeida and Costa, 2019). The ontology is made available through `https://github.com/brunoalmeida81/OntoAndalus`.

6. Available from `https://protege.stanford.edu`.

7. "DOLCE" is an acronym for Descriptive Ontology for Linguistic and Cognitive Engineering (Masolo *et al.*, 2003). Although slightly outdated, Mascardi *et al.* (2007) provide a good overview and comparison of similar top-level ontologies.

8. Available from `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`.

4.1. *Modelling lighting artefacts*

In this section, the category of lighting artefacts will be described as an example of how artefact types are modelled in OntoAndalus. This category includes some of the more representative artefacts of the Islamic period in the Iberian Peninsula (Gómez Martínez, 2004).

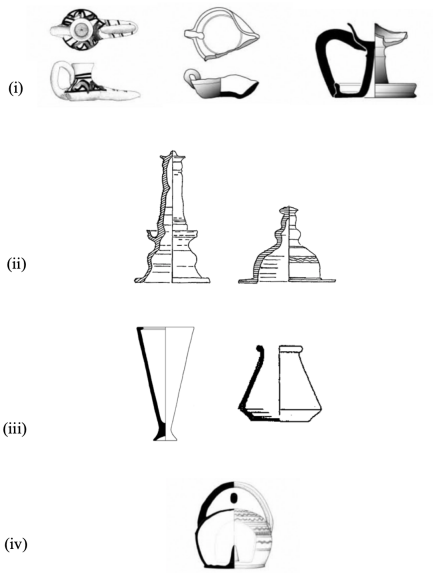

Figure 1. *Graphical representation of types of lighting artefacts.*

According to Gómez Martínez (2004), archaeologists usually acknowledge the existence of four kinds of lighting artefacts, denoted by the Spanish terms (i) *candil*; (ii) *policandela* (or *almenara*); (iii) *lamparilla*; and (iv) *fanal*. Figure 1 shows graphical representations of typical instances of each series. [9]

In order to maintain consistency with DUL and facilitate international communication, each class in OntoAndalus has an English identifier, namely: (i) `Lamp`; (ii) `MultipleLamp`; (iii) `StationaryLamp`; and (iv) `Lantern`. The formal definitions of these classes put forward in OntoAndalus adhere to the following pattern: **superordinate class + collection + function + part(s) or component(s)**. The following paragraphs explicate each class through natural-language definitions derived from formal definitions in the ontology, along with references to relevant texts in the corpus.

9. These illustrations are reproduced from Bugalhão *et al.* (2010, p. 471), Rosselló-Bordoy (1991, p. 174), Vallejo Triano and Escudero Aranda (1999, p. 165) and Gómez Martínez (2000, p. 433).

`Lamp` (*candil*$_{es}$) *Def.* Artefact for lighting in closed spaces composed by at least one spout and a single chamber for liquid fuel (Coll Conesa *et al.*, 1988; Gómez Martínez, 2004; Rosselló-Bordoy, 1991).

`Multiple lamp` (*almenara*$_{pt}$, *almenara*$_{es}$, *policandela*$_{es}$). *Def.* Artefact for stationary lighting in closed spaces composed by more than one chamber for liquid fuel unified by a structure (Gómez Martínez, 2004; Rosselló-Bordoy, 1991).

`Stationary lamp` (*lamparilla*$_{es}$). *Def.* Artefact for stationary lighting in closed spaces composed by a single chamber for liquid fuel (Vallejo Triano and Escudero Aranda, 1999).

`Lantern` (*fanal*$_{pt}$, *lanterna*$_{pt}$; *fanal*$_{es}$, *linterna*$_{es}$). *Def.* Artefact for lighting in open spaces composed by a single chamber for solid fuel (Bugalhão *et al.*, 2010; Gómez Martínez, 2004).

The `Lamp` class is the more complex part of the ontology with regard to lighting artefacts (figure 2). OntoAndalus includes four criteria of subdivision of `Lamp`: (i) vessel form, (ii) type of spout, (iii) inclusion of a discus or neck and (iv) inclusion of a tall foot. The multiple criteria of subdivision are represented through pairwise disjoint defined classes.

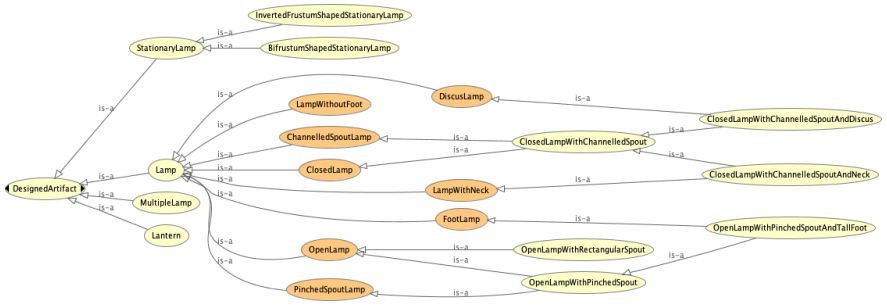

Figure 2. *Types of lighting artefacts in OntoAndalus.*

The vessel form (i.e. open or closed) is generally considered to be the more salient criterion for distinguishing between types of lamps (Bugalhão *et al.*, 2010; Gómez Martínez, 2004). This was chosen as the primary criterion of subdivision of the `Lamp` class. Therefore, only the disjoint classes `ClosedLamp` and `OpenLamp` are further subdivided in the asserted hierarchy of OntoAndalus, while the remaining classification can be inferred by a reasoner.

We have seen how artefact types can be modelled in OntoAndalus in order to constitute the language-independent layer of an ontology-based terminological resource. This includes formal descriptions and definitions of classes, which guide the drafting of natural-language definitions. The following section will describe the work carried out with regard to the extraction and representation of Portuguese and Spanish terms

from specialised corpora, as well as the relationship between conceptual and linguistic aspects.

5. Extracting and modelling information at the term level: the case of lighting artefacts

5.1. *Criteria for corpus design*

The purpose of constituting a corpus was twofold with regard to the work described in this paper. On the one hand, the corpus facilitated the understanding of the domain and assisted the modelling of OntoAndalus. On the other hand, the corpus was also paramount for compilling terminological and lexicosemantic information. The criteria for corpus design represented in table 1 were adopted based on Bowker and Pearson (2002) and Cabré (2008).

Domain	Pottery of the al-Andalus
Language	Portuguese, Spanish
Time period	1970 and later
Level of specialisation	Medium to high
Text integrity	Full texts
Medium and modality	Digitised written texts with visual information
Text genres and discourse types	Heterogenous

Table 1. *Criteria for corpus design.*

The texts constituting the corpus are primarily about al-Andalusian artefacts and were originally written in Portuguese or Spanish. [10] The time period is justified by the fact that the archaeology of al-Andalus only established itself as a domain from the 1970's onward. The corpus includes full texts with a medium to high level of specialisation, i.e. produced by domain specialists and intended for actual or future specialists of the domain. This implies the exclusion of works intended for the general audience, since the terminology used therein could be significantly different from that employed in specialised discourse (Bowker and Pearson, 2002).

Regarding the medium and modality, the corpus is formed by written texts with visual information. There are several reasons for the latter requirement: (i) the importance of the visual modality in the domain, (ii) facilitating the understanding of the

10. Following Costa (2001), a specialised text is understood here as a stable linguistic product resulting from the professional activity of experts within a specialised community. Specialised texts are, therefore, constrained by social and communicative conditions which determine, for example, their boundaries and internal structure.

domain (e.g. artefact kinds, vessel morphology) and (iii) future work (e.g. enrichment of a future terminological resource with visual information). [11]

Lastly, the corpus is heterogenous with regard to genres and discourses. It ranges over transdisciplinary discourses (e.g. archaeology, history) as well as multiple text genres (e.g. thesis, journal articles). The justification for this option lies in our focus on domain knowledge, which can be expressed across several genres and discourses.

5.2. *Structure and composition of the corpus*

A simple structure was devised for our corpus according to language and text genre. This was based on previous work carried out by Costa (2001). A review of the scholarly communication of the domain allowed to identify the more significant text genres, namely: (i) theses and dissertations, (ii) monographs, (iii) catalogues, (iv) articles in scientific journals, (v) papers presented in conferences and other events.

Although the proposed typology is methodologically useful, it does present several issues. For one, the same text (or a very similar version) may range over more than one genre. For example, the content of a thesis may be later published as a monograph or even as a series of monographs. This is the case of the doctoral thesis of Rosa Varela Gomes, entitled *Silves (Xelb): uma cidade do Gharb al-Andalus: arqueologia e história (séculos VIII-XIII)*. This thesis, originally presented in 2000, was later published in four volumes of the monograph series *Trabalhos de arqueologia*, which are included in the corpus (Gomes, 2003; Gomes, 2011; Gomes, 2006; Gomes, 2002).

A further difficulty was posed by museum and exhibition catalogues. These often include articles by several authors along with the actual catalogue description of the artefacts. For this reason, it was decided to include a separate category for catalogue articles in the structure of the corpus. The descriptions themselves were not included in the final typology, since we did not have access to digitised catalogues of artefacts in both languages. It is common, however, for other genres to include catalogues in a separate section. The discourse of catalogue descriptions is, therefore, present in the corpus. [12]

Tables 2 and 3 show the composition of the Portuguese and Spanish corpora. The data shown was gathered from Sketch Engine, the corpus manager used in the work carried out in this paper. Both corpora are comparable in terms of size: the Portuguese

11. The visual modality plays an important role in archaeology. While photography allows for a realistic depiction of an object, drawing is paramount in conveying selective (or *diagrammatic*) information about an artefact (Adkins and Adkins, 2009; Caballero Zoreda, 2006). With regard to terminology work, the importance of visual information in corpora design and terminological resources has already been noted (Prieto Velasco and Faber, 2012).

12. For example, the monograph of Rosselló-Bordoy (1978) about the classification and terminology of Majorcan pottery includes a catalogue in the final section. This work is available in an electronic format, which facilitated its inclusion in the corpus.

Text genre	**Texts (N)**	**Tokens (N)**	**Word-forms (N)**	**Word-forms (%)**
Theses, dissertations	6	324,817	242,597	36.20
Monographs	5	485,225	362,401	54.08
Journal articles	6	54,171	40,458	6.04
Catalogue articles	1	6,789	5,070	0.76
Conference papers	4	26,257	19,610	2.93
Total	22	897,259	670,136	≈100

Table 2. *Composition of the Portuguese corpus.*

Text genre	**Texts (N)**	**Tokens (N)**	**Word-forms (N)**	**Word-forms (%)**
Theses, dissertations	3	747,865	575,011	81.45
Monographs	1	76,037	58,462	8.28
Journal articles	7	66,684	51,271	7.26
Catalogue articles	1	3,977	3,057	0.43
Conference papers	4	23,604	18,148	2.57
Total	16	918,167	705,949	≈100

Table 3. *Composition of the Spanish corpus.*

corpus has over 670,000 word-forms while the Spanish corpus has over 705,000 word-forms. [13]

There is some asymmetry with regard to text genre: monographs are the more represented genre in the Portuguese corpus, while theses and dissertations are predominant in the Spanish corpus. Furthermore, all of the Portuguese texts included in the latter category are master's dissertations, while the Spanish texts are doctoral theses. These circumstances are compensated by the fact that four of the five monographs in the Portuguese corpus correspond to the doctoral thesis of Rosa Varela Gomes.

13. Sketch Engine distinguishes word-forms from tokens. The former are the several forms assumed by lexemes (i.e. the English verb "to go" has the word-forms "go", "went", "gone"), while the latter are instances of word-forms and punctuation symbols.

5.3. *Extraction of linguistic information from the corpora*

The NLP part of our work was carried out using Sketch Engine, a corpus manager and text analysis tool, following the constitution of a Portuguese and Spanish corpus of specialised texts.

The first stage of this consisted in extracting frequency lists of word-forms in both languages and selecting simple terms (i.e. with a single root) for our case study of lighting objects. Only words occurring more than once in a single text were considered for analysis.

A further stage consisted in extracting complex terms, which denote subtypes of the above-mentioned artefact types. The extraction was carried out based on the collocational strength between the head-word and its modifiers. A "collocate", in this context, describes any co-occurring words within a specified pattern. These collocates were automatically ordered by Sketch Engine according to their logDice score, which measures the collocational strength between two words in a corpus (Rychlý, 2008). LogDice has a theoretical maximum of 14, in which every instance of X in a corpus co-occurs with Y (and vice-versa). A score above 10 represents a strong collocation. Only collocates with a logDice score above 10 were considered for analysis.

The examples of complex terms presented in this paper adhere to the following patterns, which involve the more relevant collocates in the corpus (N = noun, P = preposition, A = adjective):

– `candil:N de:P N A?` (for both languages);

– `candeia:N de:P N A?` (only relevant in the case of Portuguese).

5.4. *The names of lighting artefacts in Portuguese*

Table 4 shows the Portuguese simple terms for lighting artefacts employed in the corpus.

The `Lamp` concept and its subordinates are paramount for the linguistic expression of lighting artefacts. There are, however, some issues regarding more generic lighting artefacts concepts in Portuguese. In one text of the corpus, *luminária* denotes lighting artefacts in general. Other generic terms are *lâmpada* and *candeeiro*, although the latter is much less relevant in the corpus. *Lâmpada* seems to approximate the `Lamp` concept more clearly, distinguishing it from *lanterna* and other terms denoting domestic or non-domestic lighting artefacts.

A dichotomy is established between closed and open lamps with, respectively, *candil* and *candeia*. This is in line with the terminological proposals put forward by several archaeologists (Bugalhão *et al.*, 2010; Torres *et al.*, 2003).

Term	Frequency	Texts
candil	551	15
lucerna	495	13
candeia	217	9
lamparina	68	8
luminária	67	1
lâmpada	10	2
lanterna	7	3
vela	5	1
tocha	3	1
fanal	2	2
candeeiro	2	1

Table 4. *Simple terms for lighting artefacts in Portuguese.*

The `candil:N de:P N A?` pattern provides an insight of how possible types of lamp are named in Portuguese, most notably the `ClosedLamp` concept. Table 5 shows the more significant noun collocates in the corpus.

Collocate	Frequency	logDice
disco	19	11.97
bico	17	11.3
depósito	7	10.45
pé	10	10.43

Table 5. *Noun collocates following* candil:N de:P *in the Portuguese corpus.*

As we can see, these collocates denote parts of the artefacts, represented by the `Discus` (*disco*), `Spout` (*bico*), `LampFuelChamber` (*depósito*) and `Foot` (*pé*) concepts in OntoAndalus. The collocates are present in the following complex terms in the corpus:

– *candil de disco impresso* (19 occurrences);

– *candil de bico* (17 occurrences);

– *candil de depósito aberto* (7 occurrences);

– *candil de pé alto* (10 occurrences).

Candil de disco impresso denotes an established type of lamp. [14] The second collocate, *bico*, occurs in the terms *candil de bico*, *candil de bico comprido* and *candil de bico curto*. These expressions are used to distinguish between closed lamps based on salient characteristics, namely the discus and spout.

14. This type, characterised by the discus surrounding the pouring hole (Zozaya, 1999), is represented in our ontology through the `DiscusLamp` concept.

With one exception, *depósito* and *pé* only collocate with *candil* when the *candil/candeia* dichotomy is not followed. These nouns take part in the terms *candil de depósito aberto* and *candil de pé alto*, respectively. The exception is a context where *candil de pé alto* is explicitly rejected in favour of *candeia de pé alto*.

The `candeia:N de:P N A?` pattern is equally informative in Portuguese. As we can see in table 6, the more relevant collocates also denote parts, which are represented by the `Foot` (*pé*) and `LampFuelChamber` (*depósito*/*câmara*) concepts in OntoAndalus. Contrary to the case of *candil*, *candeia* is used almost exclusively to denote an open lamp.

Collocate	Frequency	logDice
pé	25	12.05
depósito	10	11.41
câmara	4	10.2

Table 6. *Noun collocates following* candeia:N de:P *in the Portuguese corpus.*

These word-forms are employed in the following complex terms:

– *candeia de pé* (25 occurrences);

– *candeia de depósito aberto* (10 occurrences);

– *candeia de câmara aberta* (4 occurrences).

The first collocate, *pé*, occurs in the terms *candeia de pé* (5 occurrences) and *candeia de pé alto* (20 occurrences). These expressions are used indiscriminately in the corpus to refer to the same type of artefact, which is represented by the `FootLamp` concept in our ontology. This seems to indicate that the foot of these lamps is always 'tall' when compared to similarly-sized artefacts. A proposed term for the appendage of this type of lamp is *pé alto sobre prato de sustentação*, as opposed to *pé alto maciço*, which characterises other kinds of artefacts (Bugalhão *et al.*, 2010). *Candeia de pé* is, therefore, an abbreviation of longer and more precise terms.

The remaining collocates of *candeia* denote the fuel chamber of the artefacts. These motivate the seemingly redundant expressions *candeia de câmara aberta* and *candeia de depósito aberto*. Both expressions highlight the open fuel chamber as a distinctive quality of these artefacts.

5.5. *The names of lighting artefacts in Spanish*

The `Lamp` concept and its subordinates are also predominant naming-wise in Spanish (table 7). Contrary to Portuguese, *candil* is consensually used to denote both open and closed forms of al-Andalusian artefacts. *Lámpara* and *luminaria* are the more employed generic terms for lighting artefacts. The remaining terms apply to less studied or controversial artefact kinds (i.e. *policandela*, *fanal*, *linterna*, *lamparilla*, *almenara*) and artefacts emanating from the Roman period (i.e. *lucerna*).

Term	Frequency	Texts
candil	489	15
lámpara	25	5
fanal	22	2
lucerna	14	7
almenara	14	3
lamparilla	13	3
linterna	7	3
policandela	7	2
luminaria	3	1
candelabro	2	3

Table 7. *Simple terms for lighting artefacts in Spanish.*

As in the case of Portuguese, the `candil:N de:P N A?` pattern is of import for types of domestic lamps. As we can see in table 8, the noun collocates denote partitive concepts, namely `Spout` (*piquera*), `Foot` (*pie*), `Discus` (*disco*) and `LampFuelChamber` (*cazoleta* and *depósito*).

Collocate	Frequency	logDice
piquera	58	12.41
pie	32	11.76
disco	22	11.45
cazoleta	26	11.40
depósito	11	10.39

Table 8. *Noun collocates following* candil:N de:P *in the Spanish corpus.*

The more frequent complex terms containing these word-forms are the following:

– *candil de piquera* (57 occurrences);

– *candil de pie alto* (30 occurrences);

– *candil de disco impreso* (18 occurrences);

– *candil de cazoleta abierta* (9 occurrences);

– *candil de depósito abierto* (9 occurrences).

The first term denotes a thoroughly studied type of lamp in al-Andalusian archaeology (Zozaya, 2007). Although *piquera* may be used for denoting any kind of spout for holding a wick, it is often used for denoting the characteristic spout of closed lamps (i.e. the `ChannelledSpout` concept). The term *pellizco*, which only co-occurs once with *candil* in the corpus, denotes the spout that is typical of open lamps (i.e. the `PinchedSpout` concept). The term *piquera de pellizco*, with 18 occurrences in 4 texts of the corpus, is however a more precise denomination, since it makes clear the reference to a part (i.e. the pinched spout).

However, there does not seem to be any parallel for spouts applied to closed lamps. Instead, *piquera* is further modified by adjectives (e.g. *larga*, *corta*) or prepositional phrases (e.g. *de entronque suave*, *de quilla de barco*, *de bañera*), all of which denote sizes or shapes of the spout used in closed lamps.

While the spout is the more salient characteristic for referring to a closed lamp, the fuel chamber is predominant for reference to an open lamp. This is the more likely explanation for the fact that *candil de pellizco* and *candil de cazoleta/depósito cerrada(o)*, which follow the opposite motivation, are less relevant in the corpus.

5.6. *Representation and comparison of the Portuguese and Spanish terms*

The Portuguese and Spanish terms can be represented through lexical networks. The latter are prominent devices for representing language-specific information, which may be used for creating a concept-based terminological resource (Santos and Costa, 2015).

Figures 3 and 4 show the more relevant terms in Portuguese and Spanish. Terms motivated by the same criteria of subdivision (e.g. vessel form) are represented in the graphs through divided taxonomic arcs (e.g. *candil* and *candeia* in Portuguese).

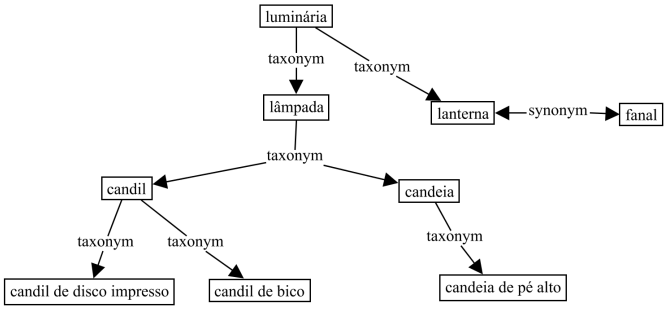

Figure 3. *Lexical network of lighting artefacts in Portuguese.*

The lexical relations employed in both networks are taxonomy, a specialisation of the hyponymy relation (Cruse, 1986), and synonymy. The latter is restricted here to what is assumed to be absolute synonymy. Including the relation of quasi-synonymy in the graphs would imply a much more complex network, since any two terms denoting closely related concepts could be considered as quasi-synonyms.

The generic terms *luminária*$_{pt}$/*luminaria*$_{es}$ and *lâmpada*$_{pt}$/*lámpara*$_{es}$ pose several difficulties. We have followed the assumption that the former are superordinates of the latter. While these terms are not directly relevant for the Islamic period, we have included them in the networks to clarify that *candil* can be seen as a subordinate term of *lâmpada*$_{pt}$/*lámpara*$_{es}$ in both languages. This leads us to argue that *lucerna*$_{pt}$/*lucerna*$_{es}$ are not synonyms of *candil*$_{pt}$/*candil*$_{es}$ in neither language. In-

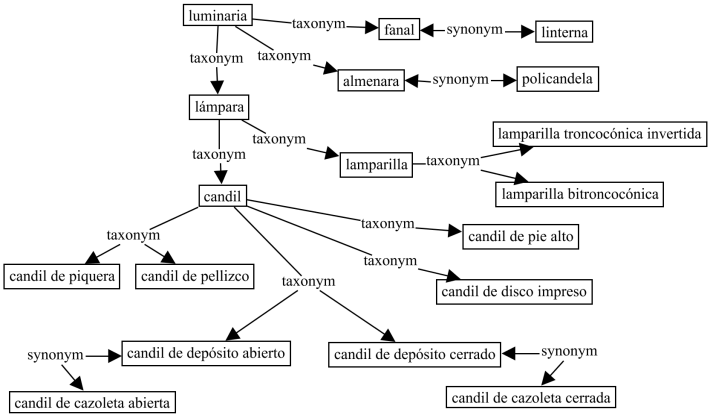

Figure 4. *Lexical network of lighting artefacts in Spanish.*

stead, they are subordinates of *lâmpada*$_{pt}$/*lámpara*$_{es}$: while *lucerna*$_{pt}$/*lucerna*$_{es}$ denote an artefact kind of the Classical period, *candil*$_{pt}$/*candil*$_{es}$ denote a related artefact kind of the Islamic period (as assumed by most specialists).

As we can see in the graphs, the situation is markedly different in both languages. A significant difference in the networks lies in the *candil*/*candeia* dichotomy in Portuguese. As pointed out before, *candil* is used in Spanish to denote both open and closed lamps, whereas in Portuguese it only denotes closed lamps.

A further difference lies in the use of criteria of subdivision, which is more apparent in Spanish. In Portuguese, the most obvious case is the *candil*/*candeia* dichotomy, which is based on the overall form of the vessels. The term *candil de bico* is motivated by the spout, but there are no converse terms based on this criterion in the corpus. Furthermore, its adequacy as a term is doubtful, since every kind of lamp in the domain should have a spout of some sort for holding the wick in place. *Candil de disco impresso*, on the other hand, is motivated by the portion of the chamber surrounding the orifice, which – in this kind of closed lamp – has a discus instead of a neck. Finally, the presence of an applied foot is the only productive criterion in the case of open lamps in Portuguese (i.e. *candeia de pé alto*).

In Spanish, the chamber is also an important criterion, as attested by *candil de depósito abierto*/*cerrado* and its respective synonyms *candil de cazoleta abierta*/*cerrada*. Other denominations are based on the spout, namely *candil de piquera* (57 occurrences) and *candil de pellizco* (1 occurrence). Furthermore, there is the already pointed out ambiguity in interpreting *piquera*, since *piquera de pellizco* is present in the corpus (cf. section 5.5). Finally, there are terms motivated by the presence or absence of a discus and applied foot: *candil de disco impreso* and *candil de pie alto*. A summary of the terms in both languages is shown in table 9.

Criteria of subdivision	**Portuguese term**	**Spanish term**
Chamber	*candil* *candeia*	*candil de depósito (cazoleta) cerrado(a)* *candil de depósito (cazoleta) abierto(a)*
Spout	*candil de bico* (?)	*candil de piquera* *candil de pellizco* (?)
Discus / neck	*candil de disco impresso*	*candil de disco impreso*
Foot	*candeia de pé alto*	*candil de pie alto*

Table 9. *Terms for lamps according to different criteria of subdivision.*

With regard to the other kinds of lighting artefacts, only the lantern is represented in both languages through the archaism *fanal*. The latter is, however, used more extensively in the Spanish corpus, while *lanterna* is the preferred denomination in the Portuguese corpus. *Almenara* and *policandela* are used interchangeably for denoting the same artefact kind in Spanish. Subtypes of the stationary frustum-shaped lamps are denoted in Spanish using adjective modifiers. The latter denote approximate geometrical shapes (i.e. *bitronconcónica* and *troncocónica invertida*).

5.7. *Relationship between linguistic and conceptual information*

We have shown how language-specific information can be represented by means of lexical networks. This brings into question the relationship between the lexical networks and OntoAndalus. Making this relationship explicit allows to contrast both languages with regard to the conceptualisation of the domain.

Table 10 summarises the relationship between the more established terms in the corpus and the concepts of lighting artefacts in OntoAndalus. The table highlights the asymmetry in both languages in expressing concepts from this section of the ontology.

Finding Portuguese denominations for the `Lamp` concept as well as for the artefact types not studied by the Portuguese archaeologists remain open questions. As shown in the Portuguese lexical network, the `Lamp` concept may remain unnamed in this language, since the generic term *lâmpada* is a suitable hypernym of *candil* and *candeia*. The second issue can be more problematic. A possible solution is using the terms *candeeiro* and *lamparina* for, respectively, the `MultipleLamp` and `StationaryLamp` concepts in our ontology, since both terms are already present in the corpus. Another possibility is to use the term *almenara* for the latter concept since the term exists in both languages (Gómez Martínez, 2004). It is, however, not represented in the Portuguese corpus.

Concept	Portuguese term	Spanish term
Lamp		*candil*
ClosedLamp	*candil*	*candil de depósito (cazoleta) cerrado(a)*
ClosedLampWithChannelled Spout		
ClosedLampWithChannelled SpoutandDiscus		
OpenLamp	*candeia*	*candil de depósito (cazoleta) abierto(a)*
OpenLampWithPinchedSpout		
OpenLampWithPinchedSpout andTallFoot		
OpenLampWithRectangular Spout		
ChannelledSpoutLamp	*candil de bico* (?)	*candil de piquera*
PinchedSpoutLamp		*candil de pellizco* (?)
FootLamp	*candeia de pé alto*	*candil de pie alto*
DiscusLamp	*candil de disco impresso*	*candil de disco impreso*
Lantern	*fanal* *lanterna*	*fanal* *linterna*
MultipleLamp	*almenara* *policandela*	
StationaryLamp	*lamparilla*	
BifrustumShaped StationaryLamp	*lamparilla bitroncocónica*	
InvertedFrustumShaped StationaryLamp	*lamparilla troncocónica invertida*	

Table 10. *Relationship between concepts of lighting artefacts and their terms.*

In this section, extracted term candidates in each language were described, which highlighted several inconsistencies in each lexical network in relation to the domain ontology. This is but one step of the overall process, since the extracted data would require expert validation before it can be included in a future terminological resource, in order to follow a quality-based approach to terminology management (Silva, 2014). Future work involving domain specialists will be carried out with regard to validation and terminology harmonisation, in which quantitative and/or qualitative methods, such as surveys or focus groups, will be employed.

5.8. *Modelling linguistic information with Lemon*

Lemon is an acronym for "Lexicon Model for Ontologies." The purpose of the model is to provide a linguistic grounding for computational ontologies. Most notably,

Lemon can be used to represent how ontology elements (e.g. classes, object properties, instances) are expressed in natural language.

Lemon consists of the following modules:

– *Ontolex*. It allows to establish an interface between a lexicon and an ontology;

– *Synsem*. It allows to represent information at the syntactic and semantic levels;

– *Decomp*. It allows to represent information on the decomposition of complex expressions;

– *Vartrans*. It allows to represent information regarding variation and translation;

– *Lime*. It allows to represent linguistic metadata.

Ontolex, the core module of Lemon, is especially important, since it allows to establish a relationship between a lexicon (or terminology) and an ontology. Lemon is structured around lexical entries, which are either single words, multiword expressions or affixes. A lexical entry is realised as a series of forms in a language. In Lemon, each entry needs to be linked to at least one form and, at most, to one canonical form. The latter is typically the lemma of a lexical entry in a dictionary. Each form may have written and/or phonetic representations.

There are several ways for relating a lexical entry to an ontology element. The relationship can be established directly through the `denotes/isDenotedBy` object properties. Another option is to establish a mediated link through the "lexical sense" construct. The latter allows to model the fact that a single lexical entry may have several distinct senses in a language, as traditionally represented in dictionaries. For example, "consumption" in English may be used in the everyday sense of "act of consuming" or in several specialised senses (e.g. in economics). Each of these senses may be linked to different ontology elements via the `reference/isReferenceOf` properties. Lemon also allows to represent pragmatic information regarding the usage of lexical entries. For example, the different senses of the French words *rivière* and *fleuve* may be clarified, although both of them can still point to the same ontology element, if this is deemed useful in support of a particular modelling decision.

Lemon further introduces the "lexical concept". The latter allows to model a unit of thought or collection of senses which are not directly represented in an ontology. A lexical concept may be associated to a lexical entry and to an ontology element via the available object properties. It may also be associated with a natural language definition through the `skos:definition` property. Finally, concept sets may be defined in order to organise a lexicon according to the concept (i.e. onomasiologically).

In the present case, language-specific information can be represented using Lemon while conceptual information is left in OntoAndalus. The former consists of information at the term level, including grammatical information and lexicosemantic relations. Conceptual information, on the other hand, pertains to domain knowledge. This approach allows to distinguish between the linguistic and conceptual dimensions of terminology work while still drawing relationships between each dimension.

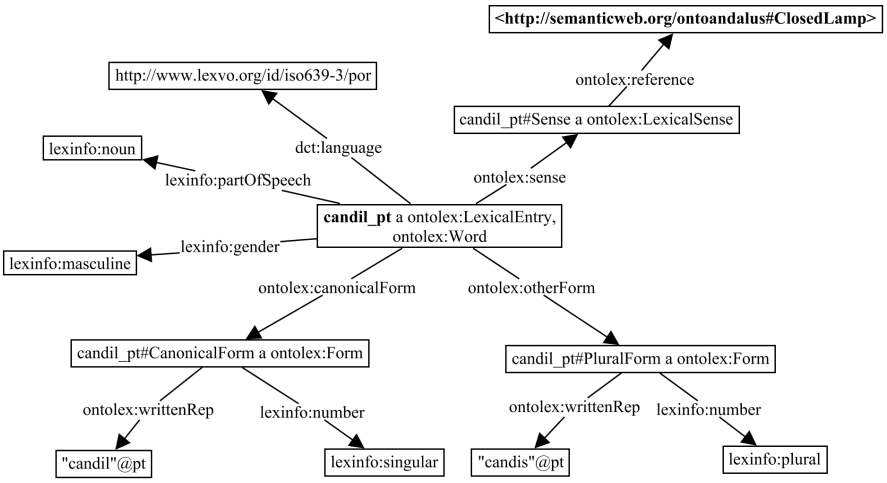

Figure 5. *Term entry in Lemon.*

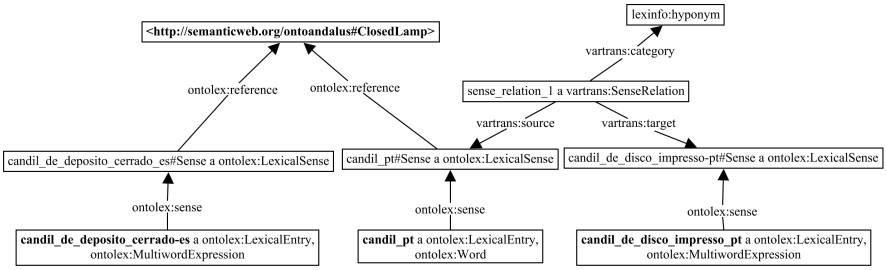

Figure 6. *Representing lexicosemantic relations in Lemon.*

Figure 5 shows a possible term entry for the Portuguese term *candil*. Relevant linguistic information includes the language, part of speech and grammatical gender. As recommended in Lemon, this information is represented via Dublin Core metadata and the LexInfo model. [15] Canonical and plural forms are provided along with their respective written representations.

The links between terms and the predicates of OntoAndalus are mediated by the lexical sense construct. This approach is required for the representation of lexicosemantic relations at the term level in each language. As we can see in figure 6, the equivalence between *candil*$_{pt}$ and *candil de depósito cerrado*$_{es}$ can be represented simply by pointing both lexical senses to the same class in OntoAndalus.

15. Available from, respectively, `http://dublincore.org/documents/dcmi-terms` and `https://www.lexinfo.net`.

Since LexInfo does not include the relation of taxonomy, it has been replaced in the graph above with the broader hyponymy relation. In Lemon, the relation is established by indicating the source and target terms as well as the respective category in the LexInfo model. In this example, the Portuguese term *candil de disco impresso* is asserted as a hyponym of *candil*.

The operationalisation of this data requires its expression through a suitable formalism. The following RDF code in Turtle syntax represents grammatical and semantic information about the Portuguese term *candil*:

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .

:candil_pt a ontolex:LexicalEntry, ontolex:Word ;
    dct:language <http://www.lexvo.org/id/iso639-3/por> ;
    lexinfo:partOfSpeech lexinfo:noun ;
    lexinfo:gender lexinfo:masculine ;
    ontolex:canonicalForm :candil_pt#CanonicalForm ;
    ontolex:otherForm :candil_pt#PluralForm ;
    ontolex:sense :candil_pt#Sense .

:candil_pt#CanonicalForm a ontolex:Form ;
    ontolex:WrittenRep ‘‘candil’’@pt ;
    lexinfo:number lexinfo:singular .

:candil_pt#PluralForm a ontolex:Form ;
    ontolex:WrittenRep ‘‘candis’’@pt ;
    lexinfo:number lexinfo:plural .

:candil_pt#Sense a ontolex:LexicalSense ;
    ontolex:reference <http://semanticweb.org/ontoandalus#ClosedLamp> .

:senseRelation1 a vartrans:SenseRelation ;
    vartrans:source :candil_pt#Sense ;
    vartrans:target :candil_de_disco_impresso_pt#Sense ;
    vartrans:category lexinfo:hyponym .
```

As we can see, Lemon enables the representation of diverse information at the term level. Grammatical information includes the gender, part of speech and singular and plural forms of the term. Semantic information includes reference to a class in OntoAndalus as well as the hyponymy relation between the Portuguese terms *candil* and *candil de disco impresso*.

6. Conclusion

This paper presented our work towards multilingual terminological resource aimed at experts and students of the archaeology of al-Andalus. OntoAndalus provides a language-independent conceptualisation of the domain, which can be shared across multiple communities of practice, while the language-specific components can be represented with Lemon, a model for the linguistic grounding of computational ontologies. The constitution and subsequent analysis of the corpus with Sketch Engine was paramount for representing both language-specific and language-independent information.

The case of lighting artefacts was highlighted in this paper, starting with the conceptualisation of these artefact types in OntoAndalus and leading to the extraction and representation of their Portuguese and Spanish terms, with a special emphasis on complex terms derived from collocational patterns in the corpus. Lemon provides the necessary means for the representation of rich grammatical and semantic information on the terms, including lexicosemantic relations and reference to ontology elements. The approach described in this paper, therefore, is able to distinguish between the linguistic and conceptual dimensions of terminology work while drawing useful relationships between each dimension. Combining methods from NLP and ontology engineering allowed to better meet the needs of digital humanities research, in particular in the archaeology of al-Andalus. Such an approach helps the archaeologist to clearly differentiate matters pertaining to the terms used in each language from matters pertaining to domain knowledge, which may help guide future initiatives in terminology harmonisation as well as facilitate the dissemination of knowledge for research and educational purposes.

Acknowledgements

Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2019.

7. References

Adkins L., Adkins R., *Archaeological illustration*, University Press, Cambridge, 2009.

Almeida B., Terminology and knowledge representation: ceramic artefacts of al-Andalus, Doctoral thesis, Universidade NOVA de Lisboa, 2019.

Almeida B., Costa R., "OntoAndalus: an ontology of Islamic artefacts for terminological purposes", 2019, Manuscript submitted for publication.

Almeida B., Roche C., Costa R., "Terminology and ontology development in the domain of Islamic archaeology", *in* H. Thomsen, A. Pareja-Lora, B. Madsen (eds), *Term bases and linguistic linked open data: TKE 2016*, Copenhagen Business School, Copenhagen, p. 147-156, 2016.

Bosque-Gil J., Gracia J., Aguado-de Cea G., Montiel-Ponsoda E., "Applying the OntoLex Model to a multilingual terminological resource", *in* F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker, A. Zimmermann (eds), *The Semantic Web: ESWC 2015 Satellite Events*, Springer International Publishing, p. 283-294, 2015.

Bowker L., Pearson J., *Working with specialized language : a practical guide to using corpora*, Routledge, London, 2002.

Bugalhão J., Catarino H., Cavaco S., Covaneiro J., Fernandes I. C., Gomes A., Gómez Martínez S., Gonçalves M. J., Grangé M., Inácio I., Lopes G., Santos C., "CIGA: projecto de sistematização para a cerâmica islâmica do Gharb al-Ândalus", *Xelb*, vol. 10, p. 455-476, 2010.

Caballero Zoreda L., "El dibujo arqueológico: notas sobre el registro gráfico en arqueología", *Papeles del partal*, vol. 3, p. 75-95, 2006.

Cabré M. T., "Constituer un corpus de textes de spécialité", *Cahier du CIEL*, vol. 2007-2008, p. 37-56, 2008.

Cabré M. T., Palatresi J., "Acquisition of terminological data from text: approaches", *in* R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (eds), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume*, De Gruyter Mouton, Berlin, p. 1486-1497, 2013.

Carvajal López J. C., "The archaeology of al-Andalus: past, present and future", *Medieval Archaeology*, vol. 58, p. 318-339, 2014.

Cimiano P., McCrae J. P., Buitelaar P., Lexicon model for ontologies: community report, Technical report, Ontology-Lexicon Community Group, May, 2016.

Cimiano P., McCrae J. P., Rodríguez-Doncel V., Gornostay T., Gómez-Pérez A., Siemoneit B., Lagzdins A., "Linked terminologies: applying linked data principles to terminological resources", *in* I. Kosem, M. Jakubíček, J. Kallas, S. Krek (eds), *Electronic lexicography in the 21st century: linking lexical data in the digital age: Proceedings of the eLex 2015 conference*, Trojina, Ljubljana, p. 504-516, 2015.

Coll Conesa J., Martí Oltra J., Pascual Pacheco J., *Cerámica y cambio cultural: el tránsito de la Valencia islámica a la cristiana*, Dirección General de Bellas Artes y Archivos, Madrid, 1988.

Condamines A., "Terminological knowledge bases", *in* P. A. Fuertes-Olivera (ed.), *The Routledge Handbook of Lexicography*, Routledge, London, p. 335-350, 2018.

Costa R., Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas, Doctoral thesis, Universidade Nova de Lisboa, 2001.

Costa R., "Terminology and specialised lexicography: two complementary domains", *Lexicographica*, vol. 29, n^{o} 1, p. 29-42, 2013.

Covaneiro J., Fernandes I. C., Gómez Martínez S., Gonçalves M. J., Inácio I., Santos C., Coelho C., Liberato M., Bugalhão J., Catarino H., Cavaco S., "Cerâmica islâmica em Portugal : 150 anos de investigação", *Arqueologia em Portugal : 150 anos*, Associação dos Arqueólogos Portugueses, Lisboa, p. 73-80, 2013.

Cruse D. A., *Lexical semantics*, Cambridge University Press, Cambridge, 1986.

Doerr M., "Tailoring the Conceptual Model to archaeological requirements", *ARIADNE: the way forward to digital archaeology in Europe*, ARIADNE, [S.l.], p. 65-74, 2014.

Faber P., Léon-Araúz P., Reimerink A., "Representing environmental knowledge in EcoLexicon", *in* E. Bárcena, T. Read, J. Arús (eds), *Languages for specific purposes in the digital era*, Springer, Cham, p. 267-301, 2014.

Gangemi A., "Dolce+D&S Ultralite and its main ontology design patterns", *in* P. Hitzler, A. Gangemi, A. Janowicz, A. A. Krisnadhi, V. Presutti (eds), *Ontology engineering with ontology design patterns: foundations and applications*, IOS Press, Amsterdam, p. 81-103, 2016.

Gomes R. V., *Silves (Xelb), uma cidade do Gharb Al-Andalus: território e cultura*, Instituto Português de Arqueologia, Lisboa, 2002.

Gomes R. V., *Silves (Xelb), uma cidade do Gharb Al-Andalus: a alcáçova*, Instituto Português de Arqueologia, Lisboa, 2003.

Gomes R. V., *Silves (Xelb), uma cidade do Gharb Al-Andalus: o núcleo urbano*, Instituto Português de Arqueologia, Lisboa, 2006.

Gomes R. V., *Silves (Xelb), uma cidade do Gharb Al-Andalus: a zona da Arrochela, espaços e quotidianos*, Instituto Português de Arqueologia, Lisboa, 2011.

Gómez Martínez S., "Contenedores de fuego en el Garb al-Andalus", *in* V. O. Jorge (ed.), *Actas do 3º Congresso de Arqueologia Peninsular*, vol. 7, ADECAP, Porto, p. 421-434, 2000.

Gómez Martínez S., La cerámica islámica de Mértola: producción y comercio, Doctoral thesis, Universidad Complutense de Madrid, 2004.

Guarino N., Musen M., "Applied ontology: focusing on content", *Applied ontology*, vol. 1, nº 1, p. 1-5, January, 2005.

Hitzler P., Krötzsch M., Rudolph S., *Foundations of semantic web technologies*, CRC Press, Boca Raton, 2010.

Jockers M. L., *Macroanalysis: digital methods and literary history*, University of Illinois Press, Urbana, Il, 2013.

L'Homme M. C., "A lexico-semantic approach to the structuring of terminology", *Proceedings of CompuTerm 2004*, COLING, Geneva, p. 7-13, 2004.

Mascardi V., Cordì V., Rosso P., "A comparison of upper ontologies", *Proceedings of WOA 2007*, Seneca Edizioni, Torino, p. 55-64, 2007.

Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., Schneider L., WonderWeb deliverable D17: the WonderWeb library of foundational ontologies: preliminary report, Technical report, The WonderWeb Project, 2003.

Meeks E., Weingart S. B., "The digital humanities contribution to topic modeling", *Journal of Digital Humanities*, 2012.

Melby A. K., "Terminology in the age of multilingual corpora", *The Journal of Specialised Translation*, vol. 18, p. 7-29, July, 2012.

Meyer I., "Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework", *in* D. Bourigault, C. Jacquemin, M. L'Homme (eds), *Recent advances in computational terminology*, John Benjamins, Amsterdam, p. 279-302, 2001.

Meyer I., Skuce D., Bowker L., Eck K., "Towards a new generation of terminological resources: an experiment in building a terminological knowledge base", *COLING '92 Proceedings of the 14th conference on Computational linguistics*, vol. 3, Association for Computational Linguistics, Stroudsburg, PA, p. 956-960, 1992.

Munn K., Smith B. (eds), *Applied ontology: an introduction*, Ontos, Heusenstamm, 2008.

Nazarenko A., Hamon T., “Structuration de terminologie : quels outils pour quelles pratiques?”, *TAL*, vol. 43, n° 1, p. 7-18, 2002.

NF ISO 704, *Terminology Work – Vocabulary*, AFNOR, La Plaine Saint-Denis, 2009.

Prieto Velasco J., Faber P., “Graphical information”, *in* P. Faber (ed.), *A cognitive linguistics view of terminology and specialized language*, De Gruyter, Berlin, p. 225-248, 2012.

Roche C., “Ontoterminology: how to unify terminology and ontology into a single paradigm”, *LREC 2012*, ELRA, Paris, p. 2626-2630, 2012.

Roche C., “Ontological definition”, *in* H. J. Kockaert, F. Steurs (eds), *Handbook of terminology: vol. 1*, John Benjamins, Amsterdam, p. 128-152, 2015.

Rosselló-Bordoy G., *Ensayo de sistematización de la cerámica árabe en Mallorca*, Institut d’Estudis Baleàrics, Palma de Mallorca, 1978.

Rosselló-Bordoy G., *El nombre de las cosas en al-Andalus: una propuesta de terminología cerámica*, Museo de Mallorca, Palma de Mallorca, 1991.

Rychlý P., “A lexicographer-friendly association score”, *in* P. Sojka, A. Horák (eds), *Proceedings of Recent Advances in Slavonic Natural Language Processing*, Masaryk University, Brno, p. 6-9, 2008.

Santos C., Costa R., “Domain specificity: semasiological and onomasiological knowledge representation”, *in* H. J. Kockaert, F. Steurs (eds), *Handbook of terminology*, vol. 1, John Benjamins, Amsterdam, p. 153-179, 2015.

Silva R., Gestão de terminologia pela qualidade: processos de validação, Doctoral thesis, Universidade Nova de Lisboa, 2014.

Sure Y., Staab S., Studer R., “Ontology engineering methodology”, *in* S. Staab, R. Studer (eds), *Handbook on ontologies*, 2nd ed edn, Springer, Berlin, p. 135-152, 2009.

Torres C., Gómez Martínez S., Ferreira M. B., “Os nomes da cerâmica medieval: inventário de termos”, *Actas das 3as Jornadas de Cerâmica Medieval e Pós-Medieval*, Câmara Municipal de Tondela, Tondela, p. 125-134, 2003.

Vallejo Triano A., Escudero Aranda J., “Aportaciones para una tipología de la cerámica común califal de Madinat al-Zahra”, *Arqueología y territorio Medieval*, vol. 6, p. 133-176, 1999.

W3C OWL Working Group, OWL 2 Web Ontology Language document overview: 2nd ed, Technical report, W3C OWL Working Group, December, 2012.

Zozaya J., “Una discusión recuperada: candiles musulmanes de disco impreso”, *Arqueología y territorio medieval*, vol. 6, p. 261-278, 1999.

Zozaya J., “Candiles de piquera”, *Tierras del olivo*, Fundación El Legado Andalusí, Granada, p. 125-135, 2007.

Note de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Anne LACHERET-DUJOUR, Sylvain KAHANE, Paola PIETRANDREA. A prosodic and syntactic treebank for spoken French. *John Benjamins publishing company*. 2019. 396 pages. ISBN 978-9-027-20220-8.

Lu par **Fanny KRIMOU**

Université d'Orléans / LLL

L'ouvrage expose au lecteur, qu'il soit spécialiste ou novice, une approche et une méthodologie de l'étude du discours, de la prosodie et la syntaxe à partir du corpus Rhapsodie, une banque de données orales en français collectées, structurées, transcrites et annotées dans le cadre du projet Rhapsodie.

Cet ouvrage collectif synthétise les travaux issus du projet Rhapsodie coordonné par Anne Lacheret-Dujour, Sylvain Kahane et Paola Pietrandrea.

Articulé en dix-neuf chapitres, il expose la construction d'un modèle d'interface entre la prosodie et la syntaxe pour l'analyse du discours en français, en s'inspirant des théories fondatrices et des méthodes d'analyse contemporaines.

Les chapitres 1 et 2 rendent compte de la conception et de la création du corpus, de la description des métadonnées et du choix des conventions orthographiques et phonologiques.

Le corpus Rhapsodie, composé de plus de 33 000 mots, centralise au total trois heures d'échantillons de productions orales de français natifs d'une durée moyenne de cinq minutes. Ces extraits sonores sont triés par types de discours : monologues et dialogues, communications privées et publiques, discours spontanés, semi-spontanés et planifiés, degré d'interactivité (interactif, non interactif et semi-interactif), types (argumentatif, narratif, descriptif, oratoire...) ainsi que le type de projet dont ils sont extraits (interviews, films, commentaires sportifs...). Les données sociologiques telles que l'âge, la profession ou le genre du locuteur sont précisées.

La matrice de cette banque de données repose sur la diversité de données issues de sept corpus externes ainsi que sur un corpus interne regroupant des enregistrements collectés à partir de plusieurs supports multimédias.

Transcrit à la fois orthographiquement et phonétiquement, ce corpus ne comporte pas de ponctuation. Il a été conçu de façon à être utilisé sur le logiciel Praat pour

l'étude de la prosodie. De plus, il a été segmenté semi-automatiquement à la fois en mots, syllabes et phonèmes grâce à l'extension EasyAlign du logiciel Praat.

Les linguistes, et plus particulièrement ceux qui s'intéressent à la syntaxe, pourront se référer aux chapitres 3 à 7.

Le schéma Rhapsodie, décrypté dans cette partie, s'appuie sur la syntaxe de dépendance et la syntaxe de l'oral. Les mécanismes syntaxiques sont ainsi disséqués en traitant indépendamment la microsyntaxe (unités de rection) et la macrosyntaxe (unités illocutoires). Ce choix est argumenté dans le chapitre 3.

Le traitement d'annotation semi-automatique de la banque de données pour la microsyntaxe est expliqué au chapitre 4. La démarche d'annotation manuelle sur le plan macrosyntaxique est décrite dans le chapitre 6. Les différents outils utilisés sont présentés au chapitre 7.

Les auteurs proposent, chapitre 5, une analyse fine des phénomènes d'entassements (coordination, reformulation, etc.) tout en classifiant les phénomènes de marqueurs de discours repérés. Enfin, ils développent un niveau supplémentaire d'annotation en relevant les relations de dépendance entre les unités.

Les prosodistes, comme les non-initiés dans ce domaine, s'intéresseront aux chapitres 8 à 14.

Les caractéristiques de la structure prosodique annotée sont tout d'abord détaillées dans le chapitre 8 afin que le lecteur puisse saisir l'enjeu et la difficulté de l'annotation prosodique.

Avant de procéder à la transcription des phénomènes prosodiques, les données ont néanmoins été nettoyées (chapitre 14). Il est en effet nécessaire, pour mettre en évidence les paramètres acoustiques et particulièrement la fréquence fondamentale (F0), d'effacer les bruits parasites qui faussent les résultats et ce, grâce au logiciel WinPitch.

Le chapitre 9 expose ensuite la méthodologie pour le codage des proéminences syllabiques et des disfluences. Intrinsèquement liée aux phénomènes syntaxiques, cette annotation a été menée par des étudiants, annotateurs à l'« oreille semi-naïve », puis des experts, annotateurs à l'« oreille avertie ».

L'objectif principal du projet Rhapsodie étant la génération automatique d'une structure prosodique, le chapitre 11 rappelle l'hétérogénéité des éléments composants la structure prosodique et met en valeur les pauses prosodiques.

Le chapitre 10, quant à lui, présente le modèle Analor et la segmentation en périodes intonatives et en unités d'intégration prosodiques maximales fondée sur le paramètre de hauteur (la fréquence fondamentale).

Le chapitre 12 introduit l'extension Prosogram de Praat qui permet la stylisation de la fréquence fondamentale (F0) et l'étiquetage automatique conçu à partir du découpage en syllabes de l'échantillon.

Il est complété par le chapitre 13 dédié à la présentation de SLAM, un algorithme pour la stylisation et l'étiquetage automatique du contour mélodique à partir des travaux de Delattre et de ceux d'Aubergé.

La dernière partie propose aux spécialistes du traitement automatique du langage des solutions au traitement structurel des données (chapitre 15) en détaillant les outils et les méthodes d'analyse, d'annotation et d'exploitation automatique stabilisés utiles au traitement quantitatif de la syntaxe orale (chapitre 16) et de la prosodie (chapitre 17). Cette dernière partie est l'occasion également d'exposer et de discuter des résultats obtenus par l'approche quantitative des données.

Le chapitre 18 étudie le rôle que jouent les indices intonosyntaxiques pour la compréhension du discours en explorant notamment les relations entre les unités prosodiques majeures et les unités macrosyntaxiques du continuum sonore.

Le chapitre 19, conclusion de l'ouvrage, démontre que les techniques de traitement des données orales développées par Rhapsodie peuvent être utilisées sur des corpus de plus grande échelle et pour d'autres langues.

Issu d'un travail rigoureux et de la mise en place de techniques pour l'annotation manuelle et automatique de l'oral, le projet Rhapsodie livre, dans cet ouvrage, différents protocoles efficaces accompagnés de l'explication des choix qui ont été opérés pour la constitution d'un corpus échantillonné transcrit et annoté pour le traitement syntaxique et prosodique. Il sera, à ce titre, très utile à tous les linguistes travaillant sur le français oral et les questions de prosodie et de syntaxe. On rappellera que ce projet rend disponible le corpus Rhapsodie sur le site www.projet-rhapsodie.fr

Si l'origine grecque du mot Rhapsodie fait référence à la « couture des chants », l'ouvrage témoigne des progrès accomplis en ce qui concerne l'étude des propriétés prosodiques et syntaxiques du discours, en fournissant notamment les premiers résultats quantitatifs de l'analyse intonosyntaxique et une visualisation des relations prosodicosyntaxiques.

Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Elise BIGEARD : bigeard.elise@live.fr

Titre : Détection et analyse de la non-adhérence médicamenteuse dans les réseaux sociaux

Mots-clés : fouille de texte, apprentissage automatique, domaine médical, non-adhérence.

Title: *Detection and Analysis of Drug Non Compliance in Social Media*

Keywords: *text mining, machine learning, medical domain, non compliance.*

Thèse de doctorat en sciences du langage, Maison de la Recherche, Université de Lille, sous la direction de Natalia Grabar (CR HDR, CNRS) et Frantz Thiessard (MCU-PH, Université de Bordeaux). Thèse soutenue le 16/10/2019.

Jury : Mme Natalia Grabar (CR HDR, CNRS, codirectrice), Mme Farah Benamars (MC HDR, Université Toulouse III Paul Sabatier, rapporteur), M. Mathieu Roche (Chercheur HDR, Cirad, TETIS, rapporteur), M. Luigi Lancieri (Pr, Université de Lille, président), M. Fabien Torre (MC, Université de Lille, examinateur), Mme Anne-Lyse Minard (MC, Université d'Orléans, examinatrice), Mme Lorraine Goeuriot (MC, Université Grenoble Alpes, examinatrice), M. Frantz Thiessard (MCU-PH, Université de Bordeaux, codirecteur).

Résumé : *La non-adhérence médicamenteuse désigne les situations où le patient ne suit pas les directives des autorités médicales concernant la prise d'un médicament. Il peut s'agir d'une situation où le patient prend trop (sur-usage) ou pas assez (sous-usage) de médicaments, boit de l'alcool alors qu'il y a une contrindication, ou encore commet une tentative de suicide à l'aide de médicaments. Améliorer l'adhérence pourrait avoir un plus grand impact sur la santé de la population que tout autre amélioration d'un traitement médical spécifique. Cependant, les données sur la non-adhérence*

sont difficiles à acquérir puisque les patients en situation de non-adhérence sont peu susceptibles de rapporter leurs actions à leur médecin. Nous proposons d'exploiter les données des réseaux sociaux pour étudier la non-adhérence médicamenteuse.

Dans un premier temps, nous collectons un corpus de messages postés sur des forums médicaux. Nous construisons des vocabulaires de noms de médicaments et de maladies utilisés par les patients. Nous utilisons ces vocabulaires pour indexer les médicaments et maladies dans les messages. Ensuite nous utilisons des méthodes d'apprentissage supervisé et de recherche d'information pour détecter les messages de forum parlant d'une situation de non-adhérence. Avec les méthodes d'apprentissage supervisé nous obtenons 0,513 de F-mesure, avec un maximum de 0,5 de précision ou 0,6 de rappel. Avec les méthodes de recherche d'information, nous identifions des situations spécifiques comme la consommation d'alcool en contrindication ou l'usage psychotrope de neuroleptiques.

Nous étudions ensuite le contenu des messages ainsi découverts pour connaître les différents types de non-adhérence et savoir comment et pourquoi les patients se retrouvent dans de telles situations. Nous identifions trois motivations : gérer soi-même sa santé, rechercher un effet différent de celui pour lequel le médicament est prescrit, être en situation d'addiction ou d'accoutumance. La gestion de sa santé recouvre plusieurs situations : éviter un effet secondaire, moduler l'effet du médicament, sous-utiliser un médicament perçu comme inutile, agir sans avis médical. Additionnellement, une non-adhérence peut survenir par erreur ou négligence, sans motivation particulière.

À l'issue de notre étude nous produisons : un corpus annoté avec des messages de non-adhérence, un classifieur capable de détecter les messages de non-adhérence, une typologie des situations de non-adhérence et une analyse des causes de la non-adhérence.

URL où le mémoire peut être téléchargé :
http://elisebigeard.yo.fr/static_files/these_bigeard.pdf