N° d'ordre NNT : 2025ISAL0044

THESE de DOCTORAT DE L'INSA LYON,
membre de l'Université de Lyon

Ecole Doctorale N° 512
InfoMaths

Spécialité/ discipline de doctorat : Informatique

Soutenue publiquement le 10/06/2025, par :

Alice Brenon

Méthodes et outils pour l'étude diachronique des discours géographiques dans deux encyclopédies françaises

Devant le jury composé de :

Favre	Benoît	Professeur des Universités	Aix-Marseille Université	Rapporteur
Galleron	Ioana	Professeure des Universités	Université Sorbonne Nouvelle	Rapporteure
Doucet	Antoine	Professeur des Universités	La Rochelle Université	Examinateur
Eglin	Véronique	Professeure des Universités	INSA-LYON	Examinatrice
Puren	Marie	Maîtresse de Conférences	EPITA	Examinatrice
Laforest	Frédérique	Professeure des Universités	INSA-LYON	Directrice de thèse
Vigier	Denis	Professeur des Universités	Université Lumière Lyon 2	Co-directeur de thèse
Moncla	Ludovic	Maître de Conférences	INSA-LYON	Invité

Département FEDORA –INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
ED 206 CHIMIE	CHIMIE DE LYON https ://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	M. Stéphane DANIELE C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne directeur@edchimie-lyon.fr
ED 341 E2M2	ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION http ://e2m2.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	Mme Sandrine CHARLES Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX e2m2.codir@listes.univ-lyon1.fr
ED 205 EDISS	INTERDISCIPLINAIRE SCIENCES-SANTÉ http ://ediss.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	Mme Sylvie RICARD-BLUM Laboratoire ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
ED 34 EDML	MATÉRIAUX DE LYON http ://ed34.universite-lyon.fr Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	M. Stéphane BENAYOUN Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 stephane.benayoun@ec-lyon.fr
ED 160 EEA	ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE https ://edeea.universite-lyon.fr Sec. : Philomène TRECOURT Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	M. Philippe DELACHARTRE INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 philippe.delachartre@insa-lyon.fr
ED 512 INFOMATHS	INFORMATIQUE ET MATHÉMATIQUES http ://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	M. Hamamache KHEDDOUCI Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 direction.infomaths@listes.univ-lyon1.fr
ED 162 MEGA	MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE http ://edmega.universite-lyon.fr Sec. : Philomène TRECOURT Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	M. Etienne PARIZET INSA Lyon Laboratoire LVA Bâtiment St. Exupéry 25 bis av. Jean Capelle 69621 Villeurbanne CEDEX etienne.parizet@insa-lyon.fr
ED 483 ScSo	ScSo[1] https ://edsciencessociales.universite-lyon.fr Sec. : Mélina FAVETON Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	M. Bruno MILLY (INSA : J.Y. TOUSSAINT) Univ. Lyon 2 Campus Berges du Rhône 18, quai Claude Bernard 69365 LYON CEDEX 07 Bureau BEL 319 bruno.milly@univ-lyon2.fr

1. ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Département de la Formation par la Recherche et des Études Doctorales (**FEDORA**)

Bâtiment Charlotte Perriand 3e étage
17, av. des Arts
69621 Villeurbanne Cédex
fedora@insa-lyon.fr

Référence : TH1243 BRENON Alice

L'INSA Lyon a mis en place une procédure de contrôle systématique via un outil de détection de similitudes (logiciel Compilatio). Après le dépôt du manuscrit de thèse, celui-ci est analysé par l'outil. Pour tout taux de similarité supérieur à 10%, le manuscrit est vérifié par l'équipe de FEDORA. Il s'agit notamment d'exclure les auto-citations, à condition qu'elles soient correctement référencées avec citation expresse dans le manuscrit.

Par ce document, il est attesté que ce manuscrit, dans la forme communiquée par la personne doctorante à l'INSA Lyon, satisfait aux exigences de l'Etablissement concernant le taux maximal de similitude admissible.

Remerciements

Mes premiers remerciements vont à Denis Vigier pour avoir su voir en moi la chercheuse que je pouvais devenir quand j'ai poussé la porte de son bureau pour la première fois, il y a un peu plus de sept ans. Je tiens aussi à remercier chaleureusement Frédérique Laforest pour la qualité de son encadrement, sa disponibilité et tout ce qu'elle m'a transmis sur le plan pédagogique. J'aimerais également dire merci à Ludovic Moncla pour son exigence constante, jusque dans les moments les plus difficiles de cette thèse. Merci infiniment aux membres de mon jury pour leur patience, ayant attendu mon manuscrit bien au-delà de la date de fin initialement prévue quand nous les avons contactés il y a plus d'un an : je ne sais comment remercier Benoît Favre et Ioana Galleron pour m'avoir lue et critiquée en détail, Antoine Doucet pour son intérêt renouvelé pour mon travail après trois ans de suivi annuel ainsi que Marie Puren et Véronique Églin pour leur bienveillance et leurs encouragements.

Je suis en outre extrêmement redevable aux autres membres du projet GEODE et en particulier à Thierry Joliveau pour les échanges passionnants sur la Géographie au XIX^{ème} siècle, à Katherine McDonough pour son accueil bienveillant à Lancaster et à Olivier Kraif, toujours disponible pour me faciliter la vie avec le Lexicoscope.

Au vu de la place qu'a pris la préparation de La Grande Encyclopédie dans ce travail, j'adresse un grand merci à Christine Jacquet-Pfau pour sa connaissance précieuse de l'œuvre qu'elle communique avec enthousiasme. Les présents travaux m'ont également confrontée à des questions de reproductibilité que j'ai pu aborder sereinement grâce au travail fantastique de la communauté Guix, extrêmement accueillante et que je tiens aussi à saluer, notamment Ludovic Courtès, Céline Acary-Robert et Simon Tournier. J'étends ces remerciements aux esprits stimulants croisés dans les événements autour de la science ouverte, et aux personnes qu'ils m'ont à leur tour permis de rencontrer : deux grands mercis, à Jonathan McHugh et J.-P. Ascher, pour les discussions fertiles sur l'organisation des connaissances.

Plus généralement, je dois dire toute la gratitude que je ressens envers les personnels, chercheuses et chercheurs avec qui j'ai eu le privilège de travailler ces dernières années. Je remercie en particulier Jean-Philippe Magué pour ses conseils toujours judicieux, Matthieu Decorde, Alexei Lavrentiev ainsi que toute la communauté autour du logiciel TXM pour leur réactivité et leur aide inestimable. Je suis très reconnaissante à mes collègues ayant terminé leur thèse il y a peu pour m'avoir partagé leur expérience et permis de me projeter peu à peu dans l'après-thèse : merci à Luisa, Lucien, Santiago, Louis, Gabriel et François. Merci aussi à toutes mes collègues en master ou en thèse dont j'ai eu la chance de partager un quotidien fait de travail (beaucoup), de soucis (parfois) et de bons moments (souvent): ma demi-sœur de thèse Ghayoung, Moa et Corentin, Liv, Zeina, Lison, Luce, Alexandre, Morgane, Sinem, Antoine et Zeinab. Une mention spéciale aux meilleures co-bureaux du monde : Louise, Katherine et Taegan. Vous êtes toujours les bienvenues dans mon futur bureau, où qu'il se trouve. J'ai conscience de l'injustice que je commets en ne pouvant pas prendre davantage de place pour nommer entièrement chaque membre des laboratoires ICAR, LIRIS et LIP (que j'ai fréquenté sur la dernière année) ni dire tout ce que je leur dois mais : Matthieu, Carole, Magali, Daniel, Lu-

cie, Justine, Tarek, Audrey, Stéphanie, Lionel, (l'autre) François, Michele, Rémi, Diane, Pascal, Lois, Nathalie... merci. Je pense à vous. Enfin, je n'aurais jamais même pu envisager de faire une thèse sans ma prof de maths de sup'. Merci à Hélène Benhamou pour m'avoir inspirée, et accessoirement appris la rigueur, la logique et donc à programmer.

Je dois d'ailleurs énormément à mes camarades taupes de l'époque avec qui nous travaillions de concert en salle 356 et qui m'ont précédée dans la recherche : Thomas, Cécile et Julien. Je veux également remercier l'entourage affectif qui m'a soutenue tout au long de ces quatre années et quelques. Je pense avec émotion à Aure, Liolia, Ema, Alexandra, Amandine et toutes les amies avec qui j'ai partagé des entraînements sur glace dont la régularité m'a été salutaire. Je rends grâce à Sébastien et Maria qui ont joué les plus belles mesures de la musique qui m'a accompagnée pendant la rédaction ; à mes colocs de serveurs Martin et Sergiu qui ont tenu les murs quand je n'avais plus d'énergie pour regarder les logs ; à Grégory qui – même outre-Atlantique – sait parler de n'importe quel travail scientifique comme si c'était la découverte la plus importante de tous les temps ; à Paul, pour moi tout à la fois collègue, ami et famille ; et à Céline, Florent, Pierre ainsi qu'à la famille Maynié-Grauwin au grand complet toujours là pour parler, rire et me redonner de l'énergie. Pour finir, je remercie mes parents pour leurs encouragements continus à me dépasser sans cesse, de la maternelle à la fin de cette thèse. Enfin, ces derniers mots sont pour toi, Isa, mon amour. Merci d'avoir d'avoir traversé ça avec moi.

Table des matières

Remerciements **3**

1 Introduction **7**
1.1 Aux sources de cette thèse . . . 7
1.2 Réflexions prospectives . . . 10
1.3 Plan de l'étude . . . 14

2 Un instantané des Humanités Numériques **17**
2.1 Structuration et encodage des données . . . 18
2.1.1 Les formats XML . . . 18
2.1.2 Le format XML-TEI . . . 20
2.1.3 Succès et limites du format XML-TEI . . . 21
2.2 (Méta)lexicographie . . . 23
2.2.1 Portée conceptuelle des ouvrages lexicographiques . . . 23
2.2.2 Dictionnaires et encyclopédies . . . 25
2.2.3 L'Encyclopédie au centre des attentions . . . 26
2.3 Histoire et terminologie . . . 28
2.3.1 Rivalité historique entre les approches symbolique et statistique . . . 28
2.3.2 L'avènement des méthodes statistiques . . . 29
2.3.3 Nommer des pratiques très différentes . . . 29
2.4 Traitement Automatique de la Langue . . . 30
2.4.1 Vectorisation . . . 30
2.4.2 Classification . . . 31
2.4.3 Étiquetage morphosyntaxique et syntaxique . . . 34
2.5 Linguistiques de corpus . . . 36
2.5.1 Du texte au corpus . . . 36
2.5.2 Les outils de la linguistique . . . 38
2.5.3 Études contrastives . . . 40

3 Préparation et enrichissement du corpus **43**
3.1 Les œuvres du corpus . . . 44
3.1.1 Contextes historiques . . . 44
3.1.2 Organiser les connaissances . . . 47
3.1.3 Prétraitements . . . 58
3.2 Encodage des fichiers . . . 62
3.2.1 Des différences entre Encyclopédies et Dictionnaires . . . 62
3.2.2 Proposition d'un schéma d'encodage . . . 71
3.2.3 Le mythe du «texte brut» . . . 77
3.3 Des textes au corpus . . . 80
3.3.1 Différents formats pour différents usages . . . 80

3.3.2 Principes de structuration . . . 82
3.3.3 Reproductibilité . . . 86

4 Classification automatique en domaines de connaissance 91
4.1 Comparaison de méthodes de classification automatique . . . 92
4.1.1 Démarche . . . 92
4.1.2 Choix des modèles et préparation des jeux de données . . . 93
4.1.3 Évaluation des performances . . . 94
4.1.4 Classification non supervisée . . . 100
4.2 Relations entre *Géographie* et autres domaines de connaissance . . . 102
4.2.1 Observations préliminaires des erreurs . . . 103
4.2.2 Similarités lexicales . . . 112
4.2.3 Application de la centralité à la structure du modèle . . . 119
4.3 Annotation des articles . . . 123
4.3.1 Choix d'un classifieur . . . 123
4.3.2 Extension à *LGE* . . . 126

5 Études contrastives 133
5.1 Tracer les contours de la géographie . . . 133
5.1.1 La place de la géographie . . . 134
5.1.2 Des articles qui changent de domaine . . . 146
5.2 La biographie cachée . . . 157
5.2.1 Spécificité des lemmes «naître» et «mourir» en *Géographie* . . . 158
5.2.2 L'influence des domaines . . . 164
5.2.3 Différents profils de passages biographiques . . . 166

6 Conclusion 171
6.1 Synthèse des résultats . . . 171
6.2 Perspectives . . . 174

Glossaire 185

Bibliographie 189

Chapitre 1

Introduction

1.1 Aux sources de cette thèse 7
1.2 Réflexions prospectives 10
1.3 Plan de l'étude 14

1.1 Aux sources de cette thèse

Le genre encyclopédique

Si l'on en croit André Marcel Berthelot, la Géographie serait la science encyclopédique par excellence :

> Le géographe aborde successivement le domaine de plusieurs sciences définies ; il en prend les résultats et les place dans sa description synthétique : astronomie, physique, chimie, géologie, botanique, zoologie, anthropologie, linguistique, sociologie, statistique, démographie, histoire, toutes les branches des connaissances humaines lui apportent leur contingent de faits [...].
>
> – La Grande Encyclopédie, T18, p.767

Cet extrait de l'article GÉOGRAPHIE qu'il rédige pour *La Grande Encyclopédie, Inventaire raisonné des Sciences, des Lettres et des Arts par une Société de savants et de gens de lettres – LGE* dans le reste de cette thèse –rappelle en effet l'«enchaînement de toutes les connaissances» à l'origine du mot «encyclopédie». Plus d'un siècle après la fin de la publication de l'*Encyclopédie, Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres – EDdA* dans le reste de cette thèse – l'«esprit encyclopédique» qui a présidé à l'apparition du genre encyclopédique au tournant des XVII^{ème} et XVIII^{ème} siècles (Macary, 1973, p. 145) semble donc continuer à inspirer les auteurs.

L'*EDdA* a en effet eu de nombreux précurseurs, au rang desquels on peut compter le *Dictionnaire Universel* de Furetière (Galleron & Williams, 2022), le *Dictionnaire François-Latin* publié à Trévoux (Le Guern, 1983) et la *Cyclopedia* de Chambers dont l'*EDdA* ne devait initialement être qu'une traduction (Kafker & Loveland, 2016). Mais la portée de cet esprit encyclopédique se mesure davantage à la descendance fournie qu'il a laissée : le *Supplément* à l'*EDdA* (Doig, 1990), l'*Encyclopédie Méthodique* de Panckoucke (Groult, 2019a) puis dans le courant du XIX^{ème} siècle le *Grand Dictionnaire Universel* de Larousse et *LGE* (Jacquet-Pfau, 2022), et cela rien qu'en France. Ailleurs dans le monde,

l'*Encyclopedia Britannica*, la *Brockhaus Enzyklopädie* et l'*Encyclopedia Americana* entre autres contribuent à diffuser cet esprit. Les encyclopédies restent des objets pertinents au XXI^{ème} siècle, ayant atteint une portée universelle comme le montre le succès de projets comme *Wikipédia*.

Le genre encyclopédique fait l'objet de nombreuses études depuis la thèse de Quemada (1968), notamment en diachronie. Le XVIII^{ème} siècle est marqué par une augmentation de la taille des équipes éditoriales ainsi que par des rectifications et des augmentations des contenus comme celles observables entre le *Furetière* et le *Trévoux* (Macary, 1973, pp. 150–155). C. Rey (2022) montre que ce mouvement se prolonge au début du XIX^{ème} siècle par une professionnalisation de la rédaction d'encyclopédies et il est naturel de se demander ce que devient ce mouvement au tournant des XIX^{ème} et XX^{ème} siècles.

La Géographie, une science en recomposition

Le lien entre la Géographie et les encyclopédies n'a pourtant pas toujours autant relevé de l'évidence que ne pourrait le laisser croire la citation de Berthelot. Les contenus géographiques sont très peu présents dans le *Furetière* (Galleron & Williams, 2022, p. 38 ; Vigier et al., 2022, p. 60) et ne font vraiment leur entrée dans les encyclopédies qu'avec la deuxième édition du *Trévoux* (1721). Dans l'Avertissement des Éditeurs (L'Encyclopédie, T3, p.xj), d'Alembert doit même défendre le choix d'avoir inclus des articles de Géographie dans l'*EDdA* alors que «plusieurs personnes ont pensé que les articles de Géographie étoient de trop dans ce Livre».

Le XVIII^{ème} siècle voit émerger l'association de la Géographie aux cartes (Verdier, 2015) alors que l'activité des géographes semblait pouvoir s'en dispenser auparavant. La carte apparaît alors comme une simple technique de visualisation, intéressant davantage les mathématiciens pour les constructions géométriques qu'elles requièrent. L'article CARTE (L'Encyclopédie, T2, p.706) le montre assez clairement et, bien qu'il ne comporte pas de signature, il est d'ailleurs communément attribué à d'Alembert, un mathématicien. La valeur du travail du géographe réside dans l'exactitude de ses relevés, à partir desquels pourront être construites les cartes. Il n'y a d'ailleurs pas de cartes dans l'*EDdA* et Diderot se satisfait d'une géographie «sèche», «scientifique», «la seule qui nous suffiroit pour construire de bonnes cartes des tems anciens, si nous l'avions, & qui suffira à la postérité pour construire de bonnes cartes de nos tems» ainsi qu'il l'écrit à l'article ENCYCLOPÉDIE (L'Encyclopédie, T5, p.646). Cette thématique de la carte reste pourtant présente dans l'œuvre et permet d'introduire une métaphore frappante que d'Alembert file dans le Discours Préliminaire des Éditeurs (L'Encyclopédie, T1, p.xv) : l'Encyclopédie décrit les différentes sciences comme autant de «pays» qu'elle situe les uns par rapport aux autres, permettant d'«entrevoir même quelquefois les routes secrètes qui les rapprochent». Ses articles sont des «cartes» qui couvrent en détail des portions de la «mappemonde» que constitue le «Système Figuré des connoissances humaines», un «arbre encyclopédique» matérialisé sous forme d'une gravure au tome 1er dans le but de montrer une vue d'ensemble des liens entre sciences.

La relation entre Géographie et encyclopédies à l'origine de cette réflexion est donc en réalité à double-sens : à la Géographie comme science encyclopédique, nourrie de toutes les autres, s'ajoute la pratique encyclopédique comme une «géographie des sciences» dont elle révèle le paysage. Cependant, au-delà de cette relation, la Géographie est soumise aux XVIII^{ème} et XIX^{ème} siècles à des forces transformatrices profondes. Chevalier (1997, p. 4) distingue quatre pôles du savoir géographique, opposant les géographies scolaires et universitaires mais aussi la géographie grand public et la géographie appliquée. L'École des géographes forme depuis 1797 des topographes (Verdier, 2015, p. 41) pour dresser des cartes utiles aux états-majors dans les nombreux conflits qui parsèment le XIX^{ème} siècle. Avec la période révolutionnaire, l'école reçoit la charge de former des citoyens dont les connaissances doivent progressivement intégrer de la géographie (Chevalier, 2013, para. 2 ; Todorov, 2018, para. 15). La géographie est aussi mise à contribution, associée à l'histoire, pour développer le

sentiment national qui émerge au XIX^{ème} siècle non seulement en France dès 1833 et jusqu'aux célèbres lois «Jules Ferry» qui mettent en avant la puissance coloniale française dans les programmes de géographie (Chevalier, 2013, para. 3) mais aussi ailleurs en Europe, comme en Italie, en Allemagne ou en Hollande (Todorov, 2018, para. 1 et 9). Vers la fin du XIX^{ème} siècle la Géographie se hisse enfin au rang de véritable discipline scientifique grâce à des auteurs comme Vidal de la Blache.

Opportunités pour l'informatique

La conduite d'investigations sur des objets à la fois massifs et historiques comme peuvent l'être les œuvres encyclopédiques évoquées soulève cependant un certain nombre de difficultés, notamment techniques. La volumétrie particulièrement élevée de ce type d'ouvrage représente un premier frein évident : sans moyens automatiques, il est en pratique impossible de rechercher des phénomènes linguistiques, de les cataloguer exhaustivement ou même de naviguer rapidement de l'un à l'autre. Les encyclopédies exacerbent bien sûr ce problème mais il apparaît néanmoins dans de nombreux projets dès qu'il s'agit de travailler sur un corpus d'œuvres ou de productions langagières assez large pour prétendre à une certaine représentativité. Pour les travaux portant sur les siècles passés, le caractère historique des documents étudiés ajoute encore des contraintes : la fragilité des exemplaires s'oppose à toute consultation intensive tout en exigeant des précautions techniques lourdes pour ne pas les abîmer ; leur rareté empêche également toute étude simultanée par les membres du projet – à fortiori dans le cadre de collaborations internationales. Ces quelques raisons suffisent amplement à justifier l'intérêt de numériser les œuvres et de les exploiter via des moyens électroniques. Leur dématérialisation permet en outre de les partager sans limite au sein de la communauté scientifique mais également de les faire connaître au grand public. Les intérêts des sciences humaines et celles de l'information convergent donc naturellement, un mouvement à l'origine des «Humanités Numériques» (HN). L'ordinateur y apporte des possibilités techniques originales au service des questionnements des autres disciplines mais ne se limite pas à un rôle de soutien technique : ces thématiques nouvelles créent aussi en retour des problématiques de recherche intéressantes pour l'informatique elle-même.

La mise à disposition des équipes de grandes quantités de données et parfois de logiciels d'analyses en ligne requiert des infrastructures complexes qui soulèvent déjà des questions d'ingénierie non triviales liées entre autres au stockage, à la vitesse d'accès aux données ou même à la prise en charge du travail collaboratif. Les efforts entrepris sur ces sujets ont donné naissance à des organisations comme Ortolang (Pierrel & Parisse, 2016) ou Huma-Num (Larrousse et al., 2023) en France, et comme CLARIN (Jong et al., 2020) à l'échelle européenne. L'augmentation du nombre de plateformes de ce type et la nécessité de pouvoir les interconnecter dans des réseaux comme le European Open Science Cloud (Di Donato et al., 2020) amène à devoir développer des formats ouverts et normaliser les données et métadonnées. L'apport des sciences du numérique ne se résume donc pas à l'implémentation de services mais inclut aussi la spécification de langages formels utilisables pour encoder les données, processus aboutissant à la production de standards comme la XML-TEI (Ide & Sperberg-Mcqueen, 1995), omniprésente en HN.

De plus, le recours à des outils numériques pour le traitement et l'analyse des données fait peser une responsabilité accrue sur ces derniers : leur qualité et leur précision deviennent des enjeux pour la solidité des connaissances produites. C'est ainsi que s'engagent de véritables courses en vue de repousser l'état de l'art, par exemple de la reconnaissance optique de caractères (Patel et al., 2012 ; Wick et al., 2018) ou du repérage et de la classification d'entités nommées (Humbel et al., 2021 ; Nadeau & Sekine, 2007). Parfois, des progrès significatifs ne peuvent être obtenus qu'au prix de l'adoption d'approches radicalement différentes comme ont pu l'être les méthodes d'Apprentissage Automatique (AA) par rapport aux méthodes symboliques. Ces ruptures technologiques peuvent s'accompagner d'un bouleversement des usages (concepts observables, métriques d'évaluations) qui amène en retour à repenser le rapport aux objets examinés pour comprendre le sens du résultat

d'un calcul, comme par exemple une prédiction faite par un classifieur automatique. Cet enjeu est en lien avec la visualisation des données, qui peut demander des efforts de développement logiciel mais interroge surtout sur l'interprétation pouvant être faite des concepts statistiques ou informatiques mobilisés. Enfin, la question des chaînes de traitement peut paraître secondaire, ne constituant qu'un assemblage «évident» ou même «nécessaire» des outils requis par un projet donné mais suscite en réalité de nombreuses réflexions sur les formats des données ainsi que sur l'équilibre à trouver entre commodité d'utilisation, flexibilité et réutilisabilité des états intermédiaires. Si de très nombreux projets en HN semblent partager une structure commune allant de la numérisation de données à l'application d'outils linguistiques automatisés (Jentsch & Porada, 2020), la diversité des besoins spécifiques d'analyse ou des formats requis en entrée ou en sortie de chaîne pour chaque projet complique en réalité le réemploi de chaînes existantes. Quelle que soit la direction envisagée, les thématiques d'étude choisies apparaissent donc potentiellement fructueuses pour la recherche informatique.

Le projet GEODE

Au-delà des résultats d'analyse auquels ils parviennent, l'enjeu des présents travaux repose donc davantage sur l'exploration des solutions techniques à disposition ainsi que des manières dont elles peuvent être combinées pour répondre à des besoins d'analyse correspondant aux pratiques des HN. Cette thèse s'inscrit en effet dans le cadre du projet GEODE [1] qui adopte cette approche interdisciplinaire. Pour ce faire, il étudie un corpus de quatre encyclopédies de la deuxième moitié du XVIII^{ème} au XXI^{ème} siècle de langue française principalement (Bender, 2019), émaillées ici et là de mots et parfois de citations complètes en latin et en grec pour les deux plus anciennes. Autour de ces textes, le projet réunit des spécialistes de linguistique, d'informatique, de géographie et d'histoire.

La première des œuvres du corpus de GEODE dans l'ordre chronologique est l'*EDdA*, publiée de 1751 à 1772 et témoin de la deuxième moitié du XVIII^{ème} siècle. Le deuxième jalon de l'étude est posé par *LGE* qui permet d'observer la toute fin du XIX^{ème} siècle, de 1885 à 1902. L'étape suivante dans la progression est marquée par l'*Encyclopedia Universalis*, lancée en 1966. L'édition versée au corpus est la 23ème, datant de 2018 et tenue à jour de l'évolution des connaissances scientifiques et des événements historiques les plus récents. Toutefois certains articles tel que CONSCIENCE ou LIBERTÉ signés respectivement par Ey (décédé en 1977), et Ricœur (en 2005) sont nécessairement plus anciens et font de cette œuvre une bonne représentante de contenus encyclopédiques de la fin du XX^{ème} siècle et du début du XXI^{ème}. La dernière œuvre utilisée est la version francophone de *Wikipédia*, qui par sa nature collaborative et vivante est mise à jour en continu et contient donc des états de langues tout à fait contemporains. Les travaux conduits dans le cadre de cette thèse se sont concentrés sur le premier intervalle allant de l'*EDdA* à *LGE* et au cours duquel se sont produites les transformations de la Géographie décrites par les spécialistes de la discipline dans les références évoquées p.8.

1.2 Réflexions prospectives

Problématique

Le lien privilégié qui unit les encyclopédies à la Géographie amène à s'interroger sur les effets qu'ont pu avoir dans ce type d'ouvrages les changements subis par la discipline après le siècle des Lumières. Puisqu'il y a lieu de faire l'hypothèse que ces mutations ont eu des conséquences sur la manière dont on écrit la Géographie, il faut se donner les moyens d'observer leurs effets, ce qui constitue l'enjeu principal de cette thèse.

1. https://geode-project.github.io/

La question que soulève immédiatement une telle problématique concerne donc les moyens d'accéder aux discours géographiques présents dans les encyclopédies du corpus. De nombreux projets en HN ont déjà développé des outils pour traiter des textes antérieurs au XX^{ème} siècle ; d'autres s'intéressent à l'importance d'un ensemble d'articles pour une discipline en particulier mais sans intégrer d'analyse de discours et sans les opposer aux autres sciences. Ici, l'objectif est à la fois de pouvoir traiter les textes des encyclopédies dans leur ensemble à l'aide de méthodes automatiques et, simultanément, d'être à tout moment capable de restreindre l'étude à des sous-corpus pertinents ou d'adopter une démarche contrastive, principalement selon deux axes. Le premier, celui du champ disciplinaire, repose sur la possibilité d'identifier des discours géographiques par opposition à d'autres qui ne relèveraient pas du même domaine. Le deuxième axe est temporel puisqu'il faut comparer les discours du genre encyclopédique à deux époques : le XVIII^{ème} d'une part et le tournant des XIX^{ème} et XX^{ème} siècles d'autre part.

Cette thématique de recherche mobilise plusieurs des apports de l'informatique aux HN. Il est d'abord nécessaire de déterminer un encodage qui convienne à la fois à l'*EDdA* et à *LGE* malgré leurs différences, de façon à pouvoir regrouper les deux œuvres au sein d'un même corpus et leur appliquer les mêmes traitements puis les mêmes analyses. Cette tâche de normalisation, avec celle d'organisation des textes et des métadonnées représente un important travail d'ingénierie des données. À ces efforts s'ajoutent ensuite ceux à fournir pour identifier les discours pouvant relever de la Géographie. Les articles d'encyclopédies se répartissent généralement selon le domaine de connaissance dont relèvent les notions qu'ils abordent, ce qui offre un premier accès à ces discours. À cette échelle, associer un domaine de connaissance aux articles prend la forme d'un problème de classification. Plus près des phrases, l'annotation automatisée des textes et l'écriture de requêtes basées sur des critères lexicaux ou syntaxiques permettent de mettre en évidence des motifs utiles pour révéler une structure plus fine. En faisant appel à la fois à l'informatique et aux sciences du langage, elles rendent compte de la diversité des concepts traités dans chaque discipline.

La Géographie et ses traces

Cependant, même au niveau des articles pris dans leur ensemble une évidence apparente doit être critiquée sous peine de fragiliser la démarche : celle de l'identification implicite entre d'une part la discipline ”géographie” et d'autre part l'ensemble des articles assignés à cette discipline par le choix éditorial des encyclopédistes. Alors que la première peut s'entendre comme un ensemble vivant de pratiques et de savoirs, nécessairement fini, elle renvoie à une idée bien plus vaste. Par contraste, la seconde demeure contingente et dans une certaine mesure arbitraire. Ces articles ne sont que l'écho de la discipline, nécessairement en retard sur ses progrès : ils sont écrits après réverberation dans un ou plusieurs des dictionnaires universels dans lesquels les encyclopédistes ont puisé leurs sources, parfois d'après des récits de voyages passés. Une expression du XVIII^{ème} siècle illustre parfaitement cette source d'information : «terme de relation» (Quemada, 1968, p. 309), sous laquelle sont classées certaines entrées dans l'*EDdA* et qui sous-entend le mot «voyage». Ces entrées peuvent entre autres servir à définir du vocabulaire utilisé lorsqu'un texte relate un voyage, par exemple une francisation d'un mot de la langue parlée dans le pays visité comme PILAU (L'Encyclopédie, T12, p.618) emprunté au Turc pour parler de la cuisson du riz.

L'objet qui intéresse fondamentalement cette thèse est bien sûr la Géographie en tant que discipline (qu'on distinguera typographiquement en conservant la majuscule) mais il ne se réduit pas aux discours présents dans une encyclopédie. Les présents travaux s'efforcent de s'en rapprocher en étudiant ces discours, c'est-à-dire la trace textuelle laissée par la discipline (pour laquelle on réservera l'emploi sans marquer l'initiale : «géographie»). De la même manière, le terme de «discipline» sera pris dans le sens large de «science» alors que l'expression «domaine de connaissance» renverra plus spécifiquement à un choix éditorial de découpage dans une encyclopédie particulière.

La situation est tout à fait analogue à celle en traitement du signal d'un processus d'échantillonnage par lequel on tente de connaître une fonction continue du temps à partir d'un ensemble fini de mesures, comme représenté à la figure 1.1. Tout l'enjeu est alors de savoir si les points échantillonnés sont assez représentatifs et couvrent suffisamment l'intervalle complet pour rendre fidèlement compte du phénomène observé. Mais là où en systématisant l'échantillonnage suivant une grille régulière on parvient à connaître et limiter la quantité d'information perdue grâce aux résultats des travaux de Nyquist (1928) repris et étendus par Shannon (1949)[2], la sélection opérée par les choix éditoriaux des encyclopédistes ne constitue pas un échantillonnage car, n'ayant pas été faite dans le but de fournir de l'information sur la discipline elle-même, elle est nécessairement irrégulière et probablement lacunaire. C'est la raison pour laquelle les mesures sont réparties aléatoirement sur la figure 1.1 plutôt qu'espacées entre elles d'un pas constant comme cela aurait été le cas sur un graphe similaire destiné à illustrer un vrai processus d'échantillonnage, par exemple d'un signal acoustique. Le problème est donc de travailler à partir d'une «trace» de la Géographie, passée et finie, sans pouvoir limiter ni même seulement connaître la quantité d'information perdue par rapport au «signal» que constituerait la discipline géographique. C'est pourtant le parti qu'il faut prendre, en l'absence d'un meilleur, tout en gardant cette distinction à l'esprit.

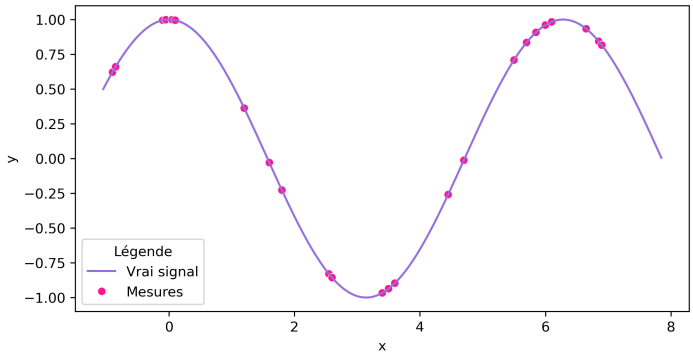


Figure 1.1 – Échantillonnage d'un processus sinusoïdal par un ensemble de points répartis aléatoirement

Faire correspondre des époques

À la lumière de la remarque ci-dessus, apparaît un deuxième questionnement sur l'identité des objets à comparer. En effet, puisqu'il est impossible d'accéder directement à la Géographie de chaque époque et qu'il faut se contenter de ses traces sous forme d'ensembles d'articles, il est naturel de s'interroger sur le bien-fondé d'une mise en regard de ces ensembles au seul motif qu'ils sont estampillés *Géographie* chacun à leur époque. Intuitivement, l'identité paraît évidente car c'est bien la même discipline qui a évolué continûment d'une époque à l'autre, et les encyclopédistes de chaque époque ont choisi les articles qu'ils considéraient relever de cette même discipline.

Mais en toute rigueur, avec les seuls éléments disponibles – c'est-à-dire des collections d'articles non seulement finies mais surtout disjointes, sans continuité temporelle – il n'est pas possible de l'affirmer : le nom aurait pu subir un glissement sémantique et désigner des disciplines entièrement distinctes aux deux époques. De plus, l'ensemble des catégories selon lesquelles les articles se répartissent diffère à chaque époque. Cela peut être dû à l'apparition d'une discipline entière comme par exemple «Industrie» dans *LGE* qui aurait été anachronique dans la deuxième moitié du XVIII$^{\text{ème}}$ siècle avant la révolution industrielle. Des redécoupages peuvent également avoir lieu : ainsi, là où la *Géographie* se retrouvait dans le «Système Figuré» des encyclopédistes des Lumières aux côté

2. le théorème de Nyquist-Shannon dit qu'en échantillonnant à la fréquence F on peut récupérer sans déformation toutes les composantes de fréquence $< \frac{F}{2}$ dans le signal initial

de l'*Uranographie* et de l'*Hydrographie* dans la *Cosmographie*, au sein des sciences mathématiques [3], elle se retrouve associée à l'*Histoire* dans l'Avant-Propos de *LGE* qui en dresse le projet (La Grande Encyclopédie, T1, p.XI), comme l'illustre la figure 1.2 [4]. D'ailleurs, si la liste des collaborateurs du projet n'associe pas chacun à une discipline précise, il est à noter que sur les 12 membres dont la qualité contient le mot «géographie», la moitié exactement d'entre eux sont des enseignants, agrégés ou professeurs, à la fois d'histoire et de géographie puisque les deux disciplines sont rassemblées à l'agrégation de 1831 à 1943, date à laquelle celle de géographie a été créée. Ce rapprochement est tout à fait cohérent avec le rôle que la Géographie joue à cette époque dans la constitution d'une identité nationale.

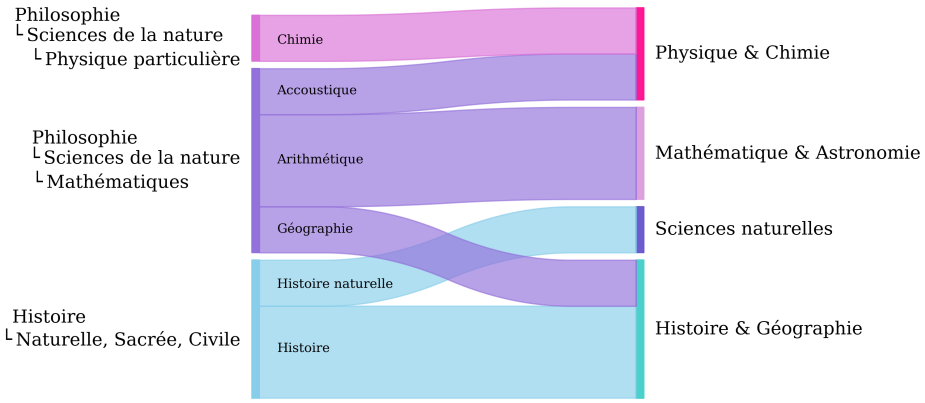


Figure 1.2 – Exemple schématique de rebrassage des domaines de connaissance

Pour reprendre la métaphore géographique de d'Alembert, les frontières des sciences bougent et se redécoupent donc et les présents travaux visent entre autres à rendre compte de ces mouvements. Pour avoir une chance d'atteindre leur objectif, ils devront donc éviter l'écueil consistant à prendre pour acquise la correspondance de deux ensembles d'articles au seul prétexte qu'ils sont regroupés sous la même étiquette à deux époques différentes et mettre d'abord en évidence assez de ressemblance pour s'assurer qu'il s'agit bien des mêmes objets. C'est seulement après avoir vérifié cette égalité «au premier ordre» que pourront se justifier des comparaisons ultérieures raffinant l'étude et la pertinence d'éventuelles différences trouvées. De plus, ils ne pourront pas se contenter de demeurer à l'intérieur des collections d'articles censés représenter la Géographie à chaque époque et devront nécessairement prendre en compte également ce qu'il y a «autour».

Mais puisqu'il faut – une fois les réserves levées – identifier la *Géographie* de l'*EDdA* et celle de *LGE*, il n'y a pas de sens à utiliser des jeux d'étiquettes différents aux deux époques pour représenter les domaines de connaissance. Et si c'est bien la nécessité technique qui motive initialement ce choix, les remarques ci-dessus montrent cette approche pragmatique sous un nouveau jour. Certes, appliquer la grille de lecture d'une époque à une autre va causer des erreurs, mais leur mesure attentive éclairera ces changements, alors qu'essayer de réconcilier deux réalités de toute façon disjointes n'aurait laissé aucune prise à l'analyse.

3. le Systême est surtout un manifeste dont les auteurs ont su s'émanciper là où il le fallait mais il témoigne en revanche assez bien des idéaux taxonomiques de ses auteurs

4. dans le préambule de l'*EDdA* déjà, un paragraphe dressait un parallèle entre *Chronologie*, «science des tems» et *Géographie* celle «des lieux». C'est peut-être là déjà un des premiers mouvements de ce qui deviendra l'*Histoire-Géographie* de *LGE*, mais les Encyclopédistes n'identifient pas *Chronologie* et *Histoire*, puisqu'ils font de celle-ci la mère de celle-là, attribuant en parallèle la même relation de parenté à la *Géographie* et à l'*Astronomie* respectivement.

1.3 Plan de l'étude

Les trois apports principaux de cette thèse concernent la préparation du corpus, un modèle de classification automatique et une étude des contenus géographiques à l'aide d'outils de textométrie. Ils correspondent à trois composantes présentes dans de nombreux projets en HN : le travail d'ingénierie nécessaire à la mise en forme des données, l'application de méthodes d'AA et enfin la conduite d'analyses expertes propres à une discipline, ici de la linguistique outillée puisqu'il s'agit de données textuelles.

Organisation

Hormis le chapitre consacré à l'état de l'art, le manuscrit est construit en suivant ces trois versants. S'il existe une relative dépendance entre ces composantes, le découpage des chapitres vise à leur conserver une certaine autonomie, bien que les premières contributions présentées soient davantage préparatoires et laissent progressivement la place aux analyses des données et métadonnées produites.

Le chapitre 2 dresse un portrait du domaine des Humanités Numériques. Il commence par décrire les formats utilisés, avant d'expliciter le lien privilégié qu'entretiennent les œuvres lexicographiques et la linguistique. Enfin, il décrit et replace dans leur contexte historique deux familles d'approches statistiques, en présentant les outils et les techniques couramment utilisées par chacune.

La majeure partie du travail de mise en forme et d'encodage du corpus fait l'objet du chapitre 3. La description détaillée des deux œuvres étudiées et de la structure des contenus que l'on peut y trouver est suivie d'une discussion des efforts entrepris pour représenter le corpus et les métadonnées qui lui sont associées. Le chapitre introduit également la notion de domaine de connaissance des articles, centrale dans le lien entre encyclopédies et géographie et sur laquelle repose un des axes d'investigation suivis dans les études contrastives.

L'importance de cette notion justifie la place centrale qui lui est accordée au chapitre 4 en tant que focale de tâches de classification. Après la description de l'entraînement de modèles d'AA capables de prédire le domaine de connaissance d'articles des deux encyclopédies étudiées à partir de leur contenu, le chapitre décrit l'application du classifieur retenu à la totalité du corpus afin de pouvoir ajouter aux métadonnées des articles le domaine dont ils relèvent.

Enfin, le chapitre 5 constitue l'aboutissement des deux précédents et se consacre aux analyses contrastives. Il commence par cerner la place accordée aux discours géographiques, de manière tout à fait quantitative d'abord puis, en s'intéressant aux transferts d'articles entre classes en diachronie, de façon plus qualitative. La dernière étude qu'il contient observe l'hétérogénéité des articles en termes de domaine de connaissance en prenant le cas notoire des biographies dans les articles de géographie.

Contributions

Les recherches présentées dans ce manuscrit se situent à l'interface de l'informatique et des sciences du langage mais la répartition des trois contributions entre ces deux sciences reste nettement en faveur de la première.

Version numérique structurée de *LGE*

Présentée au chapitre 3, la publication de la première version numérique complète de *LGE* structurée en XML-TEI représente la première contribution de cette thèse. Au-delà de l'objet produit, l'originalité de ce travail réside dans l'emploi de méthodes de graphes pour évaluer les possibilités offertes par un schéma XML.

Si les premiers succès sur ce versant avaient été obtenus dès 2021 à l'issue du projet DISCO-LGE [5], les analyses conduites sur *LGE* dans le cadre des présents travaux nécessitaient une amélioration de la qualité. La nouvelle version apporte plusieurs progrès à l'encodage de l'œuvre. D'abord, elle gagne en précision sur la segmentation des articles. Elle s'accompagne également de métadonnées permettant la navigation et le retour au texte.

Modèle de classification en domaines de connaissance

La deuxième contribution, développée dans le chapitre 4, concerne la tâche de classification automatique. Elle débute par une étude comparative de méthodes d'AA pour prédire le domaine d'articles encyclopédiques qui actualise et complète un travail similaire réalisé par Horton et al. (2009) en considérant notamment des méthodes d'Apprentissage Profond (AP).

Ensuite, au lieu de réduire les prédictions incorrectes des modèles à un horizon technique à dépasser, elle observe que ces erreurs ne ressemblent pas à un bruit aléatoire mais relèvent dans la plupart des cas d'une certaine logique. Elle les utilise comme un point d'entrée possible vers les articles en appliquant des méthodes algébriques – en particulier la notion de centralité – pour mettre en évidence des proximités thématiques ou stylistiques entre les domaines de connaissance.

Biographies et discours géographiques

Enfin le chapitre 5 apporte des éléments nouveaux pour comprendre la place des biographies dans les articles de *Géographie*. En effet, il n'y a pas officiellement de notice biographique dans les pages de l'*EDdA* et la biographie, genre discursif, ne s'intègre pas au «Systême» des domaines de connaissance des Encyclopédistes.

Plusieurs biographies comme celle de Newton sont pourtant bien connues des spécialistes de l'*EDdA*, spécifiquement sous la plume du Chevalier de Jaucourt, contributeur incontournable des entrées géographiques. En mettant en évidence un régime particulier pour les philosophes puis en définissant deux critères sur les articles contenant des biographies, l'étude propose un modèle progressif pour rendre compte des liens qui unissent les biographies et la Géographie.

Choix

La rédaction de ce manuscrit a fait l'objet d'un certain nombre de choix et de partis pris qu'il est bon d'avoir en tête avant d'entamer sa lecture et qui tiennent en partie à son caractère autoréférentiel. Au contact du genre encyclopédique, il a fini par en prendre certaines caractéristiques.

Le texte comporte des renvois fréquents, non seulement à des figures mais également à des sections ou à des passages précis dans le texte pour éviter de répéter des explications. La version PDF favorise naturellement une lecture transversale selon ces renvois grâce aux liens hypertextes qu'elle comporte ; pour faciliter une navigation semblable dans l'exemplaire papier, la plupart des renvois précisent le numéro de la page où se trouve la figure ou la section mentionnée. Ces numéros de page sont parfois omis pour les renvois «proches» à l'intérieur d'une même section ou à quelques pages d'intervalle.

En écrivant cette thèse il est en outre apparu que plusieurs sigles et acronymes mais aussi du lexique scientifique revenaient assez souvent pour justifier la création d'un court Glossaire alphabétique de la terminologie du manuscrit. À leur première occurrence, les termes sont explicités, leurs initiales étant précisées entre parenthèses pour les acronymes. Les occurrences suivantes n'utilisent plus que la version courte sous forme de lien hypertexte également. Ces liens, à la différence des renvois précédents, ne sont jamais accompagnés de la page où se trouve l'entrée : du vocabulaire qui ne

5. https://www.collexpersee.eu/projet/disco-lge/

serait pas défini juste avant son utilisation est à chercher dans le Glossaire à partir de la page 185. La version numérique est donc là encore avantagée puisque de nombreux lecteurs PDF affichent une prévisualisation de la cible d'un lien au survol avec la souris, ce qui suffit le plus souvent pour consulter ces définitions assez brèves. Le format PDF est donc celui recommandé pour lire cette thèse.

Le manuscrit ne possède en revanche pas d'Annexe alors que cela est courant dans ce type de document pour présenter des tableaux de données, des algorithmes utilisés, etc. À la place, puisqu'un certain nombre de traitements des données et de génération des figures est effectué automatiquement au moment de la génération de ce texte, le code source [6] utilisé tient lieu d'annexe (voir la section 3.3.3 sur la reproductibilité page 86 à ce sujet).

Enfin, du point de vue du langage, le texte reçoit par sa composante informatique l'influence de l'anglais, systématique dans cette discipline quelle que soient les langues maternelles des équipes qui publient alors que le français semble plus présent dans les publications de la communauté linguistique francophone. Le français est la langue du corpus, et le manuscrit a été rédigé en français. Pour la cohérence de l'ensemble, les termes français et les acronymes correspondants ont été préférés à ceux d'origine anglaise même d'un emploi répandu en France pour le vocabulaire technique, tant que cela ne donnait pas lieu à des termes trop artificiels. Certaines initiales peuvent toutefois paraître obscures quand l'habitude des termes anglophones a popularisé les leurs à la place d'un terme francophone pourtant utilisé. C'est le cas par exemple d'AA pour Apprentissage Automatique (l'expression développée est parfois utilisée en français mais l'abréviation ML pour «Machine Learning» est bien plus courante dans la littérature que «AA») ou pour HN qui peut surprendre par rapport à DH pour «Digital Humanities» alors qu'«Humanités Numériques» se rencontre couramment. Par souci d'homogénéité, ce sont pourtant les versions francophones qui ont été préférées ; dans tous les cas, le Glossaire est là (p.185) si besoin. Pour certains termes comme les noms de méthodes de classification automatique où aucune traduction ne parvient à s'imposer, des concessions à ce principe ont été faites.

6. https://gitlab.liris.cnrs.fr/abrenon/manuscrit

Chapitre 2

Un instantané des Humanités Numériques

2.1 Structuration et encodage des données . . . 18
2.1.1 Les formats XML . . . 18
2.1.2 Le format XML-TEI . . . 20
2.1.3 Succès et limites du format XML-TEI . . . 21
2.2 (Méta)lexicographie . . . 23
2.2.1 Portée conceptuelle des ouvrages lexicographiques . . . 23
2.2.2 Dictionnaires et encyclopédies . . . 25
2.2.3 L'Encyclopédie au centre des attentions . . . 26
2.3 Histoire et terminologie . . . 28
2.3.1 Rivalité historique entre les approches symbolique et statistique . . . 28
2.3.2 L'avènement des méthodes statistiques . . . 29
2.3.3 Nommer des pratiques très différentes . . . 29
2.4 Traitement Automatique de la Langue . . . 30
2.4.1 Vectorisation . . . 30
2.4.2 Classification . . . 31
2.4.3 Étiquetage morphosyntaxique et syntaxique . . . 34
2.5 Linguistiques de corpus . . . 36
2.5.1 Du texte au corpus . . . 36
2.5.2 Les outils de la linguistique . . . 38
2.5.3 Études contrastives . . . 40

Les travaux présentés dans ce manuscrit reposent sur des considérations issues d'un spectre de disciplines centré sur l'informatique (encodage, complexité algorithmique, AA) mais allant des mathématiques (statistiques, théorie des graphes) à la linguistique (syntaxe, cooccurrences). Cette pratique pluridisplinaire constitue une illustration représentative des HN, un courant interdisciplinaire né à la fin du XX^{ème} siècle, ou au moins constitué sous ce nom à cette époque à partir de la pratique en développement depuis les années 1960 consistant à appliquer des techniques de calcul automatique à des thématiques des humanités (Beaudouin, 2016, p. 17). Cette rencontre entre l'informatique d'une part et les sciences humaines et sociales d'autre part a rendu possible l'émergence de projets ambitieux, capables d'aborder de nouveaux objets d'études et de révéler ceux déjà connus sous un jour nouveau.

L'encodage des données pour permettre leur manipulation par des outils informatisés, rappelé par

l'adjectif «Numérique» dans l'expression «Humanités Numériques» est un point commun à tous les projets du domaine. La première section de ce chapitre s'intéresse aux efforts entrepris pour standardiser les formats utilisés dans le but de rendre les fichiers produits facilement exploitables, interopérables et réutilisables. La place privilégiée qu'occupent les ouvrages lexicographiques dans les HN et leur lien étroit avec les questionnements linguistiques sont ensuite traités à la section 2.2. Des travaux existants sur les dictionnaires et les encyclopédies y sont présentés, notamment sur l'*EDdA*, œuvre majeure qui intéresse de nombreuses branches des HN. Puis la section 2.3 dresse un bref historique du pan informatique de la discipline et, enfin, les sections 2.4 et 2.5 présentent les appareils conceptuels, formats et outils déployés respectivement dans les chapitres 4 (classification automatique) et 5 (linguistique de corpus).

2.1 Structuration et encodage des données

Les formats de fichiers les plus manipulés au cours de cette thèse appartiennent principalement à deux familles présentes très fréquemment en HN : ceux structurés en tableaux et ceux possédant des structures arborescentes représentées par un langage de balisage.

Les métadonnées des fichiers du corpus ont été représentées par des formats tabulaires de la famille CSV, un format très simple en usage depuis le début des années 1970 et formalisé bien plus tard (Shafranovich, 2005). L'acronyme signifie *Comma-Separated Values* («des valeurs séparées par des virgules») qui décrit la manière d'organiser les cellules au sein d'une ligne (les lignes elles-mêmes sont séparées par un caractère «nouvelle ligne» `'\n'`). De nombreuses variantes de ce format utilisant différentes conventions sont en usage, si bien qu'on parle également parfois de «SSV», *Something-Separated Values* («des valeurs séparées par quelque chose») pour désigner plus généralement cette famille où le caractère délimiteur peut varier. La variante utilisée en pratique dans les présents travaux est TSV qui se sert du caractère tabulation (`'\t'`) pour séparer les cellules.

L'autre format principal dans cette thèse est le format XML-TEI, un langage à balises (*markup language*, c'est l'origine de la séquence «ML» dans les noms de langage mentionnés dans cette partie). Ce format possède une histoire bien plus riche ponctuée d'étapes importantes de normalisation et de standardisation qu'il faut expliciter pour donner un sens aux travaux exposés dans la section 3.2. Il joue un rôle majeur actuellement en HN et c'est à lui que s'intéresse cette section.

2.1.1 Les formats XML

Les langages de balises constituent des outils précieux pour représenter des informations structurées et détaillées. Dans cette famille, la lignée du langage SGML – dont les origines remontent via COCOA et GML jusqu'aux années 1960 – se distingue par sa grande fertilité en ayant donné naissance à un très grand nombre de formats dont une partie est visible à la figure 2.1. Ces formats trouvent des applications dans des domaines très variés comme le dessin vectoriel pour SVG, la messagerie instantanée pour XMPP et bien sûr les pages web pour le format HTML.

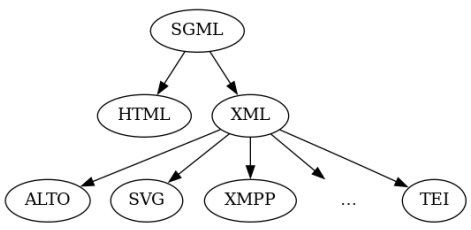


Figure 2.1 – Quelques membres éminents de la famille SGML

SGML (Standard Generalized Markup Language, 1986) est en fait un métalangage, c'est-à-dire qu'il

décrit une syntaxe pour définir des structures de fichier plutôt que de donner des règles sur un format précis. Il se caractérise par l'emploi de chevrons (ouvrant `'<'` et fermant `'>'`) et d'une barre oblique (`'/'`) pour donner un sens spécial à une séquence de caractères, en vert sur la figure 2.2. C'est ainsi qu'il est possible de créer des *balises*, ouvrantes (figure 2.2a) et fermantes (2.2b), dont le nom est défini par la séquence entre chevrons (et éventuellement la `'/'` pour une balise fermante). Une balise ouvrante et une balise fermante de même nom s'associent pour former ensemble l'équivalent paramétrable d'une paire de parenthèses servant à délimiter une portion du document : un *élément* comme celui visible à la figure 2.2c. Il est en outre possible d'ajouter des propriétés aux balises à l'aide d'attributs qui associent des *valeurs* à des *clefs*. Sur la même figure 2.2c, la clef `id` est ainsi associée à la valeur `doc0` (le symbole `=` est syntaxique et ne fait pas partie de la clef ni de la valeur, de même que les guillemets à l'intérieur desquels doit être placée la valeur).

```
<text>
```

(a) Une balise ouvrante (dont le nom est «text»)

```
</author>
```

(b) Une balise fermante (dont le nom est «author»)

```
<p id="doc0">
Contenu
</p>
```

(c) Une paire de balises «p» délimitant le texte «Contenu» et dont l'attribut «id» vaut «doc0»

Figure 2.2 – Des balises SGML

Quand un élément ne contient ni texte ni aucun autre élément – un cas qui peut se rencontrer assez régulièrement, un élément vide reste utile par rapport à son élément parent ou pour ses attributs – il est également possible d'utiliser une syntaxe dite «autofermante» qui permet de le représenter en une seule balise. Ainsi un élément `<p></p>` peut se noter `<p/>`. C'est cette notation qui sera utilisée dans toute la thèse pour désigner les éléments afin de distinguer la balise et son nom. Ainsi `<p/>` pourra être utilisé pour parler de l'entité abstraite que constitue un élément (la classe de tous les éléments XML qu'il est possible de construire avec des balises dont le nom est «p») aussi bien que pour une de ses instances particulières («l'élément racine de la figure 2.2c est un `<p/>`»).

Le langage XML (Bray et al., 2008) forme un sous-ensemble de SGML qui ajoute des contraintes pour rendre le langage plus simple à analyser automatiquement mais il reste lui aussi au niveau de la structure. Pour définir un langage concret utilisable dans un cadre précis pour représenter un certain type de données, il faut décrire un ensemble de balises, d'attributs et de règles indiquant quelles balises peuvent en inclure combien de quels autres types, et quels attributs elles peuvent recevoir (affectés de quelles valeurs). Ces informations constituent un *schéma* de données, qui est spécifié historiquement sous la forme d'une *Document Type Definition* (Bray et al., 2008, sec. 2.8), format de référence des schémas défini en même temps que XML lui-même. D'autres formats plus récents et plus expressifs ont vu le jour pour décrire des schémas : XSD (Gao et al., 2012) publié également par le consortium *W3C* à l'origine de XML, Schematron (Jelliffe, n.d.) ou encore RELAX NG (Clark & Murata, 2001).

Deux formats XML définis par des schémas particuliers, ALTO et TEI, sont principalement utilisés dans les présents travaux. Le format ALTO (*Analyzed Layout and Text Objects*, c'est-à-dire «Mise en page analysée et objets textuels») est publié par la Librairies du Congrès des États-Unis d'Amérique (*ALTO*, 2023). Il sert à représenter des pages numérisées après l'application de systèmes de reconnaissance optique de la disposition (*Optical Layout Recognition* ou OLR) puis de reconnaissance optique de caractères (*Optical Character Recognition* ou OCR). Il contient essentiellement des éléments de structure à différents niveaux (page, bloc de texte, ligne, mot) comme le montre l'extrait de la figure 2.3. L'imbrication de ces éléments modélise la structure du document représenté, et leurs attributs donnent des informations sur leurs coordonnées pour les positionner sur la page (`HPOS` et `VPOS`) et

sur les séquences de caractères lues (`CONTENT` pour le contenu lui-même et `WC` – *word confidence* ou «confiance dans le mot» – qui donne un indice de qualité d'après l'OCR). Le standard, qui a atteint sa version 4.4 est spécifié par un schéma XSD [1]. Il est utilisé par exemple pour encoder des journaux (Gutehrlé & Atanassova, 2022) et sert de format de référence pour des outils développés dans le but d'améliorer la qualité d'OCR (Maurer et al., 2023).

```
<Layout>
  <Page ID="P44" PHYSICAL_IMG_NR="44" HEIGHT="4884" WIDTH="3353" ACCURACY="0.7655755"
        QUALITY_DETAIL="1228 words, 6373 chars, 4988 with recognition greater than 50% (78,27%)">
    <TopMargin ID="P44_TM00001" HPOS="0" VPOS="0" WIDTH="3353" HEIGHT="509"/>
    <LeftMargin ID="P44_LM00001" HPOS="0" VPOS="509" WIDTH="433" HEIGHT="3555"/>
    <RightMargin ID="P44_RM00001" HPOS="2661" VPOS="509" WIDTH="692" HEIGHT="3555"/>
    <BottomMargin ID="P44_BM00001" HPOS="0" VPOS="4064" WIDTH="3353" HEIGHT="820"/>
    <PrintSpace ID="P44_PS00001" HPOS="433" VPOS="509" WIDTH="2228" HEIGHT="3555">
      <TextBlock ID="P44_TB00001" HPOS="1435" VPOS="513" WIDTH="201" HEIGHT="50" IDNEXT="P44_TB00002">
        <TextLine ID="P44_TL00001" HPOS="1443" VPOS="521" WIDTH="184" HEIGHT="34">
          <String ID="P44_S00001" HPOS="1443" VPOS="540" WIDTH="44" HEIGHT="8" STYLEREFS="StyleId-0" CONTENT="-" WC="0.1" CC="9"/>
          <SP ID="P44_SP00001" WIDTH="0" HPOS="0" VPOS="0"/>
          <String ID="P44_S00002" HPOS="1514" VPOS="521" WIDTH="39" HEIGHT="34" STYLEREFS="StyleId-1" CONTENT="11" WC="0.255" CC="68"/>
          [...]
        </TextLine>
      </TextBlock>
    </PrintSpace>
  </Page>
</Layout>
```

Figure 2.3 – Un court extrait d'un document XML-ALTO

Dans le cadre de cette thèse, XML-ALTO représente le format d'entrée dans lequel une des deux œuvres du corpus, *LGE*, est représentée. C'est à partir de ce format qu'il a fallu travailler pour gagner l'accès au texte de l'œuvre, opération décrite à la section 3.1.3, en vue de pouvoir en produire un encodage dans le format de référence pour le corpus d'étude : XML-TEI.

2.1.2 Le format XML-TEI

De la même façon qu'ALTO, XML-TEI est donc un dialecte du XML obtenu en définissant un schéma. Celui du format TEI est publié et maintenu par la *Text Encoding Initiative* [2], un collectif de volontaires issus du monde de la recherche. Cette organisation a été créée en 1988 (Huitfeldt, 1994, p. 237 ; Ide, 1995, para. 5) suite aux principes adoptés en novembre 1987 à la conférence de Vassar (Ide & Sperberg-Mcqueen, 1995, p. 2). La TEI publie ses premières recommandations en novembre 1990 (Ide & Sperberg-Mcqueen, 1995, p. 11), la TEI P1 (*Proposal 1*, «proposition n°1»). Les années suivantes voient se poursuivre l'essentiel des travaux de normalisation et deux autres versions (P2 et P3) sont publiées, respectivement en avril 1992 et à l'automne 1993 (Ide & Sperberg-Mcqueen, 1995, p. 12). La TEI est alors stabilisée pour l'essentiel et le rythme de sortie des recommandations diminue fortement.

Fondée dès l'origine sur le standard SGML, la TEI est reformulée en termes de XML peu après sa normalisation (les premières versions de XML datent de 1996), conduisant à la TEI P4 publiée en juin 2002. L'étape suivante, la TEI P5, sortie le 2 novembre 2007 est toujours la version officielle du consortium TEI et continue d'être mise à jour (la version 4.8.0 [3] a été publiée le 8 juillet 2024).

Un des objectifs à l'origine du projet est de proposer un format unifié pour favoriser la réutilisabilité des données de recherches et éviter l'apparition d'encodages incompatibles au fil des projets, une démarche en accord avec ce qui sera par la suite formalisé sous le nom des principes FAIR [4] pour des données découvrables, accessibles, interopérables et réutilisables (*Findable*, *Accessible*, *Interoperable* and *Reusable*). Dès ses débuts, le consortium vise un format unique pour représenter tous types de

1. https://www.loc.gov/standards/alto/v4/alto.xsd
2. https://tei-c.org/
3. https://tei-c.org/Vault/P5/current/doc/tei-p5-doc/readme-4.8.0.html
4. https://www.go-fair.org/fair-principles/

documents – nativement numériques, imprimés, manuscrits ou même épigraphes – dans des genres allant du théâtre aux écrits religieux en passant par les dictionnaires et la correspondance, en mêlant potentiellement plusieurs langues à la fois. Les documents peuvent être représentés avec une grande finesse, en marquant par exemple des passages rayés ou reformulés ou en faisant coexister dans un même document des versions alternatives d'un texte à des fins de comparaisons. Il est également possible d'enrichir un document d'annotations compilées dans le cadre des recherches comme des parties de discours, des liens entre les entités mentionnées dans le texte et une base de connaissance voire des informations géo-sémantiques (Moncla & Gaio, 2023). Ces couches additionnelles donnent une profondeur supplémentaire aux textes et permettent leur analyse suivant différents axes, allant au-delà de l'unidimensionalité imposée par le format numérique (Huitfeldt, 1994, p. 236).

Pour rendre cela possible sans trop accroître la complexité du schéma, la TEI opte pour une approche modulaire combinant des composantes très génériques communes à tout type de documents tels que les modules *core* et *tei* et d'autres au contraire très spécialisées et adaptées aux besoins spécifiques de certaines disciplines. La TEI P5 comporte ainsi 587 éléments répartis en 21 modules et définit 84 attributs pour ces éléments [5]. Cette modularité est implémentée en morcelant le schéma en plusieurs fichiers indépendants qui peuvent être inclus ou non sans remettre en cause la cohérence de l'ensemble. Le consortium TEI a mis en ligne l'interface ROMA [6] pour permettre de générer des schémas personnalisés adaptés aux besoins de chaque projet et il existe même une telle personnalisation du schéma TEI dont le but est d'aider à en écrire d'autres directement dans un éditeur XML sans avoir besoin de ROMA (Bauman, 2019). Des schémas alternatifs au schéma standard de la TEI sont repartagés au sein de la communauté et rencontrent une certaine popularité comme le *Comic Book Markup Language* (CBML, le «langage à balises pour les bandes dessinées») de Walsh (2012) ou la TEI Lex-0 (Romary & Tasovac, 2018) pour les ouvrages lexicographiques.

Un module est présent depuis la TEI P3 (Ide & Véronis, 1995 ; Sperberg-McQueen & Burnard, 1999) : le module *dictionaries*. Il comprend 33 éléments et se focalise sur les entrées, niveau à partir duquel les ouvrages lexicographiques diffèrent des autres types d'œuvres (Ide & Véronis, 1995, p. 168). Il permet également de représenter les sous-entrées, objets utilisés par les lexicographes pour rendre compte de la polysémie des mots. Ce module a servi de point de départ à la détermination d'un encodage pour représenter *LGE* (travail présenté à la section 3.2).

Les besoins de la TEI en outillage spécifique restent modestes : s'appuyant sur des technologies standards (SGML puis XML) pour lesquelles existent de nombreux outils, la communauté a accès à un riche panel d'outils pour travailler sur ses fichiers. Il existe ainsi déjà des éditeurs de texte spécialisés pour les langages de balisage, des moteurs de validation vérifiant la conformité d'un document à un schéma donné ainsi que des feuilles de transformation XSLT (Kay, 2017) permettant d'appliquer des règles de réécriture automatique pour changer la structure d'un document et éventuellement le convertir vers ou à partir d'autres formats XML ou HTML. D'autres outils de conversion comme pandoc [7] peuvent prendre en entrée d'autres formats que des dialectes SGML et permettent de produire de la TEI automatiquement à partir de fichiers dans de nombreux formats tels que Markdown, HTML, LaTeX et même ODT (format de traitement de texte).

2.1.3 Succès et limites du format XML-TEI

La communauté qui s'organise autour du format TEI possède sa propre revue, le *Journal of the Text Encoding Initiative* (JTEI) [8] depuis le début des années 2010 (Schreibman, 2011). Le volume de données encodées en XML-TEI a connu une croissance rapide au cours des années 1990, pour dépasser

5. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html
6. https://roma.tei-c.org/
7. https://pandoc.org/
8. http://jtei.revues.org/

les centaines de millions de mots (Ide & Sperberg-Mcqueen, 1995, p. 5).

Le format s'est imposé comme un standard incontournable dans de nombreux logiciels en HN. C'est par exemple le format utilisé en interne par le logiciel d'analyse textométrique TXM (Heiden et al., 2010) pour représenter les textes d'un corpus. Le format d'entrée pour importer les corpus dans le Lexicoscope (Kraif, 2016) est également basé sur le format XML-TEI sans en respecter exactement le schéma : les fichiers ne possèdent pas l'en-tête minimal requis et utilisent certains éléments non conformes comme `<doc/>` ou `<meta/>`. De plus, au lieu de représenter directement un texte, les éléments `<meta/>` et `<s/>` contiennent des «îlots» d'autres formats de données, respectivement TSV et CoNLL-U (voir la section 2.4.3 à propos de ce format). Enfin, des interfaces de visualisation existent pour exposer des corpus encodés en TEI sur le web. Le projet TAPAS [9] permet ainsi de parcourir des fichiers dans ce format en alternant entre différents modes d'affichage, de l'édition diplomatique au code XML des fichiers. Philologic [10](Allen et al., 2013) développé à l'ARTFL permet non seulement d'afficher les fichiers d'un corpus au format TEI mais offre en outre la possibilité d'effectuer des requêtes sur leurs contenus et de visualiser les résultats dans un concordancier simple avec un contexte autour des motifs trouvés.

La version de l'*EDdA* de l'ARTFL est encodée en TEI et publiée à l'aide de Philologic4. D'autres projets portant sur des ouvrages lexicographiques ont utilisé le même format pour représenter leurs corpus. C'est le cas par exemple de Nenufar [11] qui s'intéresse à plusieurs éditions du *Petit Larousse* (Bohbot et al., 2019) qu'il permet de comparer dans une interface dédiée ou de BASNUM [12] qui s'intéresse ainsi au *Dictionnaire Universel* de Furetière, ou plutôt à sa seconde édition, améliorée par Henri Basnage de Beauval (Galleron & Williams, 2022 ; Williams, 2017). Le format est répandu pour des textes de différentes époques et l'énumération qui suit se contente de citer quelques exemples parmi la grande variété qui existe. Le format TEI a ainsi été utilisé pour encoder des textes du moyen-âge comme les «Chroniques rimées du Mont Saint-Michel» [13] (Saint Pair, 2022) ou la Base du Français Médiéval [14], mais aussi des textes de la Renaissance avec le corpus Epistemon [15] qui utilise Philologic4 et la version web de TXM pour proposer des interfaces publiques de parcours et de requête des textes. Dans d'autres langues que le français, Salgado et al. (2024) ont publié une version en TEI Lex-0 du *Diccionario da Lingua Portugueza* d'António de Morais Silva (fin du XVIII^{ème} siècle); il existe également une édition de la correspondance de Mary Hamilton [16]. L'*Oxford Text Archive* [17] des *Bodleian Libraries* rassemble des textes dans plus de 25 langues différentes sur plus de 2 000 ans. Enfin, le projet Perseus [18] présente la spécificité d'intégrer les œuvres qu'il étudie dans un système d'information géographique (GIS) à l'aide d'une couche d'annotation géo-sémantique riche.

Malgré ces succès et une large adoption, certaines réserves ont toutefois été émises à l'encontre de la TEI en général et du module *dictionaries* en particulier. Ainsi, Ide & Véronis (1998, p. 6) déplorent que le haut degré de généricité de la TEI entraîne une forte abstraction qui complique l'encodage. De plus, la très grande flexibilité de la TEI offre souvent une multiplicité de solutions pour représenter un phénomène donné dans une œuvre. Si cette souplesse peut être vue comme un avantage certain, elle accroît les risques que des projets différents utilisent des conventions incompatibles. Le reproche sur la complexité est partagé par Tutin & Véronis (1998, p. 364) dans leurs travaux sur l'encodage du *Petit Larousse Illustré*.

9. https://tapasproject.org/
10. https://artfl-project.uchicago.edu/philologic4
11. http://nenufar.huma-num.fr/presentation/
12. https://anr.fr/Projet-ANR-18-CE38-0003
13. http://catalog.bfm-corpus.org/ChronSMichelBo
14. https://txm-bfm.huma-num.fr/txm/
15. https://www.bvh.univ-tours.fr/Epistemon/index.asp
16. https://www.maryhamiltonpapers.alc.manchester.ac.uk/edition/overview/
17. https://ota.bodleian.ox.ac.uk/repository/xmlui/
18. https://www.dlib.org/dlib/july00/crane/07crane.html

Enfin malgré l'expressivité de la boîte à outils que représente la TEI dans son ensemble, il est à noter que la version de l'*EDdA* encodée par l'ARTFL n'utilise pas le module *dictionaries*. Ce point ne constitue pas en soi une critique mais conduit à s'interroger sur l'expressivité du module *dictionaries* et sa pertinence pour les œuvres encyclopédiques. Quelques angles morts demeurent dans ses règles, empêchant parfois l'utilisation d'éléments simples dans certains contextes, un point que n'hésitent pas à reconnaître Ide & Véronis (1995, pp. 172–173), encourageant même à ne pas respecter tout à fait le schéma du module quand cela est nécessaire. Il est ainsi impossible d'utiliser l'élément `<p/>` représentant les paragraphes dans une définition (`<def/>`) comme le rapportent Hagen et al. (2020, p. 6). Parmi les autres projets cités précédemment, Nenufar est moins concerné par ces limites d'expressivité car le *Petit Larousse* qu'il traite et encode est un dictionnaire et ne possède pas de développements structurés sur plusieurs niveaux comme ceux des encyclopédies. Le projet prend le parti d'utiliser l'élément `<note/>` pour représenter les quelques paragraphes présents dans certaines entrées (voir par exemple l'entrée AIR [19]) en dépit de la sémantique qui lui est associée par la documentation. Le projet BASNUM le rejoint sur ce choix qui, avec les problèmes d'expressivité du module *dictionaries* en général, seront développés à la section 3.2.

2.2 (Méta)lexicographie

En s'intéressant à l'encodage en XML-TEI, la section précédente a montré qu'un module spécifique était dédié aux dictionnaires et mis en évidence plusieurs projets consacrés à l'étude de ce type d'ouvrages ou aux encyclopédies. Au-delà du seul prisme des présents travaux, le genre lexicographique semble occuper une place particulière dans les HN et jouer un rôle fondateur pour la linguistique de corpus. La section qui débute tâche de cerner ces liens.

2.2.1 Portée conceptuelle des ouvrages lexicographiques

Dictionnaires et encyclopédies représentent, malgré leur apparence de simple table associative entre termes et définitions, des objets intrinsèquement complexes, dont la difficulté d'encodage a pu dérouter des spécialistes des HN et de la lexicographie tels que Ide & Véronis (1995, p. 177). L'identification de ces complexités et leur prise en compte dans le schéma de la TEI s'est montrée fructueuse non seulement pour ce genre mais aussi pour les autres (Ide & Véronis, 1995, *ibid*) ce qui constitue un premier signe de l'importance du genre lexicographique en HN.

La structure associative de ces ouvrages implique également une discontinuité dans «le» texte. Un dictionnaire rassemble en réalité un *ensemble* de textes distincts dont l'adjacence apparente n'est due qu'à l'arbitraire de l'ordre alphabétique et qui ne se lisent pas linéairement du début à la fin (Ide & Véronis, 1995, p. 167). La dernière phrase d'un article ne vient pas «avant» la première de l'article suivant ; chacune existe dans son propre espace.

Ces textes sont séparés mais pas indépendants : des renvois invitent à un parcours transverse des entrées basé typiquement sur la synonymie dans les dictionnaires et des proximités thématiques dans les encyclopédies (Blanchard & Olsen, 2002, p. 46). Cette structure transversale se prête bien à des études de graphes surtout pour les dictionnaires de synonymes (Ploux & Victorri, 1998). D'autres approches basées seulement sur les occurrences de mots dans les entrées de dictionnaires de langues bénéficient également des apports de la théorie des graphes (Loiseau et al., 2011).

Mais à cette seule complexité structurelle s'ajoute une portée conceptuelle extraordinaire, autant pour les dictionnaires de langues que pour ceux de choses, pour suivre la distinction couramment faite. Les premiers, en ayant pour but de définir les mots, constituent un relevé de signes : la vedette qui sert de clef aux articles représente un signifiant, son signifié étant apporté ensuite par la

19. http://nenufar.huma-num.fr/?article=1211

définition qui forme le reste de l'entrée. Les énoncés qu'on trouve dans les dictionnaires de langue ne portent pas tant sur le monde que sur le langage lui-même d'après Sinclair (1966, p. 6). Le choix du lexique lui-même reflète la façon de parler puisque «la nomenclature d'ensemble est imposée par l'usage de la langue» (A. Rey, 2006, p. 39). Les définitions dans les dictionnaires de chose au contraire donnent «accès au domaine à décrire», toujours selon A. Rey (2006, *ibid*) et, en cela, concernent davantage le référent des signes. Dans les deux cas le dictionnaire constitue ainsi en lui-même un «traité métaphysique sur le sens» (Willinsky, 2001, pp. 189–190) : le genre lexicographique intéresse donc la sémantique la plus abstraite jusqu'à la logique et la philosophie. Le problème du sens des mots représente un sujet de réflexion vertigineux. D'Alembert avait déjà compris son ampleur en écrivant à l'article DICTIONNAIRE (L'Encyclopédie, T4, p.958) :

> un dictionnaire de langues, qui paroît n'être qu'un dictionnaire de mots, doit être souvent un dictionnaire de choses quand il est bien fait : c'est alors un ouvrage très-philosophique.

Il n'est pas possible de définir les mots en demeurant au niveau du seul langage, sous peine de produire des tautologies selon Haiman (1980, p. 330), ce qui l'amène à attaquer lui aussi la distinction faite entre dictionnaires et encyclopédies. Russell (1940, p. 25) en adoptant une approche parallèle à la déduction mathématique basée sur des hypothèses, distingue des mots qui ne peuvent s'expliquer qu'au moyen d'autres mots et des mots dont le sens provient au contraire de l'expérience sensible du monde : les «mots-objets», qui ne présupposeraient aucun autre mot (Russell, 1940, p. 26) et joueraient ainsi le rôle d'axiomes du langage (tant du point de vue logique que de celui de la chronologie individuelle puisqu'il écrit que ce seraient les premiers mots appris par les enfants lors de l'acquisition du langage). S'il reste possible de définir certains d'entre eux en intension (au prix d'une description introduisant plusieurs nouveaux mots pour chacun) d'autres, typiquement ceux rattachés à des qualia comme la couleur ne peuvent espérer recevoir qu'une définition extensionnelle (Lehmann & Martin-Berthet, 2018, chap. 2, §2) nécessairement incomplète (il est possible d'énumérer *des* objets de couleur rouge mais pas tous ces objets). Mais sans explications ni liens, une expérience du monde ne constitue pas une définition, ce qui fait écrire à Teubert (2005, p. 137) que les descriptions des livres d'images pour enfant jouent sans doute un rôle supérieur à celui des dictionnaires dans l'acquisition du langage.

Élève de Russell, Wittgenstein ne se satisfait pas des «mots-objets» et refuse le statut d'évidence que Russell accepte de leur concéder. Pour lui, le vrai sens des mots échappe nécessairement aux définitions du dictionnaire car elles établissent des frontières nettes entre eux alors que la réalité est bien plus floue : les mots tirent selon lui leur signification de la manière dont ils sont employés dans le langage (Willinsky, 2001, p. 196). À la suite de Wittgenstein, Firth (1957, p. 11) base le sens des mots sur leur relation avec la notion de *collocations* :

> You shall know a word by the company it keeps !

(«vous reconnaîtrez un mot à ses fréquentations»). Harris (1954, p. 146) se propose même de quantifier ces fréquentations en comptant les occurrences conjointes de différentes classes de mots dans une approche distributionnelle. Cette approche ainsi que l'aphorisme de Firth sont à la source de l'application de méthodes statistiques à la linguistique et trouvent un écho jusque dans la façon de modéliser les mots dans la plupart des outils actuels (voir les sections 2.4 et 2.5). L'héritage de Firth, en passant par Halliday puis Sinclair (Léon, 2015, p. 161) développe une pratique centrée sur l'étude des collocations d'un point de vue probabiliste et distributionnel qui aboutira plusieurs décennies plus tard à la conception de dictionnaires radicalement nouveaux (Krishnamurthy, 2008, p. 239).

2.2.2 Dictionnaires et encyclopédies

Identifier des collocations exige des corpus assez grands (Teubert, 2005, p. 134) pour pouvoir observer assez d'occurrences des phénomènes étudiés pour qu'elles puissent être significatives statistiquement, et c'est pourquoi l'approche de ce qui restera connu sous le nom d'*École britannique* a pu révolutionner la lexicographie une fois que les capacités des ordinateurs leur ont permis de traiter automatiquement du texte. La section 2.3 revient plus en détail sur les grandes étapes qui ont jalonné le XX^{ème} siècle dans ce domaine pour donner un sens plus précis à l'expression «Traitement Automatique des Langues» (TAL) mais l'acronyme sera utilisé dès la présente section pour désigner tous les traitements statistiques de la langue opérés à l'aide d'ordinateurs.

Équipé de ces moyens nouveaux, Sinclair (1990) concrétise en quelque sorte les idées de Wittgenstein avec le projet COBUILD qui aboutit à la création d'un dictionnaire dont les définitions sont fondées sur des observations d'occurrences réelles trouvées dans des grands corpus de textes. Dans d'autres pays aussi des projets similaires voient le jour.

En France, le *Trésor de la Langue Française*[20] (TLF) est lancé au début des années 1960 (Surmont, 2006, p. 56) sur un corpus reprenant les données de Frantext. Il comporte 270 000 définitions et 430 000 exemples et emploie dès ses débuts des techniques de TAL et des ordinateurs Bull Gamma 60 (Surmont, 2006, p. 57). Les français restent toutefois à la traîne sur les britanniques et les américains sur le versant informatique (Léon, 2015, p. 159) et ces moyens ne sont déployés qu'en vue d'écrire le contenu du dictionnaire dont la finalité est d'être imprimé, l'idée d'un dictionnaire comme fichier étant reléguée à un futur relevant de la science-fiction (Surmont, 2006, *ibid*). L'informatisation du TLF (le TLFI) n'interviendra que bien plus tard dans les années 1990, et il est frappant de constater que la première version mise en ligne l'a été grâce à l'ARTFL, en 1998 (Surmont, 2006, p. 57). En Allemagne, Teubert (2005, p. 127) rapporte l'existence d'un dictionnaire basé sur un corpus, le *Schlüsselwörter der Wendezeit* paru en 1997.

La linguistique de corpus nourrit et guide donc la conception de dictionnaires, mais à l'inverse les dictionnaires sont des objets utiles pour le TAL. Il y a donc deux mouvements contraires simultanés, mais en reprenant la distinction faite par Lehmann & Martin-Berthet (2018, p. 255) il est même possible de diviser le mouvement «retour» de l'apport des œuvres lexicographiques au TAL en deux flux comme le montre la figure 2.4 : les produits de la lexicographie sont nécessaires dans de nombreuses branches du TAL, mais les dictionnaires et les encyclopédies constituent également un genre dont les œuvres sont étudiées pour elles-mêmes, ce que les deux autrices distinguent sous le nom de métalexicographie.

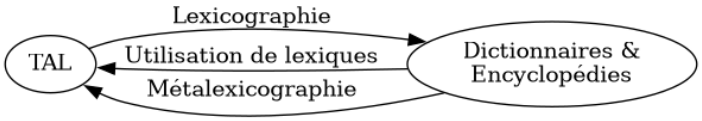


Figure 2.4 – La triple relation qu'entretiennent le TAL et les œuvres lexicographiques

Le premier de ces deux mouvements concerne l'emploi de lexiques et de bases associatives de mots en TAL, qui semble remonter à ses origines. En 1959, alors qu'il est encore anachronique de parler de TAL et que l'ensemble des activités qui relèveraient de ce domaine se résume à la Traduction Automatique (TA), des objets comme la Machine de Trojanski reposent sur des dictionnaires pour des tâches de traduction automatique depuis et vers un langage symbolique servant de passerelle centrale entre les langues naturelles (Léon, 2015, pp. 99–100). Plus récemment de nombreux outils font usage de lexiques pour la segmentation du texte en tokens ou leur étiquetage morphosyntaxique (Léon, 2015, p. 143). C'est le cas par exemple de la chaîne de traitement utilisée dans le projet PRESTO (Diwersy

20. http://atilf.atilf.fr/

et al., 2017, p. 33) et d'outils comme Talismane (Urieli & Tanguy, 2013, p. 191) ou Stanza [21] (Qi et al., 2020). Leur emploi est tel que de nombreux efforts ont été déployés pour rendre l'information contenue dans des dictionnaires accessible à des traitements automatiques avec des succès parfois en dessous des attentes initiales selon Ide & Véronis (2000, p. 9).

La métalexicographie s'est quant à elle développée en tant que discipline à part entière parmi les sciences du langage depuis la fin des années 1960 (Vigier, 2022, p. 9). Zhu (2022, p. 129) fait observer que les dictionnaires constituent «un excellent terrain d'expérimentation» et les encyclopédies représentent également des sources d'information précieuses non seulement pour les nombreuses disciplines dont elles traitent mais aussi pour la philosophie et l'histoire des sciences. La métalexicographie partage donc ses objets d'étude – les répertoires alphabétiques – avec la lexicographie historique dont elle diffère cependant par ses méthodes empruntées à la linguistique outillée. C'est dans ce rapport entre encyclopédies et TAL que s'inscrivent les présents travaux.

Des projets de recherche couvrent les époques marquant des phases majeures dans l'évolution des dictionnaires et des encyclopédies. Quemada (1968) remonte ainsi jusqu'au *Dictionnaire françois-latin* de Robert Estienne, en 1539, premier relevé alphabétique de mots en français à offrir, outre leur traduction latine, une définition rédigée en français (Quemada, 1968, p. 12). L'apparition d'un genre encyclopédique en tant que tel est un phénomène progressif mais un palier important a été atteint avec le *Dictionnaire Universel* de Furetière en 1690 qui dépasse le cadre du seul langage pour s'intéresser aux choses (Leca-Tsiomis, 2009, paras. 5–9). Bien qu'on puisse lui reprocher certaines lacunes, sa deuxième édition, parue en 1701 et évoquée à la section 2.1.3, est largement revue et augmentée par Basnage de Beauval ce qui achève d'offrir au *Furetière* sa place dans l'histoire de l'encyclopédisme en Europe et en fait un précurseur important de l'*EDdA*. Le mouvement se poursuit dans le reste du XVIII^{ème} siècle, période très étudiée pour la richesse des innovations dans les dictionnaires universels et les encyclopédies (Leca-Tsiomis, 1995). La première édition du *Dictionnaire Universel François et Latin* (*Dictionnaire de Trévoux* ou juste *Trévoux* dans le reste de cette thèse) est publiée par les jésuites (catholiques au contraire de Basnage qui était protestant) en 1704, peu après la parution de la seconde édition du *Furetière*, celle de Basnage, dont il reprend assez d'éléments pour être accusé de plagiat (Le Guern, 1983, p. 51 ; Macary, 1973, p. 152). S'il y a tout de même quelques additions, il s'agit en effet surtout de le purger de tout ce qui contredit la doctrine de Rome (Le Guern, 1983, p. 56). Le *Trévoux* connaîtra au total 9 éditions jusqu'à 1771 au fil desquelles il fait l'objet d'augmentations importantes et se démarque progressivement du *Furetière* (Macary, 1973, pp. 153–154). Grand rival de l'*EDdA*, ses auteurs se montrent de redoutables adversaires des philosophes des Lumières.

Au travers de ces quelques exemples glanés parmi une myriade d'œuvres et de projets, une régularité se dégage : si les dictionnaires semblent intéresser le TAL aussi bien comme outil que comme objet d'étude, en revanche les encyclopédies ne sont exploitées que comme objet d'étude.

2.2.3 L'Encyclopédie au centre des attentions

Parmi toutes les encyclopédies étudiées en HN, l'*EDdA* se singularise par le nombre de travaux qui y ont été consacrés dans de nombreux domaines. Il existe même une revue en ligne spécialisée sur le sujet, «Recherches sur Diderot et sur l'Encyclopédie» [22]. Des publications retracent l'histoire romanesque de sa publication entravée de censure royale (Moureau, 2001). D'autres s'intéressent aux personnages centraux dans le projet (Kafker & Loveland, 2016), en particulier à leurs motivations et à leurs stratégies en tant que contributeurs de l'*EDdA* (Lojkine, 2013) ou bien au contraire au reste de leur vie. Pruvost (2022) montre ainsi la place majeure qu'a occupée l'Académie française dans la

21. https://stanfordnlp.github.io/stanza/lemma.html#improving-the-lemmatizer-by-providing-key-value-dictionary
22. https://journals.openedition.org/rde/

vie de d'Alembert en dehors de sa participation à l'*EDdA* et l'influence déterminante qu'il a eue sur la cinquième édition du *Dictionnaire de l'Académie*.

Plusieurs études comparatives mettent en valeur l'importance de cette œuvre centrale du siècle des Lumières au travers de ses liens avec les autres grandes entreprises encyclopédiques du même siècle ou du suivant. Morin (1989) confronte ainsi l'*EDdA* à son prédécesseur et rival le *Dictionnaire de Trévoux* en explicitant le dialogue qui existe entre les deux œuvres malgré une ignorance feinte. Bien que les dernières éditions du *Trévoux* soient contemporaines de l'*EDdA*, cette rivalité se double en réalité d'une relation de filiation via la *Cyclopedia* anglaise de Chambers (Leca-Tsiomis, 1995) que Diderot et d'Alembert devaient initialement se contenter de traduire. Le caractère innovant de son titre aurait suscité chez son lectorat français des espoirs d'originalité infondés d'après Leca-Tsiomis (2023, p. 437), et ce malgré l'avertissement de Chambers lui-même dans sa préface. Le processus de traduction s'avère décevant pour les encyclopédistes en révélant de nombreux emprunts parfois passés sous silence (Leca-Tsiomis, 2023, p. 441), notamment au *Dictionnaire de Trévoux* [23] qui s'était lui-même servi copieusement chez Basnage. Certains travaux donnent même des éléments pour identifier les versions des différentes œuvres auxquelles ont pu avoir accès les contributeurs (Passeron, 2006). La lignée se prolonge après l'*EDdA* avec des œuvres comme la monumentale *Encyclopédie Méthodique* que Panckoucke envisage d'abord comme une simple amélioration de l'«ancienne Encyclopédie» (Groult, 2019a, p. 8) mais qui deviendra une collection de dictionnaires en plusieurs tomes pour chaque discipline (Groult, 2019b, p. 123) et dont la publication se poursuivra bien après sa mort jusqu'en 1832 grâce à ses descendants (La Cotardière, 2019). L'héritage de l'esprit encyclopédique passe ainsi au XIX^{ème} et donnera lieu à d'autres tentatives d'actualisation des connaissances de l'*EDdA* à travers des œuvres comme le *Grand Dictionnaire Universel* de Pierre Larrousse et *LGE* (Jacquet-Pfau, 2022), l'autre encyclopédie du corpus auquel s'intéresse cette thèse.

Naturellement, l'organisation de la connaissance dans l'*EDdA* fait également l'objet d'études. Blanchard & Olsen (2002) décrivent ainsi les renvois et la composition hiérarchique des domaines de connaissance qui, en sus de l'ordre alphabétique que les encyclopédies partagent avec les dictionnaires, structurent les articles de l'*EDdA*.

Ces articles constituent ainsi une source d'information précieuse sur l'évolution de nombreux domaines scientifiques. En s'intéressant à l'Histoire, Hardesty Doig (2006) souligne la part d'arbitraire résidant dans tout système d'organisation de la connaissance et montre que sa valeur ne se limite pas à la vue synthétique qu'il offre mais qu'il possède aussi une composante analytique des rapports entre sciences. L'appartenance d'un article à un domaine est marquée par un ou plusieurs désignants, dont les combinaisons, loin d'être de simples juxtapositions, obéissent à des logiques révélatrices (Cernuschi, 2006). Les désignants peuvent ainsi être des lieux d'innovation, faisant par exemple se rencontrer les termes «économie» et «politique» dans des articles qui, s'ils n'inventent pas le domaine *ex nihilo* semblent néanmoins fondateurs pour la discipline (Salvat, 2006). Reflets de l'état de la connaissance au XVIII^{ème} siècle, les désignants mettent en évidence des différences fondamentales dans le traitement de concepts aujourd'hui compris comme proches, et qui expliquent les limites des théories de l'époque, par exemple en Physique (Viard, 2006).

La Géographie est bien sûr concernée par ce type de travaux au même titre que les autres sciences. Laboulais (2003) relate comment elle peine d'abord à s'imposer dans les discours encyclopédiques [24] puis à trouver sa place parmi les autres domaines de connaissance, d'une affiliation théorique mais quelque peu artificielle aux mathématiques à une pratique bien plus ancrée dans les réalités humaines sous la plume d'auteurs comme Robert de Vaugondy ou de Pâris de Meyzieu (voir la section

23. ce qui vaudra aux encyclopédistes des accusations de plagiat de la part des mêmes jésuites qui avaient pourtant encensé la *Cyclopedia*

24. elle n'est vraiment introduite en masse dans une œuvre à vocation encyclopédique qu'à partir de l'édition de 1721 du *Dictionnaire de Trévoux*

3.1.2 p.58 sur ce point). L'importance des approches et des styles personnels des différents auteurs est également discutée par Laramée (2017) qui saisit le contraste édifiant entre la géographie de Diderot et celle du chevalier de Jaucourt avant d'expliciter la vision du continent américain qui ressort implicitement des pages de l'*EDdA*. Enfin, des études contrastives entre l'*EDdA* et d'autres œuvres, de manière similaire à celles mentionnées précédemment dans cette partie, existent aussi pour la Géographie. Vigier et al. (2022) montrent par exemple la continuité dans les profils discursifs des articles entre les dictionnaires de géographie spécialisés et les encyclopédies, tout en soulignant les différences d'approches entre le *Trévoux* et l'*EDdA*.

Après avoir présenté un format dans la section précédente et un objet d'étude essentiel pour les HN dans la présente, il reste avant de pouvoir clore cet état de l'art à s'intéresser à deux approches utilisées en HN. Pour bien les comprendre et tenter de les nommer, il est utile de commencer par un bref historique afin de montrer leurs origines communes.

2.3 Histoire et terminologie

Il est difficile de trouver un terme pour regrouper ces deux approches sans risquer d'en effacer une en donnant l'impression qu'elle n'est qu'un cas particulier de l'autre. Le terme «TAL» a été utilisé dans la section précédente (p.25) mais il vaut mieux bien comprendre ce qu'il recouvre avant de l'adopter. Pour cette raison, on se contente pour l'instant de désigner ces familles de méthodes par ce qu'elles ont en commun : une approche plus statistique que symbolique du langage. Pour bien cerner cette distinction, il est utile de revenir à la chronologie esquissée dans la section 2.2 et de la développer.

2.3.1 Rivalité historique entre les approches symbolique et statistique

Aux origines de cette lignée, il est d'abord question de «Traduction Automatique» (TA) et le domaine de recherche est surtout exploré pour ses applications potentielles, notamment militaires (Léon, 2002, para. 5). Ainsi ça n'est pas «TAL» mais seulement «TA» qu'il faut reconnaître dans «ATALA», le nom de l'Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée [25] créée en 1959 (Léon, 2002, para. 1).

Tout comme le terme plus général d'Intelligence Artificielle (IA) dont il constitue un champ, le TAL ne faisait initialement aucune hypothèse quant au moyen d'automatiser le traitement de la langue. Depuis ses origines, le domaine a d'après Church (2011, p. 2) connu une alternance de phases empiristes (axées sur des méthodes statistiques) et rationalistes (axées sur des méthodes symboliques). Après des succès rapides, l'accroissement des difficultés amenait un ralentissement des progrès pour l'approche dominante permettant à l'autre approche de reprendre l'initiative.

Au sortir de la guerre, l'IA est saisie d'un véritable engouement pour les réseaux de neurones artificiels. McCulloch & Pitts (1943) montrent comment ils peuvent être utilisés pour implémenter des fonctions logiques, puis Minsky (1952) met leurs idées en pratique en construisant une machine (physique) sur ces principes : le SNARC. Avec le perceptron de Rosenblatt (1958), un réseau de neurones monocouche, ces travaux établissent les fondations de l'Apprentissage Automatique (AA), dont les perspectives semblent alors illimitées.

Dans les années 1960 Chomsky, Pierce et même Minsky lui-même ont critiqué les gains techniques sans vision scientifique d'ensemble de ces méthodes empiristes et statistiques. Le rapport ALPAC (*Language and Machines*, 1966) – auquel Pierce a d'ailleurs contribué (Church, 2011, p. 16) – est à l'origine d'une réduction des financements pour ces approches, un «hiver» (Church, 2011, p. 4) au

25. https://www.atala.org/

profit des rationalistes. C'est à cette époque que fleurissent les travaux sur les grammaires formelles, décrites et catégorisées notamment par Chomsky (1956).

2.3.2 L'avènement des méthodes statistiques

Au cours des décennies suivantes, l'augmentation exponentielle des puissances de calcul disponibles selon la loi de Moore (1998) a permis de traiter de plus grandes sources de données ce qui a graduellement redonné l'avantage aux empiristes. Il s'est ainsi écoulé une trentaine d'années entre les premières solutions statistiques mentionnées ci-dessus et l'apparition d'architectures vraiment utilisables en pratique telles que les Réseaux de Neurones Récurrents (RNN), décrits par Werbos (1974) puis implémentés séparément par Rumelhart et al. (1986) et Le Cun (1986) ou les Réseaux de Neurones Convolutionnels (CNN), descendants du Neocognitron de Fukushima (1980).

Mais depuis 2010, le balancier semble cassé si l'on en croit Church (2011, p. 2), qui s'inquiète de ne pas voir les méthodes rationalistes faire leur retour à l'issue de la période d'une vingtaine d'années qui séparait les pics d'activité dans chacune des deux approches lors des oscillations précédentes. Loin de revenir dans les recherches pour combler les lacunes des méthodes statistiques, elles se retrouvent éclipsées jusque dans les contenus pédagogiques (Church, 2011, p. 19 et seq.).

Plus d'une décennie plus tard, ses inquiétudes semblent on ne peut plus fondées : l'approche empiriste occupe plus que jamais le devant de la scène et le terme «IA» est désormais utilisée de manière interchangeable avec «AA», voire parfois seulement pour désigner des modèles génératifs. L'Association pour l'Avancement de l'Intelligence Artificielle [26] semble ainsi ne plus s'occuper que d'AA, contrairement à ses débuts. Hors du monde académique l'IA est devenue un «catalyseur de croissance» (Verdegem, 2024) à l'origine d'un type d'économie, le «capitalisme d'IA» (Dyer-Witheford et al., 2019). Le mot, devenu un but en soi, est suffisant pour définir le cap de politiques publiques (Aghion & Bouverot, 2024).

Les travaux conduits dans cette thèse s'inscrivent dans l'approche empiriste. Ils utilisent des outils produits par ce courant aux deux époques où le «pendule» de Church oscillait de son côté, et qui correspondent à deux mouvements distincts dans la pratique de l'informatique linguistique. Après avoir identifié leurs origines communes, il reste à caractériser ce qui les sépare pour pouvoir convenir d'une dénomination à utiliser dans le reste de cette thèse.

2.3.3 Nommer des pratiques très différentes

La linguistique n'a pas attendu l'arrivée de réseaux de neurones artificiels utilisables pour que sa pratique bénéficie des apports des méthodes statistiques. Avant la période rationaliste autour des années 1970, les linguistes ont déjà vu leur boîte à outils s'étoffer au fil d'une mathématisation progressive de la discipline dans les deux décennies qui ont suivi la 2nde guerre mondiale (Léon, 2015, p. 6).

Au départ, ces nouvelles méthodes représentent seulement un moyen d'accélérer les recherches en décuplant la capacité des linguistes à aller trouver des exemples. En effet à partir des années 1960 la lecture attentive de textes pour repérer des occurrences de motifs, par exemples syntaxiques, est jugée trop inefficace (Léon, 2002, para. 63) et les équipes s'emploient à automatiser leur détection. Le projet de Wittgenstein de fonder le sens des mots sur leur emploi peut enfin être réalisé en étudiant leur «vrai» usage dans de grands échantillons de productions des locuteurs plutôt que de se satisfaire d'expériences de pensée (Willinsky, 2001, p. 197). L'école britannique, de Firth à Sinclair, s'inscrit naturellement tout à fait dans cette approche. En France aussi, mais de façon tout à fait différente, les mathématiques s'intéressent à la linguistique à cette époque et vont apporter énormément à la discipline à travers les contributions de chercheurs comme Benzécri.

26. https://aaai.org/conference/aaai/

Il ne s'agit donc pas d'un mouvement homogène mais bien d'une constellation de pratiques, ce qui complique déjà le choix d'un nom satisfaisant pour cette seule sous-partie du domaine et amène Léon (2015, p. 157) et Nazarenko et al. (1997) à parler de «linguistiques de corpus» au pluriel. C'est cet usage qui a été retenu pour ce manuscrit car il possède en effet une qualité précieuse : en faisant de «linguistique» le noyau du syntagme nominal, il rappelle qu'il s'agit d'une «automatisation des sciences du langages» pour reprendre le titre de Léon (2015). Il s'agit certes de l'utilisation de techniques nouvelles, mais par des linguistes et dans le cadre d'une pratique centrée sur la linguistique.

Pour revenir au terme «TAL» sur lequel s'ouvrait cette brève section, il est maintenant possible de lui donner un sens. Il semblerait parfait pour désigner l'ensemble du domaine : l'arrivée des machines dans les tâches linguistiques initiée dans les années 1960 et mentionnée ci-dessus consistait effectivement déjà à traiter le langage de manière automatique. Toutefois, le terme reste peu employé par les linguistes et semble en réalité bien plus populaire sur l'autre versant de la discipline : dans une pratique de l'informatique dont la linguistique représente un objet d'application. Il y est revendiqué autant par des (rares, désormais) héritiers de Chomsky établissant de nouveaux modèles formels de la langue que par des empiristes développant de nouveaux outils statistiques.

Au vu des succès écrasants des approches empiristes, le sens de «TAL» tend à se réduire à «application de méthodes d'AA à des problèmes linguistiques». C'est pourtant le sens plus général d'«informatique appliquée à la linguistique» sans hypothèse sur le type de méthodes utilisée qui a été retenu dans ces pages, non pas par refus d'une évolution du langage mais pour ne pas contribuer à occulter une réalité qui a été productive par le passé et pourrait l'être à nouveau. S'il fallait un terme pour distinguer la pratique du TAL restreinte à l'AA, il serait alors tentant d'exploiter la remarque faite à la section 1.3 (p.16) en employant «NLP» (*Natural Language Processing*) tant l'emploi de l'acronyme anglais équivalent est massif même chez des francophones quand il s'agit d'AA.

Ainsi, à l'issue des considérations historiques et réflexions précédentes la place relative des approches couvertes dans les deux sections suivantes de cet état de l'art devient claire. La section 2.4 s'intéresse au versant informatique de la discipline, majoritairement à des méthodes empiristes mais sans exclure quelques techniques rationalistes. La section 2.5 sera au contraire centrée sur les pratiques linguistiques enrichies d'outils informatiques, héritières des travaux initiés dans les années 1960 par les écoles britannique et française.

2.4 Traitement Automatique de la Langue

La section qui s'ouvre présente les différents outils employés pour une part lors de la préparation du corpus (voir chapitre 3) mais plus encore dans les travaux de classification (présentés au chapitre 4). Comme le laissent présager les remarques historiques de la section précédente, l'AA y occupe une part importante.

Les méthodes basées sur des réseaux de neurones travaillent sur des entrées sous forme numérale. Pour pouvoir travailler sur des textes, ils doivent nécessairement à un moment de leur traitement être représentés par des nombres ou plutôt des tableaux de nombres : des vecteurs.

2.4.1 Vectorisation

Pour vectoriser un document, une première approche consiste à le considérer comme un «sac de mots» (Salton & McGill, 1986). L'ensemble des mots dans un corpus (tokens ou lemmes selon les implémentations) constitue le vocabulaire utilisé comme base d'un espace vectoriel de très grande dimension dans lequel chaque document est représenté. Pour chaque vecteur de cette base (correspondant donc à un mot unique du corpus), la composante associée dans le vecteur qui représente un document est un entier positif ou nul égal au nombre d'occurrences du mot (sa fréquence) dans

ce document. Le terme anglophone consacré est «*Bag of Words*», et bien que ce terme soit le seul pour lequel l'équivalent français s'emploie, l'acronyme *BoW* sera également utilisé dans ce qui suit par souci d'homogénéité avec les autres méthodes. Cette approche produit des représentations vectorielles très grandes (beaucoup de nombres) et avec peu de valeurs non nulles (la plupart valent 0) car chaque document n'utilise qu'une fraction de l'ensemble du vocabulaire défini par le corpus entier. On parle de représentation *creuse* (*sparse* en anglais).

La deuxième approche, TF-IDF (*Term Frequency - Inverse Document Frequency*, «fréquence du terme - fréquence inverse de document») est un type de représentation en *BoW* qui au lieu de simplement pondérer les mots par leur fréquence dans le document considéré tempère ce nombre en le divisant par la proportion de documents du corpus qui contiennent ce mot. De la même manière que les *BoW* «purs», les vecteurs produits par cette méthode sont creux, ce qui tend à dégrader les performances des algorithmes d'AA en augmentant leurs consommation de mémoire et leurs temps d'exécution.

Par contraste avec les deux approches précédentes, les plongements de mots produisent des vecteurs de plus petite dimension et denses en coefficients non nuls. Ils constituent une famille de méthodes fondamentalement différentes des deux précédentes du fait qu'elles capturent le contexte des mots dans leurs représentations vectorielles. Il y a deux approches pour entraîner des plongements de mots : CBOW (*Continuous Bag of Words*, «sacs de mots continus») et *skip-gram* («fenêtres à trou»). La première consiste à prédire un mot en fonction de son contexte. À l'inverse, une architecture *skip-gram* apprend à prédire le contexte à partir d'un mot. Initialement pensée au niveau des mots individuels, *Word2Vec* (Mikolov et al., 2013) permet déjà de représenter un texte en combinant les vecteurs des mots qu'il contient (par exemple en faisant leur moyenne). Peu adapté aux textes longs, on lui préfère souvent la méthode voisine *Doc2Vec* (Le & Mikolov, 2014) sur ce type d'entrées. Ces méthodes de plongements nécessitent un entraînement sur le corpus sur lequel elles vont être utilisées car elles ne peuvent pas gérer de mots spécifiques à ce corpus et qu'elles n'auraient jamais rencontrés. D'autres telles que *FastText* (Bojanowski et al., 2017) sont disponibles préentraînées car elles considèrent des fenêtres de quelques caractères plutôt que des mots entiers, ce qui leur permet de calculer un vecteur même pour des termes absents de leur vocabulaire d'entraînement. On parle pour ces méthodes de plongement «statique» parce qu'elles produisent un vecteur unique pour représenter un mot, en combinant tous les différents contextes dans lesquels il peut apparaître. Plus récemment, *BERT* (Devlin et al., 2019) – *Bidirectional Encoder Representations from Transformers* («Représentations d'Encodeur Bi-directionnels à partir de Transformeurs», voir p.32) – utilise un plongement contextuel pour la phase de vectorisation qu'il intègre, c'est-à-dire où la représentation d'un mot dépend du contexte dans lequel il apparaît dans une phrase. *BERT* utilise des réseaux de neurones de type *transformer* et le concept de masque pour prédire les mots qui complètent une amorce de phrase donnée.

2.4.2 Classification

La classification de documents est un problème général en IA qui consiste à regrouper ensemble des objets jugés plus proches ou ressemblants pour une certaine sémantique ou une certaine métrique par opposition à ceux qui aboutissent dans les autres groupes. Il existe dès le début des années 1970 une littérature fournie portant sur les stratégies de classification et les mesures de similarités utilisées que Cormack (1971) entreprend d'organiser.

Les problèmes de classification couvrent un grand nombre d'applications pratiques variées. On trouve des études portant sur la classification de sons (Nogueira et al., 2022), d'images (Adriyanto et al., 2022) – même hors du spectre visible (Yu et al., 2020) – ou bien d'événements sur un système informatique en vue de détecter des attaques (Gavari Bami et al., 2022).

Classer des textes

En TAL, classer un texte peut signifier lui attribuer par exemple un sujet («infrastructure» ou «politique étrangère»), un type de contenu («dépêche» ou «publicité») voire une sensibilité politique («gauche» ou «droite») dans le cas d'une étude de Peterson & Spirling (2018).

Appliquées aux encyclopédies, les méthodes de classification permettent de prédire les domaines de connaissance des articles (voir à ce sujet la section 3.1.2 p.47). Dans le cadre des travaux de l'ARTFL sur l'*EDdA*, Horton et al. (2009) ont ainsi testé la classification Bayesienne dite «naïve» (Naive Bayes) pour prolonger la classification des auteurs de l'œuvre sur les entrées laissées sans domaine. L'Encyclopedia Britannica a de même été étudiée par le Nineteenth-Century Knowledge Project [27] qui a utilisé des méthodes d'AA mais aussi des approches basées sur des règles pour indexer 400 000 articles à travers 4 éditions de l'œuvre (Grabus et al., 2019).

Classification supervisée

Dans tous les cas évoqués ci-dessus, l'ensemble de classes à attribuer aux documents est défini en amont de l'étude. Il n'est pas pour autant indépendant du corpus mais correspond au contraire à un axe d'analyse que l'on suppose pertinent : observer une sensibilité politique peut avoir un intérêt pour classer des discours parlementaires ou des professions de foi électorales, mais serait à priori beaucoup moins adapté à des prévisions météorologiques ou des recettes de cuisine.

À partir de l'ensemble de classes, un échantillon des documents doit être annoté manuellement pour servir de référence pour l'entraînement du modèle de classification : il s'agit de classification supervisée. Plus précisément, on distingue un jeu d'entraînement qui aide directement à ajuster les coefficients du modèle et un jeu de test qui permet d'en évaluer les performances une fois l'entraînement réalisé et avant d'appliquer le modèle à l'ensemble des documents à classer.

Les méthodes les plus simples utilisées pour la classification apparaissent à la suite du renouveau de l'AA dans les années 1990 : *Naive Bayes* (Domingos & Pazzani, 1997), les *Support Vector Machines* (*SVM*) de Christianini & Shawe-Taylor (2000) puis les *Random Forests* (Breiman, 2001) et la *Logistic Regression* (Kleinbaum et al., 2002). D'autres telles que les RNN et les CNN mentionnés précédemment (voir p.29) reposent sur de nombreuses couches de neurones qui abstraient progressivement le signal d'entrée et amènent à parler d'Apprentissage Profond (AP). Les *CNN*s sont basés sur deux opérations : la convolution et un sous-échantillonnage des entrées gardant les valeurs maximales (*max-pooling*). La convolution extrait des éléments caractéristiques des données et le max-pooling compresse le résultat, permettant à ces architectures d'analyser des images avec succès. Les RNN, davantage adaptés à l'analyse de données séquentielles, trouvent un prolongement dans les *LSTM* (*Long Short-Term Memory*, «mémoire à court-terme longue») qui permettent de mieux capturer des dépendances éloignées dans les séquences d'entrées (Hochreiter & Schmidhuber, 1997), ce que les RNN ne parviennent pas à faire en pratique. Une autre technique d'AP pour garder accès aux premiers éléments d'une séquence d'entrée, nommée l'«attention», est au centre de l'architecture des transformeurs, un autre modèle d'AP qui définissait l'état de l'art au début de cette thèse. Un modèle tel que *BERT*, basé sur des transformeurs, permet à la fois la vectorisation et la classification de ses entrées. Son très grand nombre de paramètres requiert des ressources de calcul considérables pour entraîner un tel modèle à partir de rien. Puisque de tels moyens ne sont à l'heure actuelle disponibles que sur des infrastructures dédiées au calcul dont l'accès est coûteux et restreint, le modèle est distribué préentraîné sur des tâches générales et ne doit subir qu'un réentraînement superficiel sur quelques époques pour l'adapter à une tâche en particulier – on parle de *fine-tuning* en anglais et s'il est tentant de proposer la traduction «spécialisation» par analogie avec la pédagogie et l'apprentissage humain, là encore le terme anglais semble s'être imposé en français.

27. https://tu-plogan.github.io/

Topic-modeling

Au contraire des méthodes de classification précédentes, d'autres approches tâchent de révéler la structure propre à l'ensemble de documents qu'il s'agit de classer sans à priori en les groupant par «clusters». Certaines sont basées exclusivement sur des mesures de distance dans un espace métrique où sont représentés les objets (Lance & Williams, 1967), d'autres comme l'Allocation Latente de Dirichlet (LDA) créent des clusters de mots ce qui permet de mettre automatiquement en évidence les différents thèmes présents dans les documents du corpus. Ces méthodes permettent d'approcher les données avec moins de préconçus mais demandent davantage de travail dans l'interprétation des résultats.

La modélisation de thèmes par LDA connaît une certaine popularité pour déterminer les contenus saillants dans de vaste ensemble de données textuelles. Ce succès s'explique sans doute au moins en partie par le fait qu'elle ne demande pas de posséder au préalable des métadonnées ni de créer des annotations pour l'entraînement car c'est une méthode non supervisée. Dans son analyse de discours prononcés au parlement britannique, Guldi (2019) emploie le Topic Modeling pour mener une «recherche critique» des «tensions et des moments charnière» dans les débats politiques du Royaume-Uni. Barron et al. (2018) utilisent le Topic Modeling comme un point d'entrée à partir duquel mesurer la «nouveauté» et le caractère plus ou moins «éphémère» des discours prononcés lors des premières années de la Révolution Française. L'intérêt de cette approche est qu'elle ne requiert pas de métadonnées sur chaque discours outre celles disponibles à la constitution du corpus d'étude et utilisées pour organiser les textes (dans le cas précédent, le nom du député à qui il est attribué et la date à laquelle il a été prononcé).

Après les travaux de classification sur Naive Bayes évoqués plus haut, une autre étude menée à l'ARTFL a consisté à utiliser la LDA pour constituer automatiquement des groupes d'articles qui ont ensuite été analysés en les comparant aux classes originales des encyclopédistes (Roe et al., 2016). À l'issue de ces recherches, ils concluent que les thèmes identifiés par la LDA peuvent être compris en termes de discours sous-tendant l'*EDdA* de part en part et qui ne sont pas en correspondance simple avec les classes originales.

En utilisant la LDA ainsi que d'autres modèles d'AA, Underwood a examiné l'histoire et l'instabilité de genres littéraires (Underwood, 2016, 2018, 2020). Il trouve les méthodes computationnelles utiles en ce qu'elles peuvent «détecter et comparer des ressemblances de familles floues qu'il serait difficile de définir avec des mots sans être simpliste» (Underwood, 2016, p. 6). De telles approches de la classification de texte, quantitatives et prédictives offrent aux recherches en HN la possibilité de comprendre les résultats dans un cadre interprétatif nouveau.

Reconnaissance Optique de Caractères

Enfin, avant de clore complètement cet aperçu des travaux en Classification Automatique, il est intéressant de mentionner un champ de recherche qui utilise des techniques de classification sans s'y réduire. Si la localisation des caractères sur la page et la compréhension des liens qu'ils entretiennent représentent certes un type de problèmes entièrement différent, l'identification des zones trouvées constitue quant à elle un problème de classification. Les moyens classiques semblent inefficaces pour produire des résultats généralisables sur des données jamais rencontrées par l'algorithme ce qui conduit Sayre (1973, p. 216 et seq) à explorer des méthodes statistiques pour reconnaître les caractères.

En implémentant les CNN, Le Cun et al. (1989) parviennent à créer des modèles capables de lire les codes postaux écrits à la main sur des enveloppes. De nombreux systèmes récents sont basés sur les *LSTM* (Wick et al., 2018, p. 79), architecture utile en classification, ce qui souligne encore la parenté entre les deux tâches. On distingue généralement cette tâche encore plus difficile de reconnaissance

de caractères tracés à la main – donc moins réguliers – sous le nom de *Handwritten Text Recognition* (HTR) par rapport au problème plus général d'OCR. Pour favoriser l'évaluation de système de HTR, Chagué et al. (2021) proposent la diffusion de jeux de données pouvant servir de vérité terrain. Aujourd'hui, tous les systèmes d'OCR sont basés sur des méthodes d'AA.

Les travaux de classification effectués dans cette thèse n'ont pas porté sur cette tâche mais elle représente un prérequis absolu à leur existence, puisqu'on lui doit l'existence d'OCR capables de fournir automatiquement des versions numériques des textes d'encyclopédies. L'existence de systèmes d'OCR rend donc possible cette étude ainsi que d'autres semblables qui travaillent sur de grandes masses de données imprimées comme le projet NewsEye [28] qui s'intéresse à la presse imprimée des siècles passés (Doucet et al., 2020).

2.4.3 Étiquetage morphosyntaxique et syntaxique

Une autre tâche très commune en TAL concerne l'annotation automatique des textes en morphosyntaxe et en syntaxe : bien qu'avec une approche souvent différente de celle de la linguistique classique, le TAL s'intéresse aussi aux mots. D'abord il s'agit de les délimiter lors d'une phase de *segmentation* en unités lexicales (on trouve, particulièrement en TAL, l'emploi du mot «*token*» pour rendre compte du caractère artificiel et conventionnel de ces unités). Ensuite, il faut identifier à quelle classe grammaticale ils appartiennent lors de l'*annotation morphosyntaxique*. On parle aussi pour désigner les classes grammaticales de «partie de discours» (en anglais *part of speech* – POS). Enfin l'*annotation syntaxique* établit leurs relations au sein des phrases. En rendant faisables en pratique de telles annotations sur de larges volumes de texte, les méthodes automatiques constituent des outils précieux pour les linguistiques de corpus (voir section 2.5). De même que pour l'OCR, les présents travaux ne contiennent pas de contribution sur ces tâches mais elles sont indispensables pour comprendre les phases de préparation et d'analyse du corpus.

La question de la seule segmentation du texte constitue déjà un problème «délicat» (Diwersy et al., 2017, p. 29) qui ne relève en rien de l'évidence. Pour commencer, il n'existe pas de caractère qui délimite les mots sans aucune ambiguïté : l'espace qui apparaît comme un délimiteur évident fait pourtant partie de mots composés comme «compte rendu» (Lebart et al., 2019, p. 44), les caractères de ponctuation comme l'apostrophe peuvent parfois séparer deux mots («l'ami») et parfois faire partie d'un seul mot («aujourd'hui») de même que le point marquant une fin de phrase ou bien un mot abrévié (Habert, 2005, p. 13).

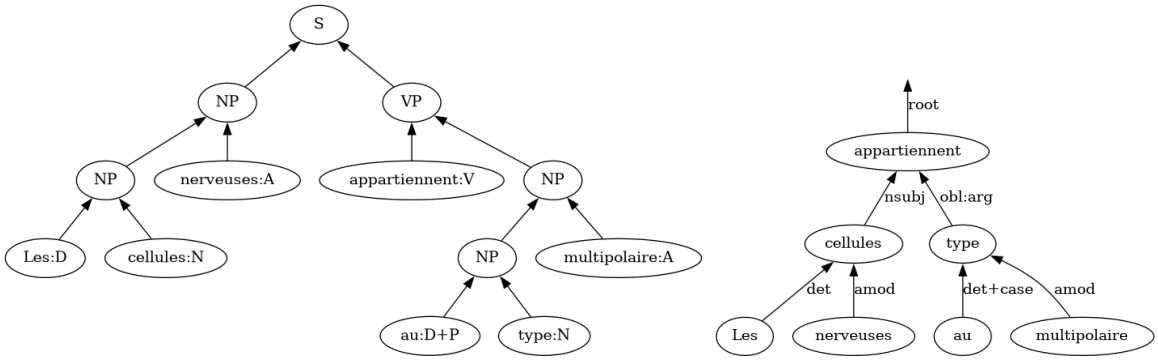
La tâche d'annotation morphosyntaxique qui vient ensuite peut s'apparenter au problème de classification décrit dans la sous-section précédente en ce qu'elle associe à chaque mot en entrée une classe possible (sa POS) parmi celles prédéfinies. Il existe plusieurs jeux d'étiquettes morphosyntaxiques de référence comme celui du Penn Treebank (Marcus et al., 1993) mais le jeu à utiliser doit être déterminé en fonction des besoins de l'étude que l'on souhaite mener. Ainsi Vigier (2017, p. 100) recourt à une catégorie «G» qui regroupe des participes présents, adjectifs verbaux et gérondifs pour tenir compte du fait que ces objets ne sont pas encore bien différenciés dans les états de langues les plus anciens de son corpus d'étude.

Toutefois, à la différence des problèmes de classification habituels, l'annotation en POS est fortement compliquée par une dépendance contextuelle aiguë. Des mots différents, auxquels il faut donc attribuer des POS différentes, peuvent revêtir la même forme et nécessiter le contexte des autres mots de la phrase pour être distingués. Ce peut être le cas par exemple pour des raisons d'homonymie : «été» peut être le nom d'une saison ou bien une forme du verbe «être». Puisqu'il s'agit de «reconnaître» les mots, cette phase d'analyse est généralement l'occasion d'associer aussi à chaque mot une forme normalisée, son «lemme». L'existence d'amalgames tels que «des» dans «le temps des cerises» ou

28. https://www.newseye.eu/

«aux» dans «une tarte aux prunes» (qui rassemblent en un seul mot en réalité deux mots – «de» et «les» pour le premier, «à» et «les» pour le second) vient complexifier les choses en montrant que la segmentation ne peut coïncider tout à fait avec le fonctionnement de la grammaire. Une stratégie pour annoter les amalgames peut consister à leur associer une combinaison des lemmes et des POS des mots qui les constituent comme le fait Vigier (2017, p. 101). Il existe de nombreux algorithmes pour calculer la POS des mots dont les implémentations ont donné lieu à des outils comme le Brill Tagger (Brill, 1992) ou TreeTagger (Schmid, 1995). Ces outils produisent généralement des formats tabulaires comme TSV (voir section 2.1 p.18) qui représentent un token par ligne, stockant dans différentes colonnes la forme précise rencontrée dans le texte, le lemme et la POS correspondante.

La dernière phase, celle d'analyse en syntaxe, révèle la structure des phrases sous une forme arborescente. On distingue généralement deux approches : l'analyse en constituants qui descend des théories de Chomsky (1965) et celle en dépendance qui est davantage fondée sur la vision de Tesnière & Fourquet (1959). La première s'inscrit dans la lignée des grammaires non contextuelles qui procèdent par réécritures successives de règles. Dans un tel système, une partie des symboles dits «terminaux» représente les mots du vocabulaire tandis que les autres symboles modélisent les différentes structures abstraites, certaines récursives, disponibles pour construire des phrases. Ces éléments non terminaux comprennent entre autres les syntagmes nominaux et verbaux ou les propositions relatives qui peuvent elles-mêmes renfermer des éléments valides au niveau d'une phrase complète. Les nœuds des arbres obtenus avec cette approche contiennent les différentes parties qu'il est possible d'isoler dans une phrase : la phrase dans sa totalité, généralement notée S, se trouve à la racine de l'arbre, qui peut avoir par exemple deux nœuds enfants NP et VP pour représenter un syntagme nominal sujet d'un syntagme verbal comme c'est le cas à la figure 2.5a qui montre une analyse en constituants de la phrase «Les cellules nerveuses appartiennent au type multipolaire» extraite de l'article SYMPATHIQUE (La Grande Encyclopedie, T30, p.759). Dans l'analyse en dépendance au contraire, les mots sont directement en relation les uns avec les autres et ce sont eux qui peuplent les nœuds de l'arbre syntaxique. Les arêtes entre nœuds expriment les fonctions syntaxiques : un adjectif peut se trouver sous un nom commun, son arête portant la fonction «amod» (qui signifie que l'adjectif modifie le sens du nom) comme le montre la figure 2.5b qui présente une analyse en dépendances de la même phrase.



(a) Analyse en constituants (b) Analyse en dépendances

Figure 2.5 – Deux analyses syntaxiques d'une phrase de l'article SYMPATHIQUE, *LGE*, T30, p.759

Ces théories apparues lors de l'âge d'or rationaliste (voir la section 2.3) s'accordent naturellement très bien avec des méthodes symboliques. Elles ont donné lieu à de nombreux formalismes grammaticaux – grammaires d'arbres adjoints (Joshi & Schabes, 1997), grammaires catégorielles abstraites (Groote, 2001) – ainsi qu'à des algorithmes tels que Cocke-Kasami-Younger (Moll et al., 1988, p. 64) pour les grammaires non contextuelles ou celui de Nivre & Scholz (2004) pour de l'analyse en dépendances. Des outils comme Talismane (Urieli & Tanguy, 2013) intègrent toutes les couches d'analyse,

de la segmentation à l'analyse syntaxique conduite par un algorithme hybride qui combine apprentissage automatique et règles expertes.

Cependant, comme pour la classification, l'état de l'art de l'analyse syntaxique est désormais défini par des méthodes d'AA. Cela explique l'importance dans le domaine des *treebanks*, ces «arboretums» syntaxiques rassemblant de grands nombres d'exemples de phrases annotées par des spécialistes et pouvant servir de vérité terrain pour entraîner et évaluer les modèles de langues sur cette tâche. Il existe des treebanks pour de nombreuses langues, souvent même plusieurs pour les langues les plus étudiées comme le French Treebank (Abeillé et al., 2003), Sequoia (Candito & Seddah, 2012) ou encore GSD (Guillaume et al., 2019) amélioré et converti à partir du French Treebank en utilisant Grew (Bonfante et al., 2018) un outil de réécriture de graphes. Ces grandes collections de données que sont les treebanks sont rendues interopérables grâce à un formalisme commun pour l'annotation syntaxique, les «dépendances universelles» (*Universal Dependencies*[29] ou UD dans le reste du manuscrit) introduites par Nivre et al. (2017). Ce formalisme qui s'inscrit comme son nom l'indique dans l'analyse en dépendances de Tesnière fournit des jeux d'étiquettes pour les POS et pour les relations syntaxiques créés dans le but de pouvoir couvrir toutes les langues naturelles. Les annotations en UD sont généralement représentées en utilisant le format CoNLL-U[30] (Buchholz & Marsi, 2006) qui s'apparente à un format TSV enrichi pour pouvoir représenter les séparations entre phrases, des commentaires ainsi que les «sous-tokens» nécessaires pour rendre compte des amalgames. Des outils tels que Stanza (Qi et al., 2020) – retenu pour annoter en syntaxe le corpus de cette thèse – ou HOPS (Grobol & Crabbé, 2021), tous deux basés sur de l'AA, permettent d'analyser un texte en UD et produisent en sortie des fichiers au format CoNLL-U. Le modèle utilisé par HOPS prédit l'existence de liens dans un graphe dont les nœuds représentent les mots de la phrase à annoter. Cette stratégie, mise en regard du fait qu'il est également possible d'utiliser des règles de réécriture de graphe pour annoter en syntaxe (Guillaume & Perrier, 2015 ; Bonfante et al., 2018, chap. 6) souligne la proximité thématique qui existe entre la théorie des graphes et l'analyse syntaxique (bien que les arbres syntaxiques, qui n'admettent pas de boucles, ne soient que des cas particuliers plus simples de graphes).

2.5 Linguistiques de corpus

La branche la plus ancienne des méthodes statistiques présentées à la section 2.3 (voir p.28) est née bien avant que des réseaux de neurones ne soient utilisables en pratique. On pourrait donc la croire éteinte : cette dernière section entend montrer qu'il n'en est rien. Puisant aux sources des mathématiques et de la statistique, les linguistes n'ont pas attendu ces technologies pour mobiliser des approches très variées sur leurs problématiques de recherche.

2.5.1 Du texte au corpus

Un dénominateur commun à toutes ces approches existe dans la notion de corpus, qui peut en première intention se comprendre comme une «collection» de textes. Avant de décrire dans le reste de cette partie les différentes manières de réunir un tel ensemble, il est utile de commencer par s'interroger sur la notion de «texte» puisque c'est par rapport à elle que sera définie celle de corpus.

Le terme est assez fortement polysémique puisqu'il peut être utilisé tant pour désigner une unité sémantique d'usage du langage (Halliday & Hasan, 1976, pp. 1–2) qu'une simple «matière» travaillée par les méthodes automatiques évoquées à la section 2.4 (p.30) voire pour signifier dans la langue courante une production écrite (par opposition à orale). Rastier (1996, p. 19) définit la notion comme «une suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque». Comme pour de nombreux objets linguistiques, l'évidence apparente

29. https://universaldependencies.org/
30. https://universaldependencies.org/format.html

dont la notion peut sembler relever s'estompe en essayant de mieux la cerner. Elle engendre dans la communauté linguistique des débats dont la définition technique précédente représente une prise de position qui refuse de «fixer dans la phrase le niveau ultime» (Benveniste, 1966, p. 131) et nourrit la réflexion de domaines comme la sémiotique (Rastier, 2001). Dans le cadre de cette section, la notion de texte envisagée se rapproche de la définition de Halliday & Hasan sans en conserver toute la porté conceptuelle puisque la cohérence sémantique des textes établit surtout la base de la segmentation des œuvres utilisée pour l'analyse. Le dernier sens évoqué ci-dessus – le texte comme «matière» – joue un rôle important dans le reste de cette thèse et il faut d'abord en dire quelques mots.

Les deux emplois sont presque contraires ; car le texte comme unité se définit implicitement par ses contours, par contraste avec les autres textes qui en sont distincts, alors qu'en considérant le texte comme une matière à travailler, la notion de frontière ne s'y applique plus. D'ailleurs, dans le premier cas «texte» est dénombrable alors que dans le second, traité véritablement comme un continu, il se retrouve manipulé comme une quantité indénombrable : Rastier & Pincemin (1999, p. 85) marquent cette opposition entre *des* et *du* texte(s). L'école londonienne perçoit l'importance cruciale de la taille des corpus pour produire des descriptions exhaustives des lexiques (Krishnamurthy, 2008, pp. 232–233) mais l'attention que porte par exemple Sinclair (2004) à l'authenticité et à l'intégrité *des* textes l'empêche de considérer le texte comme une simple substance. En revanche cet emploi prévaut dans les méthodes d'AA traitées à la fin de la section 2.4 où «texte» ne désigne souvent plus que la nature des données en entrée – par opposition à «image» ou «audio» – souvent même qualifiées de texte «brut» (voir la section 3.2.3 à ce sujet).

Une distinction semblable se retrouve dans la manière dont le texte ou les textes peuvent être assemblés pour former un corpus d'étude. Léon (2015, p. 159) oppose ainsi les mouvements *corpus-based* et *corpus-driven* déjà présents au sein de l'école britannique. Le premier correspond à l'approche de Quirk ou de Leech par exemple (Léon, 2015, p. 163) qui assemblent des textes choisis dans un but précis : l'étude est *basée sur un corpus*. À l'inverse, l'approche *corpus-driven* de Firth, Sinclair et l'école londonienne avec eux requiert des textes intégraux par souci d'objectivité et d'authenticité des productions étudiées : c'est ce qui émerge du corpus qui oriente l'analyse.

Au XXI^{ème} siècle, la notion de corpus demeure plurielle, s'appliquant à des objets, des usages, des domaines de la linguistique et même des méthodes de traitement très différentes (Cori et al., 2008, p. 6). Parmi ces méthodes, la textométrie accorde une importance décisive au corpus, qui doit être composé avec soin et équipé de métadonnées pertinentes en vue de la conduite d'analyses précises (Pincemin, 2022, p. 9). Pincemin (2012, pp. 13–14) affirme qu'il ne suffit pas de regrouper des données pour obtenir un corpus, qui doit selon elle se fonder entre autres sur des critères de cohérence, de représentativité, de complétude et d'exploitabilité. Elle singularise la textométrie par rapport à la linguistique de corpus pour laquelle le paramètre majeur serait le volume de données (Pincemin, 2012, p. 15). La profusion de textes en libre accès sur internet permet en effet l'apparition d'approches assez opposées comme le *Web as Corpus* (WAC [31]) où toute ressource est bonne à prendre pourvu qu'elle augmente la quantité de données d'une étude, partant du principe que cela la rendra plus robuste – argument rapporté mais critiqué par Pincemin (2012, *ibid*).

Dans tous les cas, le corpus joue un rôle assez déterminant dans la discipline pour justifier l'appellation *Corpus Linguistics*, «linguistique de corpus», apparue en 1984 (Léon, 2015, p. 159). Elle est parfois opposée à *Computational linguistics*, «linguistique computationnelle», préférée par les américains (Léon, 2015, p. 130) non seulement pour les approches statistiques, en particulier dans les applications de l'AA à la linguistique évoquées à la section 2.4, mais également pour les approches symboliques. L'opposition semble dépasser le cadre de la terminologie, Wilks (2010, p. 408) appelant à dépasser des «malentendus», ce qui n'empêche pas Leech d'utiliser l'hybride

31. l'acronyme n'est plus à la mode mais la pratique est en filiation directe avec celle utilisée pour entraîner les *Large Language Models* une décennie après la remarque de Pincemin.

Computer Corpus Linguistics (Léon, 2008, p. 26) – «linguistique de corpus par ordinateur». Léon (2015, p. 13) parle quant à elle d'une véritable «guerre des dénominations» (Léon, 2015, p. 130) : en plus des traductions de ces deux termes on trouve également en français linguistique «quantitative», «appliquée» et «outillée»...

2.5.2 Les outils de la linguistique

Ces termes recouvrent de nombreuses pratiques communes et semblent refléter autant des véritables différences de positions scientifiques que des traditions académiques. Elles partagent une variété de méthodologies dont la plupart connaissent des implémentations dans les outils courants du domaine tels que TXM (Heiden et al., 2010), IRaMuTeQ [32] ou encore Hyperbase (É. Brunet & Luong, 1989). Un état des lieux très complet des différents types de traitements et d'analyses utilisés est déjà dressé par Pincemin (2022, p. 2.2, pp.2-7) mais cette sous-section s'applique à les replacer dans leur contexte historique et disciplinaire.

La notion de collocation occupe une place centrale depuis les débuts britanniques de la discipline, chez Firth par exemple (Léon, 2015, p. 161) jusqu'à des travaux plus récents (Fellbaum, 2007). Le sens du mot a changé progressivement mais il reste étroitement lié à celui de cooccurrence, c'est-à-dire l'apparition conjointe de deux termes dans un même empan textuel. Motivé par la recherche de termes fréquemment associés en vue de caractériser le sens des mots dans le cadre de la production de dictionnaires (voir section 2.2.1), le concept de collocation en est progressivement venu à désigner des cooccurrences particulièrement fréquentes jusqu'à créer une «attente mutuelle» chez les locuteurs (Léon, 2008, p. 16). Caractériser cette attente demande une mesure statistique précise de leur surreprésentation (Lafon, 1981).

En pratique, l'implémentation du calcul des cooccurrences le rapproche de celui des spécificités comme le fait remarquer Pincemin (2022, p. 4). Avant d'expliciter la notion de spécificité, il est utile de remarquer que la plupart des outils utilisés dans les linguistiques de corpus reposent sur des décomptes de différents objets avec une approche statistique. En creux, ces décomptes supposent un découpage : il s'agit de recenser les occurrences d'un phénomène dans une partie du corpus par rapport aux autres pour dégager des critères quantitatifs objectifs qui caractérisent cette partie. Ce concept de «partition» est fondamental dans l'approche contrastive et dans l'ensemble de la discipline.

Guiraud le premier, à la tête des stylisticiens (Léon, 2015, p. 129) remarque que «la linguistique est la science statistique type» (E. Brunet, 2014). Bientôt rejoint par Muller, ils forment dans les années 1960 un courant que Beaudouin (2016, p. 21) nomme la statistique lexicale et dont les travaux sont initialement basés sur des décomptes de fréquence. Il s'agit seulement au départ d'établir des concordanciers répertoriant les occurrences des mots dans leur contexte, de dénombrer ces occurrences, leur distribution et d'établir un vocabulaire caractéristique des textes (E. Brunet, 2014, p. 21). On parle aussi pour cette branche de lexicométrie. L'ordinateur n'est pas encore devenu indispensable dans cette pratique de la linguistique car les calculs restent très simples, accessibles à une calculette (E. Brunet, 2014, p. 17). Le premier grand débat qui agite le champ concerne l'utilisation de formes lemmatisées ou non (Beaudouin, 2016, p. 30 ; Pincemin, 2022, p. 14). Muller prendra parti en faveur de la lemmatisation (E. Brunet, 2009, p. 1).

Cependant, considérer les seules fréquences souffre de limites intrinsèques comme le fait de ne pas pouvoir saisir les absences (Pincemin, 2022, p. 3) : impossible en effet de compter des «non occurrences» de phénomènes qui ne se réalisent pas dans une partie d'un corpus par rapport à la normale définie par les autres textes avec lesquelles elle serait mise en contraste. En rapportant les fréquences observées à la taille des différentes parties, Lafon (1980) définit la spécificité d'une

32. http://www.iramuteq.org/

forme comme une mesure de la probabilité que la distribution observée résulte d'une répartition purement aléatoire au sein du corpus. En ce sens, la spécificité quantifie la surprise que constitue la surreprésentation (ou la sous-représentation) d'un motif dans une partie du corpus. Avec la loi de Zipf, ce point constitue un deuxième lien entre la lexicométrie et les travaux menés en théorie de l'information dès les années 1950 notamment par Shannon ou Markov (Léon, 2015, chap. 4 et en particulier p.56).

Différentes lois ont été utilisées pour modéliser la distribution des mots en vue de calculer des spécificités comme la loi du Khi^2. La distribution des mots dans un langage étant fortement inhomogène (la fréquence d'un terme étant inversement proportionnelle à son rang selon la loi de Zipf), il a également été proposé pour mieux modéliser leur répartition d'utiliser une loi normale pour les formes les plus communes combinée à une loi de Poisson pour les plus rares. C'est finalement pour une loi hypergéométrique qu'opte Lafon (1980, p. 128) et c'est sur ce choix qu'est basée l'implémentation dans le logiciel TXM (Heiden et al., 2010). Dans ce logiciel, la spécificité est exprimée sur une échelle logarithmique : une valeur de spécificité de 2 signifie qu'il y a 1 chance sur 100 ($10^2 = 100$) que le hasard puisse expliquer la distribution observée. On considère généralement que le seuil à partir duquel une spécificité devient significative, dit «seuil de banalité» est de 3 (c'est-à-dire au plus 1 chance sur 1 000 que le hasard explique le phénomène). Pour des raisons pratiques d'affichage, le logiciel TXM «sature» et n'affiche pas plus de 1 000 comme valeur pour une spécificité. Peu importe la valeur réelle dans ce cas puisque cela correspond déjà à des distributions ayant seulement au mieux 1 chance sur 10^{1000} de se produire par hasard, donc l'affichage de cette valeur dans TXM suffit en pratique à exclure totalement une coïncidence.

Le calcul des spécificités permet entre autres d'étudier les associations lexicales fréquentes à travers le concept de cooccurrences. La notion classique de cooccurrent se base uniquement sur la réalisation dite «de surface» – c'est-à-dire sujette à l'ordre particulier dans lequel les mots apparaissent dans une phrase (par opposition à la «profondeur» d'un arbre syntaxique). En réitérant les mesures de cooccurrence pour former peu à peu un groupe de mots, il est même possible de repérer des motifs discursifs (Longrée et al., 2008) grâce la technique dite des «segments répétés». Comme de telles tournures de phrase ne sont pas entièrement figées, de nombreux mots peuvent occuper leurs positions encore libres et cette technique souffre donc d'un «bruit considérable» (Tutin & Kraif, 2016, p. 124). Pour lever ces difficultés, Tutin & Kraif (2016, p. 126) introduisent une méthode itérative semblable mais fondée sur des cooccurrents syntaxiques. L'idée consiste à rechercher les cooccurrents en se basant sur les relations de syntaxe plutôt que sur la proximité séquentielle des mots. À la différence d'une recherche traditionnelle, il n'y a donc pas de notion de «fenêtre» d'une taille paramétrable autour du motif considéré, la recherche se limite par construction à la phrase entière (au sens d'«arbre syntaxique», dont les éléments sont nécessairement connexes). En ajoutant progressivement à un noyau initial des cooccurrents syntaxiques forts, cette technique nommée «Arbres Lexico-syntaxiques Récurrents» (ALR) permet d'extraire des motifs courants dans un corpus. Cette approche est implémentée dans le Lexicoscope (Kraif, 2016) et a été utilisée pour mettre en évidence des routines discursives par exemple dans certains genres littéraires (Kraif et al., 2016) ou des écrits scientifiques (Ji et al., 2018).

Outre la notion de spécificité, la deuxième contribution majeure des mathématiques à l'école française de linguistique vient un peu plus tard, à partir de la fin des années 1960 avec le courant de l'analyse des données fondée par Benzécri (Pincemin, 2022, p. 1). Reprochant les approches trop centrées sur le vocabulaire de la statistique lexicale de Guiraud et Muller, il propose l'Analyse Factorielle des Correspondances – AFC – (Benzécri et al., 1973) qui est une technique de réduction de dimension. L'AFC permet de détecter les proximités et les différences entre les parties d'un corpus en représentant les associations qui existent entre le vocabulaires et les textes dans un espace à deux dimensions (Beaudouin, 2016, p. 20). L'emploi conjoint de statistiques et de techniques d'algèbre peut faire penser à un précurseur des méthodes plus récentes utilisées en AA comme les plongements

de mots. Il est intéressant de noter qu'il n'en est rien : le positionnement théorique de Benzécri est opposé non seulement à la linguistique de Chomsky mais également à l'IA (Beaudouin, 2016, p. 18). L'AFC construit une projection dans un espace abstrait qu'il appartient à la personne qui conduit les recherches d'interpréter. Pour Benzécri, l'ordinateur n'est pas «intelligent» ; bien au contraire le fait qu'il ne le soit pas est vu comme un gage d'objectivité car l'ordinateur n'a dès lors pas d'à priori et ne peut pas mentir (Beaudouin, 2016, p. 22). Cette conception est diamétralement opposée aux promesses d'intelligence des architectures les plus récentes de modèles d'AA comme les *Large Language Models* dont les «hallucinations» sont assez établies et documentées pour qu'une littérature leur soit consacrée (Benazha et al., 2024 ; Chelli et al., 2024 ; Farquhar et al., 2024).

2.5.3 Études contrastives

Forte de ces outils, la linguistique est à même de décrire les contrastes qui peuvent exister entre différents textes réunis au sein d'un même corpus. Plusieurs travaux comparent ainsi des genres textuels (Pamuksaç & Sanvido, 2023) y compris dans des domaines spécifiques comme la justice (Diot-Parvaz & Gianola, 2020) ou sur des thématiques particulières comme des questions sociétales et environnementales (Dang, 2020).

De nombreuses études explorent les différences à l'échelle d'auteurs individuels. Une des promesses des stylisticiens est bien de réussir à caractériser le style de différents auteurs (C. Labbé & Labbé, 2000 ; Muller & Brunet, 1988). Pincemin (2012, p. 19) mentionne également des travaux plus récents de textométrie faisant la synthèse des approches de la statistique lexicale et de l'analyse des données pour opposer différents genres présents au sein du corpus Frantext (voir section 2.2.2) ou dans l'œuvre d'un seul auteur – Victor Hugo – pour une étude de É. Brunet (2002). Il est intéressant de noter la ressemblance avec les applications de la classification automatique pour déterminer le sujet d'un texte ou l'orientation politique de son auteur (voir la section 2.4.2). Cependant, les deux démarches sont en réalité tout à fait opposées : alors que la finalité des tâches de classification réside complètement dans la répartition des textes dans des groupes cohérents, ces études textométriques s'attachent au contraire à identifier les critères qui permettent de rattacher un texte à une des catégories pour rendre explicite ce qui les caractérise.

À ce titre, le domaine politique semble particulièrement bénéficier des apports de la textométrie qui permet de révéler les positionnement thématiques et idéologiques qui transparaissent dans les discours. Un corpus regroupe les vœux prononcés traditionnellement au nouvel an par les présidents de la cinquième république. Mis à jour au fil des élections, il est distribué avec TXM et alimente de nombreuses études (Mayaffre et al., 2019 ; Pincemin, 2022). Les discours des syndicats font aussi l'objet d'analyses contrastives (Brugidou & Labbé, 2000). Des époques plus reculées comme l'entre-deux-guerres ne sont pas ignorées (Mayaffre, 2000). Enfin, la taille des intervalles de temps considérés permet la mise en place d'études des productions d'une seule personnalité politique où les contrastes sont envisagés selon la dimension temporelle. D. Labbé (1983) étudie ainsi le discours public de François Mitterrand dans la quinzaine d'années avant son élection au poste de président de la république puis celui de De Gaulle sur un peu plus d'une décennie (C. Labbé & Labbé, 2010).

De telles études, dites en diachronie, permettent de capturer des évolutions à des échelles variées selon la durée considérée, les changements les plus profonds dans le langage nécessitant davantage de temps pour se produire. Suivant le genre textuel considéré et l'ancienneté des époques, la finesse et la régularité des «graduations» sur l'axe temporel peut être dure à atteindre car il devient difficile de trouver assez de texte d'un genre considéré pour chaque intervalle du corpus. Vigier (2017), avec une segmentation par décennie, réussit à mettre en lumière les changements dans l'emploi des prépositions en français sur quatre siècles et demie. Quand le découpage du corpus en périodes distinctes est assez précis, il devient possible d'observer quand et sous quelle influence un changement a lieu, comme le montrent Hilpert & Gries (2016, p. 50). Les approches récentes les plus gourmandes

en données comme le WAC favorisent naturellement ce genre d'études, par exemple en sociolinguistique, permettant d'établir le cheminement d'une innovation linguistique. Tarrade et al. (2022) décrivent ainsi les différentes phases de propagation d'une innovation linguistique entre des locuteurs sur un réseau social, jusqu'à ce qu'elle devienne un changement persistant dans la langue ou bien soit abandonnée.

D'autres genres se prêtent naturellement moins à ce genre d'étude diachronique où le maillage temporel est contrôlé finement puisque certains types d'ouvrages notamment scientifiques ou lexicographiques ne peuvent être produits au même rythme et en même quantité que les messages de microblog. Pour les étudier, il faut se contenter des œuvres disponibles, parfois une seule par époque, en gardant à l'esprit les lacunes dans la couverture temporelle et le risque de mesurer des particularités propres à une œuvre donnée plus qu'à son époque. En s'appuyant de cette manière sur des manuels de physique, Mouhouche & El-Hajjami (2014) parvient à révéler les changements profonds dans la compréhension du phénomène physique de résonnance au travers de la terminologie employée pour en parler, de son origine accoustique jusqu'à ses applications modernes aux planètes pour véhiculer les notions d'accord et de transfert d'énergie. Blumenthal (2017) étudie le recours croissant aux locutions prépositionnelles dans l'Encyclopédie Universalis par rapport à l'*EDdA*. S'il ne peut pas dater finement les changements, il lui est tout de même possible de mettre en évidence les différences entre les deux époques (XVIII^{ème} et XX^{ème} siècles).

Les analyses diachroniques posent des défis techniques qui leur sont propres. Diwersy et al. (2017) ainsi que Vigier (2017) montrent par exemple les difficultés quant à l'ordre et à la segmentation des mots, et même quant à la graphie qu'implique le choix de rassembler dans un même corpus – PRESTO dans les deux cas – des états de langue aussi différents que le français des XVI^{ème} et XX^{ème} siècles. Cette question illustre elle aussi le fait qu'il n'existe pas de barrière étanche entre les approches computationnelles plus récentes considérées à la section précédente (2.4) et celles décrites dans la présente section. Les méthodes de plongement utilisées classiquement en AA pour représenter le sens des mots sont gênées par les glissements sémantiques qui surviennent dans les données d'entraînement. Étudier le sens des mots en diachronie avec ces méthodes nécessite donc des adaptations particulières (Hamilton et al., 2016 ; Kutuzov et al., 2018 ; Tahmasebi et al., 2019).

L'état de l'art qui s'achève avec ce chapitre a permis d'identifier trois axes principaux qui structurent l'ensemble de cette thèse et la situent sans ambiguïté en HN. Le premier est bien sûr celui des questions d'encodage qui sous-tendent tout le chapitre 3. La section 2.1 a décrit les formats utilisés, principalement le XML-TEI, et soulevé des difficultés qui seront discutées en particulier à la section 3.2. La section 2.2 a ensuite montré le lien étroit qui unit les œuvres lexicographiques et les travaux conduits en linguistique au cours du XX^{ème} siècle. Les quelques réflexions faites sur leur structure et sur les études déjà menées dans ce domaine alimenteront naturellement le chapitre 3, mais la section était également nécessaire pour amorcer un historique des contacts entre linguistique et informatique. Poursuivi à la section 2.4, cet historique a aidé à comprendre d'où venaient les outils conceptuels mobilisés dans les présents travaux et qui définissent les deux autres axes de l'étude. Les applications de l'AA aux tâches de classification permettent d'enrichir le corpus d'étude en associant aux articles un ensemble d'étiquettes représentant les domaines de connaissance dont ils relèvent. Cette métadonnée partage les textes suivant l'axe des disciplines en plus celui du temps, constitutif du corpus assemblé pour cette étude diachronique et qui distingue par construction les articles du XVIII^{ème} siècle issus de l'*EDdA* de ceux du XIX^{ème} provenant de *LGE*. En annotant les textes en morphosyntaxe et en syntaxe à l'aide d'autres méthodes d'AA, ce double partitionnement rend possible la conduite d'analyses contrastives. C'est ainsi que les méthodes des linguistiques de corpus décrites dans la présente section 2.5 sont utilisées au chapitre 5 pour identifier et caractériser les changements survenus dans les discours géographiques.

Chapitre 3

Préparation et enrichissement du corpus

3.1 Les œuvres du corpus . . . 44
3.1.1 Contextes historiques . . . 44
3.1.2 Organiser les connaissances . . . 47
3.1.3 Prétraitements . . . 58
3.2 Encodage des fichiers . . . 62
3.2.1 Des différences entre Encyclopédies et Dictionnaires . . . 62
3.2.2 Proposition d'un schéma d'encodage . . . 71
3.2.3 Le mythe du «texte brut» . . . 77
3.3 Des textes au corpus . . . 80
3.3.1 Différents formats pour différents usages . . . 80
3.3.2 Principes de structuration . . . 82
3.3.3 Reproductibilité . . . 86

La notion de corpus occupe une place centrale dans les HN, représentant l'objet autour duquel s'articulent de nombreux projets. Ce chapitre présente non seulement les tâches qui ont été réalisées le plus tôt dans cette thèse mais également celles qui ont requis le plus de travail. Toutefois, sa place avant les autres chapitres ne reflète pas un ordre strict de dépendance : au contraire, de nombreux allers et retours ont été nécessaires entre les données et les outils utilisés pour les exploiter. Ce chapitre est une tentative de saisir un état satisfaisant d'un effort en réalité continu et pouvant se poursuivre sans fin.

Il commence par présenter les objets de l'étude et introduire les concepts requis pour décrire ces encyclopédies et les articles qu'elles contiennent, au premier rang desquels la notion de domaine de connaissance qui sous-tend cette thèse de part en part. Puis, la section 3.2 s'intéresse aux conséquences des différences entre encyclopédies et dictionnaires sur l'encodage des fichiers, qu'elle aborde à l'aide de méthodes d'exploration systématiques de graphes. Enfin, la section 3.3, en accord avec la remarque ci-dessus sur l'impossibilité d'arrêter jamais l'amélioration de la qualité du corpus, tente de consigner les leçons apprises au fil du projet pour structurer les données de la manière la plus pratique et la plus flexible possible.

3.1 Les œuvres du corpus

Des quatre encyclopédies présentes dans le corpus du projet GEODE, cette thèse se concentre sur les deux premières dans l'ordre chronologique : l'*EDdA* et *LGE*. Ces deux encyclopédies couvrent respectivement la deuxième moitié du XVIII^{ème} siècle et la fin du XIX^{ème} siècle, la publication de *LGE* débutant en 1885, 113 ans après la fin de celle de l'*EDdA* qui s'achève en 1772.

La toute première étape des présents travaux a consisté à rassembler des versions numériques de ces textes et à les homogénéiser assez pour pouvoir leur appliquer les mêmes traitements dans les phases ultérieures. En effet, alors que sa position de symbole du siècle des *Lumières* a fait de l'*EDdA* un objet d'étude fréquent et facilement disponible en format numérique, *LGE* reste encore bien plus confidentielle avec une seule version imparfaite disponible au début de cette thèse. Cette section dresse l'état des lieux des versions des deux œuvres et retrace les étapes qui ont été nécessaires pour les constituer en corpus d'étude.

3.1.1 Contextes historiques

L'Encyclopédie de Diderot et d'Alembert

Dans le langage courant, les termes de «dictionnaire» et d'«encyclopédie» sont souvent utilisés de manière assez interchangeables pour désigner des livres qui regroupent de nombreuses connaissances et les ordonnent en listes de définitions alphabétiques. Leur similarité est visible jusque dans le titre complet de l'*EDdA*, «Encyclopédie ou dictionnaire raisonné». Si le mot «encyclopédie» fait aujourd'hui partie du vocabulaire ordinaire, il était bien plus surprenant et même controversé quand Diderot et d'Alembert ont choisi de l'utiliser pour leur œuvre.

Il ne s'agit pourtant pas d'un néologisme : le terme est déjà en usage au XVI^{ème} siècle, par exemple chez Rabelais qui l'emploie pour faire déclarer à Thaumaste que Panurge lui a ouvert «le vray puys et abisme de Encyclopédie». À cette époque, le terme renvoie encore principalement au concept abstrait de maîtrise simultanée de toutes les connaissances. C'est cette définition proche de l'étymologie grecque du terme que donne Furetière dans la première édition de son *Dictionnaire Universel* : un enchaînement de toutes les connaissances, de *κύκλος*, «cercle», et *παιδεία*, «connaissance». L'auteur critique déjà sa poursuite en tant qu'une forme d'hubris («C'est une témérité à un homme de vouloir posséder l'Encyclopédie») mais Basnage de Beauval (1702, p. 760) dans la deuxième édition de l'œuvre présente en outre le terme comme désuet et humoristique («plus en usage que dans le burlesque»). Il a pourtant été repris par Chambers qui en a fait le titre de son propre dictionnaire universel – «Cyclopedia».

Au-delà du reproche moral, le concept qui plaisait tant à Rabelais était quelque peu daté à la fin du XVII^{ème} siècle et faisait l'objet de la même critique dans le *Dictionnaire de Trévoux* (voir section 2.2.2 p.25). L'entrée *Encyclopédie* reste inchangée dans les quatres éditions parues entre 1721 et 1752, moquant le terme et déconseillant à son lectorat d'essayer de le cultiver. Dans ce but, l'article cite un poème de Pibrac qui encourage les gens à se spécialiser dans une seule discipline sans quoi ils risqueraient de ne jamais parvenir à atteindre la perfection dans aucun, avec une argumentation qui n'est pas sans rappeler le proverbe «qui trop embrasse mal étreint». Il est d'autant plus intéressant de constater que la définition demeure inchangée jusqu'à 1752, un an après la parution du premier volume de l'*EDdA*. Les Jésuites, éditeurs du *Dictionnaire de Trévoux*, désapprouvaient le projet de l'*EDdA* et ont réussi à le faire interdire la même année par le Conseil d'État en l'accusant de chercher à détruire l'autorité royale, à inspirer la rebellion et à corrompre la moralité de manière générale. Le combat idéologique ne s'arrêtait bien sûr pas aux mots, mais la volonté d'effacer le mot lui-même est tout à fait cohérent avec leur combat contre les philosophes des Lumières.

L'attaque n'est pas ignorée par Diderot qui fait commencer la définition même du mot «Encyclo-

pédie» dans l'*EDdA* par une réfutation ferme. Il commence par rejeter les inquiétudes formulées dans le *Dictionnaire de Trévoux* qui ne seraient selon lui que l'expression du manque de confiance de ses auteurs en leurs propres capacités avant de laisser l'argument principal à une citation latine du chancelier Bacon (Lojkine, 2013, p. 5) qui explique qu'une collaboration permet d'accomplir bien plus qu'il n'est possible à quiconque quel que soit son talent : ce qui ne serait sans doute pas à la portée d'une seule personne au cours d'une seule vie pourrait l'être à un effort commun sur plusieurs générations.

L'Histoire semble indiquer que les adversaires de Diderot ont pris sa défense de la faisabilité du projet très au sérieux. Six ans après la reprise de la publication, ils sont parvenus à faire révoquer le privilège de l'*EDdA* une deuxième fois (Moureau, 2001). Le résultat de cette interdiction fut que les dix tomes de texte restant ont dû être publiés illégalement jusqu'en 1765, ce qui a été rendu possible par la protection secrète de Malesherbes qui – malgré sa place à la tête de la censure royale – a sauvé les manuscrits de la destruction. Leur impression a eu lieu en cachette, en dehors de Paris, et les livres ont été (faussement) estampillés comme venant de Neufchâtel. Suite à la forte demande des libraires qui craignaient de perdre l'argent qu'ils avaient investi dans le projet, un privilège spécial fut émis pour les volumes de planches qui ont été publiés entre 1762 et 1772.

Malgré le tour romanesque qu'a dû prendre la publication de l'*EDdA*, dans leur dernière édition de 1771 les auteurs du *Dictionnaire de Trévoux* n'avaient plus d'autre choix que de reconnaître le succès des projets encyclopédiques du XVIII^{ème} siècle. Dans cette version la définition a été entièrement retravaillée, pour se contenter de reconnaître que les bonnes encyclopédies sont difficiles à écrire du fait de la somme de connaissances requises et de travail nécessaire pour se tenir à jour des avancées scientifiques. Il n'est plus possible de tourner l'idée même en ridicule. L'entrée reconnaît que la *Cyclopædia* de Chambers constitue une tentative décente avant de mentionner sans les nommer – mais de manière suffisament explicite – le projet de Diderot et d'Alembert, désignant le collectif par «Une Société de gens de Lettres» (un renvoi clair au sous-titre du projet) et précisant qu'il a débuté en 1751. Mais juste à côté de cette reconnaissance à demi-mot du succès du projet, deux nouvelles entrées font leur apparition : une pour un adjectif, «encyclopédique» et une autre pour le nom «encyclopédiste», témoignages éloquents de la portée de l'entreprise encyclopédique et de l'influence qu'elle a eu non seulement sur son époque mais sur la relation à la connaissance elle-même. Cette reconnaissance a bien entendu un retentissement particulier dans les pages du *Trévoux* mais elle ne s'y limite pas. Le terme connaît une évolution similaire dans le *Dictionnaire de l'Académie* (Pruvost, 2022, pp. 115–117).

Il existe au moins trois versions numériques de l'*EDdA* disponibles en ligne publiquement sous différentes conditions. L'Édition Numérique Collaborative et CRitique de l'Encyclopédie (ENCCRE dans le reste de ce manuscrit) édite un site web [1] offrant de nombreuses fonctionnalités dont un parcours linéaire du texte par tome qui associe à la version numérique du texte une photographie de la page dont il est extrait, ainsi qu'une interface de recherche permettant d'accéder directement à un article donné. Des métadonnées répertorient des informations objectives trouvées dans chaque article – auteur, présence de renvois, etc. – qui sont ensuite commentées par un ensemble de notes pouvant renvoyer à des fiches thématiques. Une deuxième source est l'American and French Research on the Treasury of the French Language (dorénavant ARTFL) qui diffuse également sa version du texte [2] au moyen du logiciel Philologic4 [3]. L'outil permet des recherches avancées incluant un filtrage des résultats suivant certaines métadonnées. Le texte, lisible depuis l'interface web générée par Philologic4 contient des liens vers des photos des pages pour le retour au texte. Enfin, Wikisource dispose également d'une version [4], partiellement corrigée par sa communauté, sous forme d'un ensemble de

1. https://enccre.academie-sciences.fr/
2. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/
3. https://artfl-project.uchicago.edu/philologic4
4. https://fr.wikisource.org/wiki/L%E2%80%99Encyclop%C3%A9die/1re_%C3%A9dition

pages qu'il est possible de consulter mais sans outil de recherche et sans métadonnées.

Si ces interfaces de consultation permettent une navigation facile des textes, l'application systématique d'outils de traitement requiert un accès à des fichiers contenant à la fois le plein texte et des métadonnées exploitables. L'ARTFL a accepté de partager ses données avec les membres du projet GEODE et c'est donc sa version qui a servi de base à cette étude pour le contenu textuel des articles. Chacun des 17 tomes de l'œuvre est représenté par un fichier encodé au format XML-TEI. Chaque article au sein d'un tome contient les métadonnées accessibles depuis l'interface de Philologic4. Dans le cadre de certains travaux du projet, elles ont été recoupées avec celles accessibles depuis l'interface de l'ENCCRE. Toutes les photos d'article de l'*EDdA* proviennent du site de l'ENCCRE. Le texte comprend au total 74 198 articles et 23 786 713 tokens.

La Grande Encyclopédie

La Grande Encyclopédie, Inventaire raisonné des Sciences, des Lettres et des Arts par une Société de savants et de gens de lettres fut publiée en France entre 1885 et 1902 par une équipe de plus de deux cent spécialistes organisés en onze sections. Ce texte comprend 31 tomes d'environ 1200 pages chacun et fut, d'après Jacquet-Pfau (2015) la dernière entreprise encyclopédique française majeure à marcher dans les traces de l'ancêtre prestigieux que fut l'*EDdA*, publiée environ 130 ans plus tôt.

Le titre complet de l'œuvre montre déjà la volonté de filiation avec l'*EDdA* de ses auteurs dans leur choix de sous-titre (Jacquet-Pfau, 2022, p. 96). La préface décrit la France comme une précurseure des grandes entreprises encyclopédiques, rattrapée depuis par les grandes puissances occidentales et qui n'a pas su se mettre à jour depuis l'*EDdA*. Les éditeurs mettent en avant les illustrations qu'ils placent dans la «tradition des encyclopédistes du siècle dernier», malgré l'évolution technique qu'elles apportent en étant intégrées au texte par contrastes avec les volumes de planches séparées du textes de l'*EDdA*. Le choix visionnaire des philosophes des Lumières de traiter des métiers, techniques et outils est présenté comme précurseur de la «société industrielle» du XIX^{ème} siècle.

À l'issue d'un historique qui remonte jusqu'à Démocrite en passant par Aristote, l'*EDdA* fait l'objet d'une anecdote de Voltaire rapportant un souper au Trianon. Dans ce récit, la cour expose au roi l'intérêt d'une encyclopédie et la qualité de l'*EDdA* en particulier, ce qui l'amène à revenir sur sa décision d'interdire l'œuvre. La préface s'achève par une réflexion sur la vie nécessairement «éphémère» des encyclopédies et le besoin de les renouveler pour les adapter aux progrès des sciences et aux évolutions du langage. Le tout dernier paragraphe prend la forme d'un souhait, celui de marquer son époque et de s'inscrire dans la lignée du «souvenir de Diderot et d'Alembert».

Cet espoir a été déçu dans une large mesure puisque *LGE* tombe rapidement dans l'oubli (Vigier, 2022, p. 18). Le projet connaît des difficultés financières dès les dernières années de sa publication (Jacquet-Pfau, 2015, p. 90). La Librairie Larousse, qui édite le *Grand Dictionnaire Universel*, rachète *LGE*, non pas pour développer une offre plus scientifique mais bien pour se débarasser d'un concurrent. Elle disparaît peu après malgré sa grande rigueur et ses qualités scientifiques – des articles de *LGE* sont parfois cités dans des publications (Jacquet-Pfau, 2015, p. 91) et les plus longs, véritables dossiers sur un sujet, connaissent une vie hors de ses pages en étant publiés sous formes de monographies (Jacquet-Pfau, 2022, p. 99).

Une version numérique de cette œuvre a été réalisée par la BnF et mise en ligne [5] en 2007. Basée sur une réimpression non datée de l'édition originale, elle comprend une image par page de l'œuvre, numérisée en niveau de gris à une résolution de 300x300 pixels par pouce. De ces fichiers images a été tirée une version partielle du texte par application d'un programme de reconnaissance optique de caractères (OCR). Cette version présente un certains nombre de limites empêchant de mener une étude intégrale du texte par des moyens automatiques comme la textométrie et faisait écrire à

5. http://catalogue.bnf.fr/ark:/12148/cb377013071

Jacquet-Pfau (2015, p. 88) qu'«il n'existe pas de version numérique complète et structurée en format texte de *LGE*».

D'abord, le texte est incomplet : si un grand nombre de tomes ont été OCRisés, certains comme par exemple les tomes 5 ou 18 n'ont pas été traités et aucun texte n'est disponible pour ces volumes sur le site de Gallica [6]. Cet état rend impossible une étude exhaustive mais rien ne suggère qu'une étude basée sur les tomes disponibles serait sujette à un biais particulier, puisque les tomes OCRisés ne semblent pas avoir été choisis selon une logique précise, ceux qui manquent ne sont par exemple pas contigus ni au début ni à la fin de l'œuvre. Ensuite, cette version en «texte brut» (en réalité une page HTML avec un balisage minimal) ne comporte qu'une annotation très superficielle et n'est en particulier pas segmentée en articles. Cela est un obstacle majeur pour les présents travaux, puisque l'unité d'étude du présent corpus est l'article, permettant de mettre en évidence des régularités propres à un domaine de connaissance ou à un auteur précis en vue d'analyses contrastives. Enfin, des erreurs dans la détection de l'organisation de la page (OLR) obscurcissent significativement le texte en opérant des permutations locales de son contenu qui viennent parfois mélanger des morceaux d'articles entre eux. Cela complique nettement la segmentation du texte en articles et vient dans tous les cas endommager la structure des phrases. Ces nouvelles erreurs se répercutent à leur tour sur les phases ultérieures d'analyse, causant des problèmes dans les annotations morphosyntaxiques et syntaxiques qu'il est nécessaire d'appliquer au texte pour faire de la textométrie.

Dans le but de pallier ces défauts, le projet CollEx Persée DISCO-LGE [7] (2019-2021) a entrepris de renumériser cette encyclopédie en partenariat avec la BnF et d'en produire une version encodée en XML-TEI. Cette nouvelle version a été réalisée à partir de photographies d'un exemplaire original [8] situé à la Bibliothèque de l'Arsenal à Paris [9].

Ce projet a pris fin en août 2021 et a abouti à la publication sur Nakala [10], le dépôt de données de la Très Grande Infrastructure de Recherche Huma-Num, d'une nouvelle version de l'œuvre sous différents formats, dont des segmentations par article en XML-TEI et en format texte (`.txt`). Cette version compte 229 475 articles pour un total de 54 936 266 tokens. L'identification automatique des débuts d'article y a causé davantage de faux positifs (séquence incorrectement considérée comme un début d'article donnant lieu à un surdécoupage du vrai article en plusieurs) que de faux négatifs (débuts d'article non reconnu par le système causant la fusion de deux vrais articles) et ce chiffre semble donc (largement ?) surestimé.

3.1.2 Organiser les connaissances

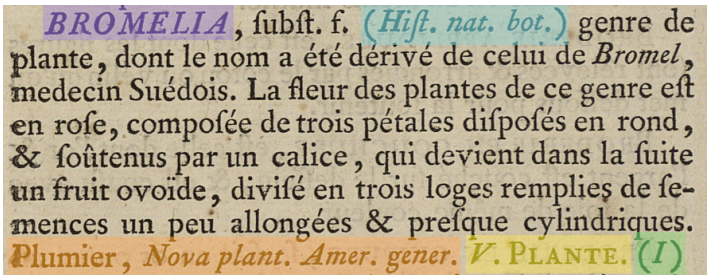
La problématique de structuration de la connaissance, comme cela a été décrit à l'introduction (voir la section 1.2 page 11), est au cœur des présents travaux. Elle a bien sûr une forte valeur intrinsèque pour les auteurs de chacune des encyclopédies du corpus mais, dans le cadre de cette thèse, elle est ce qui permet de singulariser et de problématiser la Géographie. Elle est la raison pour laquelle des efforts sont déployés au chapitre 4 pour appliquer une classification automatique aux articles, et elle fonde la démarche contrastive adoptée au chapitre 5. Pour commencer, il s'agit de définir plusieurs concepts fondamentaux liés à la notion et à sa matérialisation dans les articles.

Anatomie d'un article d'encyclopédie

Les premiers de ces concepts sont purement typographiques. En vue de pouvoir évoquer les différentes parties d'une encyclopédie, il est nécessaire de bien définir ce qui constitue les articles et leur

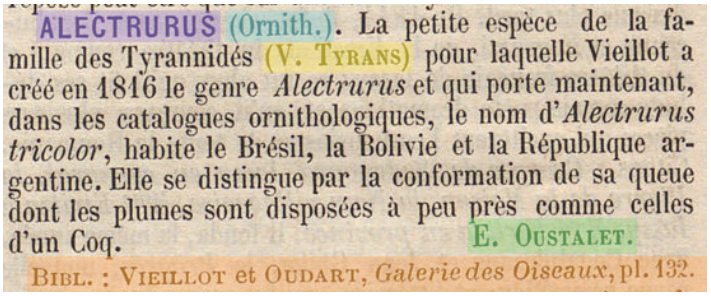
6. https://gallica.bnf.fr/ark:/12148/bpt6k246407#
7. https://www.collexpersee.eu/projet/disco-lge/
8. https://catalogue.bnf.fr/ark:/12148/cb41651490t
9. https://www.bnf.fr/fr/arsenal
10. https://nakala.fr/10.34847/nkl.74eb1xfd

environnement sur les pages. La figure 3.1 offre un aperçu des éléments pouvant apparaître dans un article (non représentatif, les deux articles ont été choisis pour leur concentration élevée).



BROMELIA, ſubſt. f. (*Hiſt. nat. bot.*) genre de plante, dont le nom a été dérivé de celui de *Bromel*, medecin Suédois. La fleur des plantes de ce genre eſt en roſe, compoſée de trois pétales diſpoſés en rond, & ſoûtenus par un calice, qui devient dans la ſuite un fruit ovoïde, diviſé en trois loges remplies de ſemences un peu allongées & preſque cylindriques. Plumier, *Nova plant. Amer. gener.* *V.* Plante. (*I*)

(a) Article BROMELIA dans l'*EDdA*, T2, p.434



ALECTRURUS (Ornith.). La petite eſpèce de la famille des Tyrannidés (V. Tyrans) pour laquelle Vieillot a créé en 1816 le genre *Alectrurus* et qui porte maintenant, dans les catalogues ornithologiques, le nom d'*Alectrurus tricolor*, habite le Brésil, la Bolivie et la République argentine. Elle se distingue par la conformation de sa queue dont les plumes sont disposées à peu près comme celles d'un Coq. E. Oustalet.

Bibl. : Vieillot et Oudart, *Galerie des Oiseaux*, pl. 132.

(b) Article ALECTRURUS dans *LGE*, T2, p.68

Figure 3.1 – Les différents constituants possibles d'un article d'encyclopédie

Dans une encyclopédie, une entrée associe un texte à un mot ou un groupe de mots, la «vedette», qui sert tout à la fois de «titre» à l'article et de point de référence dans l'ensemble du texte puisque ces vedettes engendrent l'ordre (alphabétique) des articles tout en servant de lien pour les renvois. Tout article d'encyclopédie comporte nécessairement une vedette typographiquement marquée comme c'est le cas pour les deux articles de la figure 3.1 où les vedettes, surlignées en mauve, apparaissent en majuscules. Dans l'*EDdA*, il existe même en réalité deux niveaux de division : certains termes sujets à une forte homonymie ou possédant de nombreux dérivés contiennent plusieurs entrées dont les vedettes sont graphiées en petites capitales. Pour cette raison, l'ENCCRE numérote les textes sur trois niveaux, ajoutant aux numéros de tome et d'«adresse» (nom utilisé par l'ARTFL pour désigner la structure correspondant aux articles dans le cas général) un numéro d'«entrée». L'ARTFL se limite à deux niveaux et regroupe parfois plusieurs entrées au sein d'un même «article», une différence à l'origine de certaines complications rencontrées lors des travaux de classification (voir section 4.1.1 p.92).

La vedette a une existence grammaticale à part du reste de l'article, ce qui confirme son rôle de «clef» dans la gigantesque table associative formée par une encyclopédie : quand elle n'est pas séparée de la première phrase par un point comme c'est le cas de la figure 3.1b, elle est souvent simplement apposée et tout à fait optionnelle, séparée d'une virgule du reste de la première phrase qui est alors le plus souvent nominale, déportant tout syntagme verbal dans une subordonnée relative comme le montre la figure 3.1a. Elle peut être suivie optionnellement d'un «désignant» entre parenthèses (en bleu sur la figure 3.1) qui sert à situer l'entrée dans un espace structuré de connaissances (notion centrale de cette sous-section et qui est développée plus bas dans le segment 3.1.2). Un désignant peut ainsi référer à un ou plusieurs domaines de connaissance, souvent abréviés.

Les articles ne constituent pas une collection d'informations isolées mais sont au contraire liés entre eux par un système de renvoi, qui constitue un deuxième mode d'organisation en sus du système de domaines de connaissance (Blanchard & Olsen, 2002, p. 46). Les articles peuvent ainsi contenir des renvois, figurés en jaune sur la figure 3.1 . Ils sont constitués du verbe «voir» à l'impératif

ou à l'infinitif jussif et de la vedette d'un autre article (elle est bien en majuscule, éventuellement dans une fonte plus petite) et possèdent une valeur illocutoire en invitant le lectorat à tourner les pages du livre ou à aller chercher un autre tome en vue de poursuivre la lecture dans l'article ainsi ciblé. À l'origine une simple phrase parmi les autres au milieu du texte, avec le verbe écrit en toutes lettres (cette forme reste plus fréquente dans l'*EDdA*), les renvois ont évolué et pour prendre la forme stylisée visible dans les deux exemples, où le verbe n'existe plus qu'au travers de son initiale, et où le renvoi dans son ensemble peut être mis entre parenthèses comme dans la figure 3.1b, ce qui n'est pas sans préfigurer des dispositifs modernes tels que les liens hypertextes. On trouve parfois aussi une formulation réflexive pour éviter une répétition («V. ce mot»), établissant implicitement le lien entre l'occurrence du mot dans la phrase précédente à la vedette correspondante. Il semble que cette tournure s'accompagne plus fréquemment d'une forme complète du verbe «voir» ou «voyez» que la forme directe où la vedette est explicite.

Enfin, les articles sont parfois signés (en vert sur la figure 3.1). La signature de l'article BROMELIA (figure 3.1a) n'est pas une initiale, ce 'I' entre parenthèses est utilisé par Louis-Jean-Marie Daubenton [11]. Les articles peuvent mentionner occasionnellement leurs sources, en orange sur la même figure. Celles de *LGE* sont particulièrement fréquentes et riches, faisant l'objet d'un dernier paragraphe marqué systématiquement par l'abréviation «Bibl.» (visible sur la figure 3.1b) et pouvant parfois s'étendre sur plusieurs colonnes de texte comme c'est le cas pour l'article COLONISATION (La Grande Encyclopédie, T11, p.1066) – 3 colonnes et demie de bibliographie, s'étendant de la page 1117 à la page 1119.

Pour une encyclopédie imprimée sur un support papier, les articles apparaissent les uns à la suite des autres dans l'ordre alphabétique (sans tenir compte des accents et autres diacritiques) sur une ou plusieurs colonnes (2 colonnes dans l'*EDdA* comme dans *LGE*). Autour des articles se trouvent divers éléments «péritextes» mis en évidence à la figure 3.2. Le haut d'une page comporte généralement le numéro de la page dans le tome courant, surligné en vert sur la figure 3.2. Pour aider à la recherche d'un article donné, des indicateurs visibles en bleu sur la figure 3.2 représentent l'intervalle de l'index alphabétique couvert par la page courante, en donnant les positions de l'article auquel appartiennent la première et la dernière ligne de la page. Dans l'*EDdA* (figure 3.2a), il s'agit seulement des trois premières lettres des vedettes de ces articles, dans *LGE* (figure 3.2b) des vedettes complètes. L'évolution des techniques d'imprimerie permet de positionner des figures au milieu des colonnes de texte dans *LGE* accompagnées de leurs légendes (en mauve sur la figure 3.2b). Dans l'*EDdA*, les figures et leurs légendes sont présentes mais confinées aux tomes de planches.

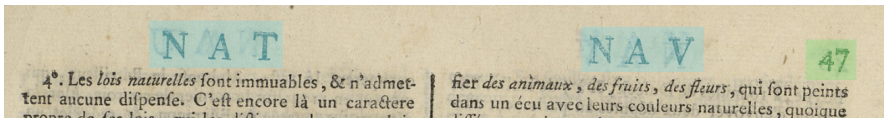
Les techniques de reliure imposent l'impression de ces encyclopédies par «cahiers» (de 8 pages pour l'*EDdA* – *in octavo* – et de 16 pages pour *LGE* – *in sextodecimo*). Chacun de ces cahiers porte une inscription qui rappelle le tome auquel il appartient (en jaune sur la figure 3.3) et la position du cahier dans l'ordre de reliure (en orange sur la figure 3.3). Cette position est exprimée par une lettre dans l'*EDdA* (ici un 'G') et par un nombre entier dans *LGE*. Par construction, ces inscriptions reviennent régulièrement : toutes les 8 pages dans l'*EDdA*, environ toutes les 16 pages dans *LGE* (la présence de cartes qui sont imprimées à part et insérées entre les pages des cahiers vient perturber cette périodicité).

Ces définitions posées, il devient possible de s'intéresser au concept principal de cette sous-section qui est étroitement lié à la notion de désignant.

Notion de domaine de connaissance

Comprendre l'organisation de la connaissance est un champ de recherche à l'intersection de plusieurs disciplines (Blair, 2010 ; Bowker & Star, 2000 ; Foucault, 1970 ; Wellmon, 2015). Les historiens,

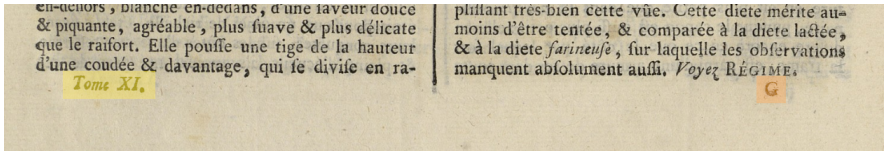
11. https ://artflsrv04.uchicago.edu/philologic4.7/kafker/navigate/1/33



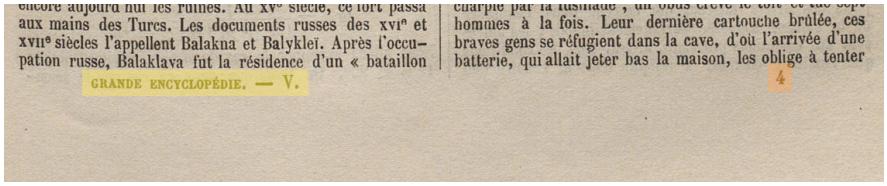
(a) Haut de la page 47 du tome 11 de l'*EDdA*

(b) Haut de la page 48 du tome 5 de *LGE*

Figure 3.2 – Les différents éléments péritextes en haut et en cours de pages



(a) Bas de la page 49 du tome 11 de l'*EDdA*



(b) Bas de la page 49 du tome 5 de *LGE*

Figure 3.3 – Les différents éléments péritextes en bas de page

par exemple, étudient quels types de documents contiennent des connaissances, comment la connaissance est divisée en catégories et quelles relations ces catégories entretiennent (Groult, 2003 ; Holmberg, 2017). Dans certains documents historiques, cette classification est visible : entre le XVII^{ème} et le XIX^{ème} siècle des encyclopédies ont ainsi commencé à la matérialiser au début des articles.

Les systèmes de classification choisis par les éditeurs d'une encyclopédie reflètent le contexte spécifique des débats intellectuels contemporains de sa publication, qui peut transparaître par exemple dans l'emploi d'un désignant plutôt qu'un autre comme le montre Viard (2006, para. 33). Les encyclopédies sont donc un terrain particulièrement intéressant pour étudier comment la classification de la connaissance change au cours du temps. Dans le cadre de cette thèse, la notion de classification permet surtout d'accéder en première approximation aux discours géographiques qui sont l'objet de cette thèse puisque tant l'*EDdA* que *LGE* considèrent la géographie comme un domaine de connaissance parmi les autres.

Là où certains philosophes tentent de trouver une organisation des connaissances satisfaisantes en elle-même en tant que description du monde dans sa totalité, formant des systèmes à priori dans lesquels n'importe quelle connaissance – contemporaine du système ou ultérieure – est censée pouvoir ensuite trouver sa place, d'autres approches plus empiriques sont également possibles. Ainsi, alors qu'il répertorie et ordonne deux années de communications scientifiques dans les *Philosophical Transactions of the Royal Society*, Lowthorp (1665, p. 405) propose une organisation tripartite. Il distingue une «Histoire naturelle» tout à fait générale, une catégorie des «Singularités» – tout ce qui échappe à la norme définie par la première catégorie rendant par construction le système complet – et une troisième catégorie méta qui traite des outils pour étudier les catégories précédentes. Une telle organisation à postériori est par construction plus facile à équilibrer mais elle a peu de chances de résister à l'épreuve du temps ou d'être adaptée à plusieurs œuvres encyclopédiques. Il est intéressant de noter que cette classification sommaire, présentée comme «a more Natural Method» sert surtout initialement à se libérer de l'arbitraire de l'ordre alphabétique.

Dans l'*EDdA*, l'organisation de la connaissance intègre non seulement les sciences mais également les arts et les métiers. Elle est fondée sur une autre partition en trois, celle proposée par Bacon qui rattache tout savoir aux facultés humaines de la mémoire, l'imagination et la raison (Jaquet, 2010, para. 10). Cette distinction est fondamentale dans la démarche des encyclopédistes : revendiquée par d'Alembert dès le Discours Préliminaire (L'Encyclopédie, T1, p.xvj), elle est aussi visible dans les trois colonnes qui partagent le «Systême figuré des connoissances humaines» reproduit à la figure 3.4. Elle témoigne d'une structuration à priori des sciences bien plus ambitieuse que la division très pragmatique de Lowthorp. Sa catégorie des Singularités semble tout de même trouver un écho dans la branche des «Écarts de la Nature» du «Systême» (voir à gauche de l'arbre de la figure 3.4 dans la moitié supérieure) mais c'est bien là la seule ressemblance. Il ne faut néanmoins pas croire que les efforts d'organisation des auteurs de l'*EDdA* s'arrêtent à cette taxonomie. Un vaste réseau de renvois structure transversalement les articles et suggère des navigations de l'œuvre, véritables liens «hypertextes» avant l'heure (Blanchard & Olsen, 2002, p. 47). Les encyclopédistes ne commettent pas non plus l'erreur de prendre le «Systême» pour plus qu'il n'est : une déclaration d'intention et non pas une «grille fixe» (Cernuschi, 2006, para. 3) dans lequel les articles devraient à tout prix trouver leur place. Des branches de l'arbre demeurent sans articles et à l'inverse des articles sont affectés à des domaines qui ne sont pas mentionnés par le «Systême figuré» (Leca-Tsiomis, 2006, para. 4).

En pratique, les domaines de connaissance se manifestent la plupart du temps au niveau des articles sous la forme des désignants présentés dans la partie précédente de cette section (p.47). Certains articles n'ont pas de désignant sous cette forme marquée typographiquement (entre parenthèses, en tête d'article, souvent abréviée), mais reçoivent tout de même une indication quant à leur domaine d'appartenance, une *marque de domaine*. Une tournure telle que «en termes de» est très utilisée à cet effet, suivie du nom d'une science – par exemple à l'article DISQUE «en termes de Botanique»

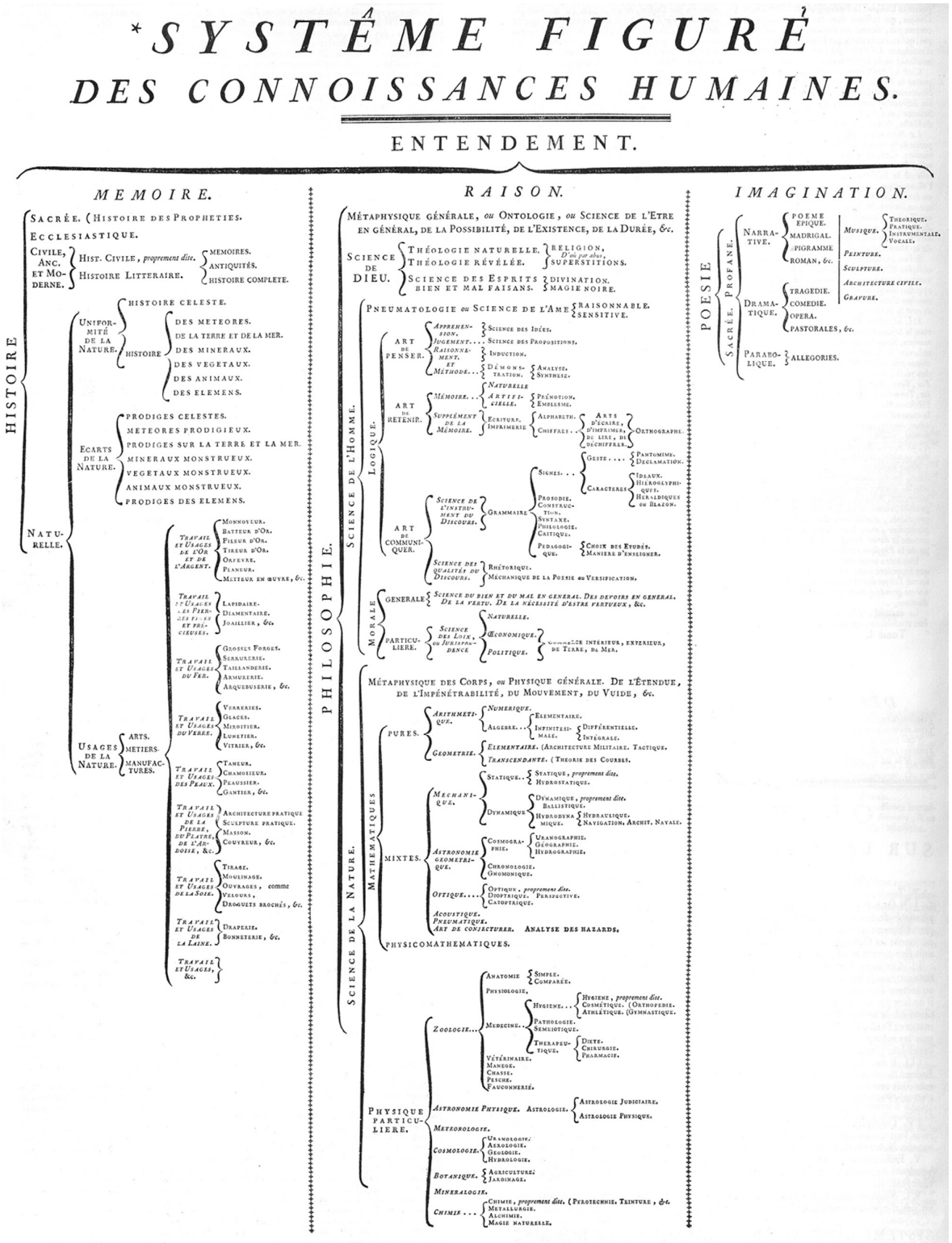


Figure 3.4 – Le «Systême figuré des connoissances humaines», une taxonomie au cœur de l'*EDdA*

(L'Encyclopédie, T4, p.1045) – ou d'une profession – comme dans l'article AMBRISE «en termes de Fleuriste» (L'Encyclopédie, T1, p.326). On trouve aussi parfois simplement «en» suivi du nom d'une science comme l'illustre l'article ÉLASTICITÉ – «en Physique» (L'Encyclopédie, T5, p.444). Une troisième tournure possible est *PRÉPOSITION* suivi de «les» («chez les», «dans les», «parmi les»...), qui peut accepter différents types de groupes humains, qu'ils soient déterminés par une activité comme à l'article A A A visible à la figure 3.5 – «chez les chimistes» (L'Encyclopédie, T1, p.5) – des pratiques par exemple dans l'entrée CALVAIRE – «chez les Chrétiens» (L'Encyclopédie, T2, p.565) – ou une simple appartenance civilisationnelle – on pourra penser à «chez les Latins» de l'article PAVE (L'Encyclopédie, T12, p.192). Outre «marques de domaine», l'ENCCRE qualifie également ces expressions de «désignants *non marqués*» ou «*flottants*» [12], le terme étant employé au moins par Boussuge & Launay (2019, p. 237). Leur usage s'est répandu à la fin du XVII^{ème} siècle dans le dictionnaire de Richelet avant d'accéder au statut de «tradition lexicographique» (Quemada, 1968, p. 307) une fois repris dans le *Furetière* et le *Trévoux*. Leur présence ou non n'obéit à aucune règle évidente, et les formulations restent très variables. Pour trouver un semblant d'ordre, on peut observer que les différents groupes répertoriés précédemment offrent en quelque sorte un dégradé de valeurs classifiantes, qu'il est possible de jauger en tentant de les reformuler. Alors que «chez les chimistes» est un équivalent assez clair du désignant «Chimie», il n'y a pas de domaine «Chrétienté» mais la position de ce «chez les Chrétiens» en début d'article, apposé à la vedette avant le verbe de la phrase évoque encore beaucoup l'usage précédent. Il ne s'agit déjà plus d'un désignant puisque l'article en est par ailleurs pourvu – «(Hist. ecclés.)» – mais une reformulation possible de l'expression en «Religion chrétienne» montre encore une certain proximité de cet emploi avec un désignant. Pour «chez les Latins», en revanche, il n'y a plus qu'une valeur circonstancielle malgré sa position relativement en début d'article : le terme défini (pavé) n'est pas constitutif d'un ensemble de pratiques qui définirait le groupe comme ce pouvait être le cas avec les chimistes et les chrétiens.

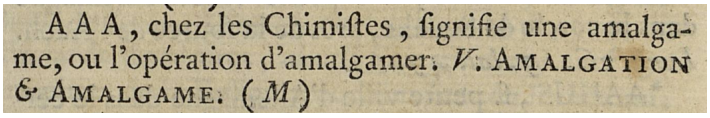

A A A, chez les Chimiſtes, ſignifie une amalgame, ou l'opération d'amalgamer. *V.* Amalgation & Amalgame. (*M*)

Figure 3.5 – Article A A A dans l'*EDdA*, T1, p.5

À la suite de Leca-Tsiomis (2006, para. 1), on réservera le terme de «désignant» aux formes marquées typographiquement, placées entre parenthèse en début d'article. Les désignants sont en toute rigueur un cas particulier de marques de domaine mais on utilisera les deux termes en opposition, pour distinguer les marques de domaine qui ne sont pas des désignants. Les désignants résultent d'une simplification et d'une codification au fil du temps des marques de domaine, leurs «ancêtres» d'après Leca-Tsiomis (2006, para. 8). Ils sont en effet absents d'œuvres antérieures comme le *Dictionnaire Universel* de Basnage ou même de la 3ème édition du *Dictionnaire de Trévoux* (celle de 1743). Dans ces œuvres en revanche, des occurrences de «en termes de» se rapprochent des débuts d'articles là où se trouverait un désignant. C'est dans la *Cyclopedia* de Chambers que le mot «terme» disparaît, restant sous-entendu puisque l'œuvre est un recueil de terminologie (Leca-Tsiomis, 2006, para. 12), laissant apparaître le nom de la discipline seule. Dans l'*EDdA*, le concept de désignant est tout à fait né et des marques de domaine se trouvent même parfois entourées de parenthèses quand elles apparaîssent à la même place, comme c'est le cas aux articles BIZEGLE – «(chez les Cordonniers)» (L'Encyclopédie, T2, p.268) – ou CANON – «(terme de Rubannier)» (L'Encyclopédie, T2, p.618). La présence de la majuscule comme dans les désignants vient également rompre la lecture normale de l'expression en transformant le nom de profession en une sorte d'objet symbolique, le nom d'une classe et plus le mot dans son acception courante. Enfin ces expressions semblent avoir quasiment disparu de *LGE*. «En termes de» présente 2433 occurrences dans l'*EDdA* contre seulement 51 dans *LGE* (pourtant 2.3 fois plus volumineuse d'après les chiffres présentés dans la sous-section 3.1.1).

12. http://enccre.academie-sciences.fr/encyclopedie/politique-editoriale/?s=31&

Cela confère au motif une spécificité d'au moins 1 000 dans l'*EDdA* (voir au sujet de la saturation de l'affichage des spécificités dans TXM la section 2.5.2 p.39). «Chez les» suivi d'un mot commençant par une majuscule est plus difficile à exploiter puisque d'après la remarque précédente la séquence n'a pas toujours la même valeur qu'un désignant mais dans tous les cas elle est également moins représentée : 713 occurrences dans l'*EDdA* contre 882 dans *LGE*, le motif obtient donc une spécificité de 33 dans l'*EDdA* ce qui est suffisant pour écarter une répartition seulement dûe au hasard.

Dans *LGE*, il n'y a pas d'arborescence des connaissances mais la préface accorde une place majeure à la notion de domaine. Dès la première page (La Grande Encyclopédie, T1, p.I), ses auteurs justifient le bien-fondé de leur entreprise par l'absence d'une œuvre française à jour en termes de connaissances scientifiques dans le paysage encyclopédique de l'époque. Ils établissent ensuite une longue liste de domaines de connaissance pour montrer la grande variété de sujets que *LGE* se propose de couvrir («en un mot, tout ce qui est de nature à jeter la lumière sur le monde physique et sur le monde intellectuel»). Plus que réellement des noms de sciences, il s'agit plutôt de thématiques choisies pour mettre en valeur les intérêts de l'époque – ce qui permet la mention des «applications nouvelles de l'électricité» alors que la physique et la chimie dont elles relèvent pourtant figurent déjà dans la liste. La géographie quant à elle n'est même pas mentionnée. Elle n'apparaît que plus bas sur la même page, ses découvertes étant associées au «développement colonial» pour intégrer les cartes, «gravées spécialement pour» *LGE*, dans un paragraphe qui met en valeur l'apport des illustrations dans l'œuvre. La place importante qu'y occupent les cartes prend un sens tout particulier quand on sait la volonté de l'œuvre de s'inscrire dans l'héritage de l'*EDdA*. Alors que cette dernière ne comporte en effet aucune carte, *LGE* incarne cette postérité qui dresse des cartes à laquelle aspirait Diderot dans l'article ENCYCLOPÉDIE (L'Encyclopédie, T5, p.646) évoqué à la section 1.1 (voir p.8).

Mais plus encore, la notion de domaine de connaissance est décrite dans les pages qui suivent comme primordiale dans tout projet encyclopédique au-delà de *LGE* seule. Après un incontournable rappel historique qui évoque chronologiquement les étapes majeures qui ont mené au concept moderne d'encyclopédie depuis l'étymologie grecque du terme, la préface revient au *Discours préliminaire* de l'*EDdA* écrit par d'Alembert pour souligner l'importance de l'organisation et de la structuration de la connaissance dans les encyclopédies, présentées comme des traits caractéristiques qui les distinguent des dictionnaires (la sous-section 3.2.1 de la section suivante explore plus en détail cette problématique et met en évidence d'autres différences mais il est intéressant de voir que cette séparation était déjà perçue au XVIII^{ème} siècle avant que le terme d'encyclopédie ne réussisse à s'imposer). D'après cette partie de la préface, la démarche encyclopédique consiste à révéler les liens qui relient tous les savoirs humains et à les organiser, partant des faits concrets pour généraliser vers l'abstraction et établir ainsi ce qui constitue les «lois naturelles». Mais à la suite de d'Alembert qui comme le fait remarquer Hardesty Doig (2006, para. 1) reconnaît la part d'arbitraire inhérente à tout «système» (toujours dans le *Discours préliminaire*), les auteurs ont bien conscience qu'il n'y a pas de preuve que l'ensemble des connaissances admette un ordre. La volonté d'y trouver une logique pourrait simplement s'expliquer par la façon dont fonctionne l'esprit humain. C'est à ce moment que l'encyclopédisme s'écarte de la science en nécessitant un acte de foi : les auteurs de *LGE* affirment que faire une encyclopédie revient à décider qu'une telle structure existe et à tenter d'y parvenir. On assiste ainsi à un certain renversement de la problématique de départ puisque la classification ne serait pas un simple aspect de la conception d'une encyclopédie mais au contraire *le* point de départ d'un projet fondamentalement positiviste et *la* question centrale à laquelle elle prétend répondre en proposant un système d'organisation des connaissances.

Une fois posé ce principe, le reste de la préface est bien plus pragmatique : il ne s'agit plus que de choisir un ensemble de domaines pour classer les articles. Fidèle à l'esprit de l'*EDdA*, les auteurs de *LGE* ancrent eux aussi leur recherche d'un système de classification dans les travaux de Bacon. Ils tiennent compte des tentatives de partition des sciences effectuées au XIX^{ème} siècle et mentionnent celle d'Ampère basée sur la dichotomie entre nature et esprit (opposition *cosmologique*/*noologique*),

celle d'Auguste Comte fondée sur l'ordre historique entre les sciences et celle de Herbert Spencer qui s'attache à la division entre science abstraite et concrète. Conscients de l'impossibilité d'«arrêter une classification définitive» les auteurs de *LGE* optent pour un système inspiré de l'approche historique d'Auguste Comte remise à jour.

L'avant-propos poursuit ce travail en fixant un objectif de taille à l'ensemble du projet avant d'accorder une place à chacun des domaines (La Grande Encyclopédie, T1, p.XI). C'est la première et seule apparition de la liste des 14 domaines retenus plus un domaine «Matières diverses» pour absorber tout élément qui mettrait en défaut ces domaines, témoin s'il était besoin du pragmatisme de l'approche à postériori de leur système – un outil pratique tout au plus mais sans illusions sur une quelconque valeur particulière d'un point de vue philosophique. Ces choix donnent le tableau présenté figure 3.6, d'aspect presque trivial par contraste avec le profond travail de réflexion épistémologique qui le précède.

Mathématiques et Astronomie.	3,000 colonnes.	219,000 lignes.
Physique et Chimie	4,500 —	328,500 —
Industrie.	3,000 —	219.000 —
Sciences naturelles	4,500 —	328,500 —
Médecine	3,600 —	262,800 —
Philosophie	3,000 —	219,000 —
Histoire et Géographie	9,000 —	657,000 —
Politique, Économie politique, Administration et Finances	4,500 —	328,500 —
Droit et Jurisprudence	4,500 —	328,500 —
A reporter. . .	39,600 colonnes.	2,890,800 lignes.

(a) Première partie du tableau dans *LGE*, T1, p.XI

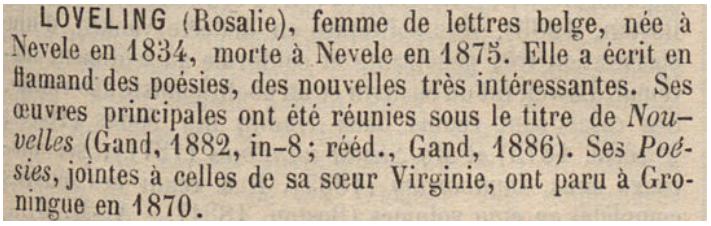
Report. . . .	39,600 colonnes.	2,890,800 lignes.
Littérature	6,000 —	438,000 —
Beaux-Arts et Archéologie.	4,500 —	328,500 —
Linguistique et Philologie	2,000 —	146,000 —
Art militaire et Marine	1,500 —	109,500 —
Théologie.	1,000 —	73,000 —
Matières diverses et Figures	5,400 —	394,200 —
	60,000 colonnes.	4,380,000 lignes.

(b) Deuxième partie du tableau dans *LGE*, T1, p.XII

Figure 3.6 – Tableau de répartition du volume de texte prévisionnel de *LGE* entre les différents domaines de connaissance

Les articles de *LGE* comportent parfois des désignants marqués typographiquement – la figure 3.1b à la page 48 en présentait déjà un exemple. Ils semblent toutefois plus rares que dans l'*EDdA* : là où environ 70% des articles de l'*EDdA* présentaient un désignant, cette proportion est (probablement très) inférieure à 45% dans *LGE*. Il n'est pas encore possible d'accéder de manière fiable aux désignants dans cette encyclopédie, mais cette borne supérieure a été obtenue de la manière suivante. Sur les 134 820 articles identifiés (voir la fin de la sous-section 3.1.3 p.61), seuls 76 786 possèdent des parenthèses sur leur première ligne ; mais pour 16 024 d'entre eux au moins ces parenthèses contiennent un renvoi. Il y a bien, en toute rigueur, quelques articles possédant à la fois un désignant et un renvoi comme l'article GELOCUS (La Grande Encyclopédie, T18, p.699) mais la conjonction des deux est assez rare pour ne pas perturber ce chiffre donné sans décimales et constituant de toute façon un majorant grossier. Les 60 762 articles restants sont à l'origine de ce chiffre de 45%. En réalité, dans *LGE* le dispositif typographique des désignants n'est pas utilisé seulement pour attribuer un domaine de connaissance. Une particularité majeure de cette œuvre réside dans la présence de no-

tices biographiques, exclues pour des raisons éditoriales des pages de l'*EDdA*. Dans les articles de biographie, le prénom de la personne suit son nom de famille qui sert de vedette à l'article, accolé entre parenthèses en tout début d'article exactement comme les désignants, ce que montre la figure 3.7 où le prénom «Rosalie», entre parenthèses, suit le nom de famille «LOVELING». Difficile donc de distinguer les deux sans posséder une liste exacte ou bien des prénoms apparaissants ou bien des domaines utilisés (car les désignants ne suivent pas les 14 domaines présentés à la figure 3.6 qui ne fournissent qu'un cadre catégorique assez vaste). Un rapide parcours de la liste de ces 60 762 mots obtenus en position de désignant suffit pour se rendre compte que les prénoms sont extrêmement fréquents, sans doute majoritaires sur les noms de domaines abréviés. Un filtrage sur le caractère '.' (caractéristique des désignants abréviés et normalement absent des noms de personnes sauf certains titres ou erreurs d'OCR) ne ramène en effet que 12 062 articles : il y a sans doute légèrement plus de vrais désignants au total, mais même une proportion de 50% par rapport aux prénoms paraît exagérément élevée (même en admettant cette proportion, il y aurait alors $0.5 \times 60\,762 = 30\,381$ articles avec un désignant, soit seulement 22.5% de l'ensemble des articles). Il y a donc un net recul de l'emploi des désignants depuis l'*EDdA*.



LOVELING (Rosalie), femme de lettres belge, née à Nevele en 1834, morte à Nevele en 1875. Elle a écrit en flamand des poésies, des nouvelles très intéressantes. Ses œuvres principales ont été réunies sous le titre de *Nouvelles* (Gand, 1882, in-8; rééd., Gand, 1886). Ses *Poésies*, jointes à celles de sa sœur Virginie, ont paru à Groningue en 1870.

Figure 3.7 – Article LOVELING dans *LGE*, T22, p.699

En revanche, un autre mécanisme est à l'œuvre pour situer le texte de *LGE* dans un système de domaines : les articles les plus longs font l'objet d'une construction complexe en sections et sous-sections (ce thème est également développé dans la section 3.2.1) et au premier niveau ces sections permettent de traiter la notion désignée par la vedette suivant ses emplois dans différents domaines. Il n'est ainsi pas rare de trouver des articles où les titres des sections feraient des désignants tout à fait acceptables comme par exemple l'article CHEMIN (La Grande Encyclopédie, T10, p.1025) dont un extrait est reproduit à la figure 3.8. Sa section III. s'intitule «Droit» et ne contient qu'un renvoi à l'article VOIRIE, la section IV. «Marine» ne comporte qu'un paragraphe (ces deux exemples montre que le découpage en section ne correspond pas nécessairement à un besoin de répartition harmonieuse du volume de texte mais bien à une logique structurelle propre) et la section V., «Art militaire» semble bien plus longue puisqu'elle s'ouvre sur une sous-section (non-numérotée) «Chemin couvert». Cette approche diffère de celle suivie dans l'*EDdA* où les termes ayant une existence dans plusieurs domaines faisaient fréquemment l'objet de plusieurs entrées distinctes – avec donc une répétition de la vedette à chacune – que seuls les désignants différenciaient, ce qu'illustre la figure 3.9. De même que certains «désignants» typographiques sont en réalité des prénoms, il n'est pas aisé de relier systématiquement ces titres de sections à des domaines de connaissance. Ainsi, les I. et II. du même article CHEMIN discuté plus haut sont respectivement intitulés Voirie et Administration ; le terme «voirie» ne renvoie pas de manière évidente à une des grandes catégories présentées à la figure 3.6, ou alors sans doute à la même catégorie qu'«administration», montrant que la division en sections ne constitue pas une partition en domaines de l'article. Certains articles tels ARGIOPE (La Grande Encyclopédie, T3, p.868) possèdent des sections qui ne sont pas du tout liées à un domaine de connaissance mais désambiguïsent un terme qui a été adopté aussi bien pour parler d'une araignée (section I.) que d'un mollusque (section II.).

règle n'ait pas été suivie d'une manière plus générale (V. Canal, Rivière navigable). L. Schmit.

III. Droit (V. Voirie).

IV. Marine. — Le chemin est l'espace parcouru par le navire, ou mieux, le nombre de milles en ligne droite parcourus par un navire, d'un midi à l'autre, mesuré sur la ligne qui joint ces deux points. On ne tient pas compte, dans l'évaluation de cette quantité, des louvoyages que peut faire un navire à voiles, quand le vent souffle du point où il se rend. On disait autrefois : nous avons fait 50 milles en bonne route, pour exprimer que le navire s'était avancé de 50 milles dans la direction à parcourir. A la mer, on évalue le chemin effectué à l'aide du *loch* (V. ce mot). Quand on navigue le long d'une côte, on opère par relèvements successifs.

V. Art militaire. — Chemin couvert. — Communication à ciel ouvert établie au-dessus et le long des contrescarpes d'une enceinte de fortification permanente ; un parapet

Figure 3.8 – Extrait de l'Article CHEMIN dans *LGE*, T10, p.1025

Air, (*Théol.*) L'air eſt ſouvent déſigné dans l'Ecriture ſous le nom de *ciel* ; *les oiſeaux du ciel* pour *les*

(a) Article AIR en Théologie dans l'*EDdA*, T10, p.236

Air ſe dit *en Peinture* de l'impreſſion que fait un tableau, à la vûe duquel on ſemble réellement reſpirer l'*air* qui regne dans la nature ſuivant les diffé-

(b) Article AIR en Peinture dans l'*EDdA*, T10, p.237

Air *en Muſique*, eſt proprement le chant qu'on adapte aux paroles d'une chanſon ou d'une petite piece de Poëſie propre à être chantée ; & par exten-

(c) Article AIR en Musique dans l'*EDdA*, T10, p.237

Figure 3.9 – Trois entrées de l'*EDdA* sur la vedette AIR

La place de la Géographie

Parmi de ces domaines, la Géographie occupe une place particulière. Elle semble du fait de ses très grandes applications pratiques – situer les villes, permettre les voyages et donc le commerce – relever d'une sorte d'évidence qui fait que sa place aux côtés des autres sciences est discutée au XVIII^{ème} siècle (voir la section 1.1 p.8). Le choix revendiqué par les éditeurs de l'*EDdA* de l'inclure dans ses pages s'impose évidemment par la suite et cette place n'est plus discutée dans *LGE*. Il n'en demeure pas moins troublant que la Géographie soit omise dans la première énumération de domaines au début de la préface de l'œuvre.

Dans le «Systême» des encyclopédistes (voir figure 3.4), la Géographie est présentée aux côté de l'*Uranographie* et de l'*Hydrographie* sous la *Cosmographie* au sein des sciences mathématiques. À un siècle où les techniques de mesure issues de la géométrie permettent des grandes avancées métrologiques telle que la détermination de la position des méridiens et de leur mesure, ce choix peut paraître relever de réalités pratiques de l'époque. Il renvoie pourtant à des considérations plus théoriques pour inscrire l'*EDdA* dans la filiation de Bacon et de Locke (Laboulais, 2003, p. 3). Une autre approche, celle de Robert de Vaugondy – géographe ordinaire du roi et contributeur de l'*EDdA* – rend compte d'une Géographie plus diverse comme le fait également remarquer Laboulais (2003, p. 2). À la fin de l'article GÉOGRAPHIE (L'Encyclopédie, T7, p.608) dont il est l'auteur, il propose de mettre la géographie astronomique à part, reconnaissant sa dette envers l'Astronomie et partitionne le reste de la Géographie en *naturelle*, *historique*, *civile* ou *politique*, *sacrée*, *ecclésiastique* et enfin *physique*. Cette division est un pas certain vers la position moderne de la Géographie au sein des Sciences Humaines et Sociales, et sa relation privilégiée avec l'Histoire. D'ailleurs, dès le préambule de l'*EDdA*, un paragraphe associe *Chronologie* et *Géographie*, dressant un parallèle entre une «science des tems» (sic) et une «des lieux» mais les encyclopédistes n'identifient pas *Chronologie* et *Histoire*, puisqu'ils font de celle-ci la mère de celle-là, attribuant en parallèle la même relation de parenté à la *Géographie* et à l'*Astronomie* respectivement. Laboulais (2003, p. 5) montre pourtant un autre lien entre les deux disciplines, mis en avant par un autre contributeur de l'*EDdA*, Paris de Meyzieu, pour des motifs pédagogiques : à l'article ÉCOLE MILITAIRE (L'Encyclopédie, T5, p.311), il présente la Géographie comme un prérequis à l'Histoire, dont elle servirait à situer les grands événements. En ce sens, il déroge à la classification du «Systême figuré» pour sortir la Géographie de la branche de l'arbre consacrée à la raison et la placer sur celle de la mémoire. Dans *LGE*, le rapprochement est consommé puisque les deux disciplines sont associées et leurs articles décomptés ensemble dans le tableau des répartition des articles entre différents domaines présenté à la figure 3.6.

Au travers des analyses présentées dans cette section, la notion de domaine de connaissance apparaît donc bien comme fondamentale. Elle joue un rôle central dans la structuration des encyclopédies et se matérialise sous plusieurs formes plus ou moins rigides, des désignants aux simples tournures de phrases en passant par des titres de section. La Géographie apparaît comme un prérequis à plusieurs autres disciplines, ce qui lui donne une place plutôt avantageuse dans l'ordre d'«enchaînement» des connaissances que les encyclopédistes – tant au XVIII^{ème} qu'au XIX^{ème} siècle – s'efforcent de saisir. Elle semble occuper en quelque sorte un rôle de pivot en permettant l'articulation de sciences aussi diverses que les Mathématiques et l'Histoire. Cette position spécifique constitue un première confirmation de la pertinence du choix de singulariser cette discipline au sein des encyclopédies dans les présents travaux.

3.1.3 Prétraitements

Après ce premier tour d'horizon de la matière textuelle disponible dans le corpus, il est temps d'évoquer quelques étapes de traitements nécessaires pour pouvoir exploiter ces données.

L'Encyclopédie de Diderot et d'Alembert

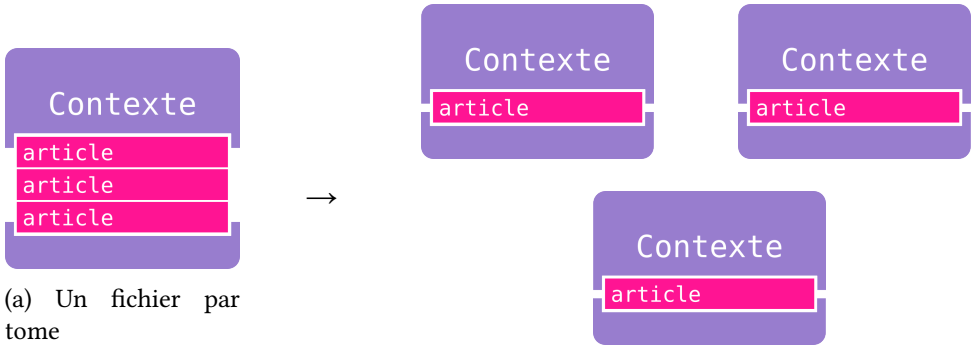
L'unité d'étude de cette thèse est l'article. En effet, il s'agit de la plus grande unité pensée par les auteurs d'un dictionnaire ou d'une encyclopédie pour être lue de manière continue du début à la fin. Les ouvrages lexicographiques ne sont pas faits pour une lecture linéaire de la lettre A à la lettre Z, ils constituent une collection de textes (voir la section 2.2) auxquels on peut se référer indépendamment : les articles. C'est au niveau des articles qu'interviennent les différents appareils de navigation de l'œuvre : renvois pour des parcours chaînés, désignants pour des parcours thématiques. Or, les fichiers livrés par l'ARTFL représentent chacun un tome. La première tâche à conduire sur ces données est donc la segmentation des tomes en articles.

Mais l'utilisation d'outils habituels pour manipuler des arbres XML est gênée par le fait que l'encodage des fichiers transmis n'est pas tout à fait valide. Certains éléments comme `sc`, `blockquote` ou `page` sont utilisés alors qu'ils n'existent pas dans le schéma TEI. Plus préoccupant, d'autres ne sont pas complets : plusieurs éléments `<index/>` utilisés pour représenter les métadonnées des articles au moyen de leurs attributs (sans contenus donc, ils sont censés être autofermants) n'ont pas leur `/` final, ce qui syntaxiquement signifie qu'ils ne se ferment jamais et englobent tout le reste du fichier après eux. Enfin, certains sont tout simplement mal formés : on trouve des balises pourvues d'attributs mais pas d'un tag (comme `<XREEF="Incorrupticoles">`, sans `<` ouvrant avant leur tag, ainsi que quelques erreurs dans les entités XML. Mais le problème le plus spectaculaire réside dans les quelques occurrences de `&` qui représenterait en XML le caractère '&' (esperluette), si seulement son dernier caractère était un ';' (point virgule) : il s'agit malheureusement en réalité du caractère UTF-8 `U+037e` utilisé en grec pour marquer l'interrogation à la place du '?', quasi identique visuellement, qui prévaut en français et dans de nombreuses langues occidentales. L'irrégularité de ces erreurs et en particulier l'homographie en jeu dans la dernière incite à penser que l'encodage de ces fichiers a été réalisé par des opérateurs humains.

Pour corriger ces erreurs ainsi que pour procéder à la division des tomes proprement dite tout en extrayant les métadonnées des articles, quelques scripts ont été écrits. Ils constituent la suite `EDdA-Clinic` [13] dont le code source est librement accessible sous license BSD-3 et hébergé sur l'instance `gitlab` d'Huma-Num. À l'aide des outils de cette suite, la division des tomes est réalisée selon la requête XPath `/TEI/text/body/div1` dont le sens est le suivant : les articles correspondent aux éléments XML `<div1/>` placés immédiatement sous le `<body/>` qui est sous le `<text/>`, commun à tout un tome de l'encyclopédie, et l'ensemble est contenu dans une balise `<TEI/>` conformément à la spécification du consortium TEI.

Le contexte de la requête, c'est-à-dire le reste de l'arbre traversé pour arriver au nœud `<body/>` contenant tous les `<div1/>` est conservé et réappliqué autour de chaque article, de manière à ce que chaque fichier d'article individuel conserve l'ensemble des métadonnées présentes sur le tome dont il est issu, comme l'illustre la figure 3.10. À l'issue de cette phase, le texte de l'*EDdA* se répartit en 74 198 fichiers. Il est à noter que les tomes 1 à 8 comportent un frontispice qui rappelle le titre de l'œuvre entre les avertissements des auteurs et le début des articles. Dans les fichiers de l'ARTFL ces frontispices sont également encodés par une balise `div1` comme les articles. Ces 8 frontispices deviennent donc 8 fichiers, mais le choix d'encodage très différent fait par l'ARTFL pour représenter ces parties qui ne sont pas à proprement parler textuelles comme peuvent l'être le discours préliminaire ou les avertissements des différents tomes empêche d'en récupérer les informations habituellement associées aux articles (auteur, domaine de connaissance...). Par conséquent, ils se retrouvent absents des fichiers de métadonnées et donc écartés dans les présents travaux, du fait de l'emploi systématique des métadonnées pour manipuler les textes. On peut donc considérer qu'il y a plutôt 74 190 articles, et des trous dans la numérotation (pas d'article 4 dans les tomes 1,2,4,5,6 et 7, pas d'article 7 dans le tome 3, pas d'article 3 dans le tome 8).

13. https://gitlab.huma-num.fr/alicebrenon/eddaclinic



(a) Un fichier par tome

(b) Un fichier par article

Figure 3.10 – Le contexte XML est dupliqué pour chaque article

La Grande Encyclopédie

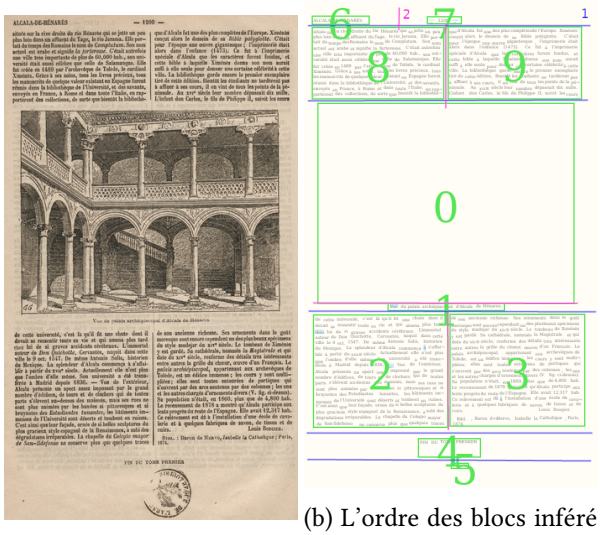
La publication de la première version numérique de *LGE* réalisée à l'issue du projet DISCO-LGE [14] repose en grande partie sur un logiciel développé spécifiquement dans le but de traiter les données fournies par la BnF au format XML-ALTO (voir la section 2.1.1). Ce logiciel, nommé `soprano` [15], permet d'obtenir des articles à partir du contenu des pages. Cette opération complexe fait intervenir trois étapes majeures. Comme l'ordre de lecture détecté par l'OLR appliqué en même temps que l'OCR est parfois erroné, un réordonnancement des blocs est d'abord appliqué. Puis, en fonction de leur position sur la page et de leur contenu, les blocs sont étiquetés pour distinguer leur rôle et différencier le corps du texte constituant les articles des autres éléments qui apparaissent fréquemment dans une page d'encyclopédie et sont décrits à la sous-section 3.1.2 p.47 (en particulier, les éléments identifiés en bleu, vert et mauve sur la figure 3.2b). Enfin le flot linéaire de texte est déduit des deux étapes précédentes et une opération de segmentation tâche de repérer les débuts d'articles pour pouvoir les répartir dans des fichiers distincts et associer à chacun la liste des métadonnées correspondantes.

Avant de pouvoir effectuer tous ces traitements, la première fonctionnalité de `soprano` est d'appliquer un filtrage pour enlever de la sortie des éléments ALTO repérés manuellement : le plus souvent des «scories», faux-positifs de l'OCR qui a cru reconnaître un caractère dans une tâche sur le papier ou dans une courbe d'une des gravures, mais aussi des données tabulaires, des équations ou des formules chimiques, abondantes dans les pages de *LGE* et susceptibles de gêner les analyses linguistiques faites sur le texte, notamment l'annotation en syntaxe. Le repérage manuel a été fait à l'aide du logiciel `chaoui` [16], également développé dans le cadre du projet DISCO-LGE. Cet outil complémentaire de `soprano` permet aussi de visualiser la structure et l'ordre des blocs tels qu'ils sont représentés par les fichiers ALTO. La figure 3.11 montre ainsi la dernière page du premier tome de *LGE* entièrement dédiée à la ville d'Alcala-de-Hénarès (avec une gravure pleine page représentant le palais archiépiscopal visible sur la figure 3.11a) et l'ordre dans lequel l'OLR appliqué à cette page a placé les différents blocs qu'il a identifiés sur la page (figure 3.11b). Les numéros représentent l'ordre des éléments XML correspondant dans le fichier ALTO qui encode la page : ils se trouvent effectivement au bon endroit sur la page mais l'ordre des numéros ne correspond pas à l'ordre de lecture. Le problème en jeu ici réside dans la difficulté à concilier un objet bidimensionnel – la page – et un objet unidimensionnel – le texte.

En effet une page possède une hauteur et une largeur ; les signes imprimés dessus peuvent se positionner les uns par rapport aux autres selon un axe horizontal (l'un est à la gauche de l'autre) mais aussi selon un axe vertical (l'un est au-dessous de l'autre). Le texte quant à lui est une abstraction

14. https://www.collexpersee.eu/projet/disco-lge/
15. https://gitlab.inria.fr/abrenon/soprano
16. https://gitlab.huma-num.fr/alicebrenon/chaoui

intrinsèquement linéaire, au moins dans ses réalisations concrètes (avant toute analyse syntaxique ou sémantique) : il est formé d'un flot de morphèmes constituant un flot de mots qui s'assemblent en un flot de phrases (pouvant à leur tour éventuellement former un flot de paragraphes, sections, chapitres suivant la complexité de la production, ou d'articles dans le cadre de cette thèse). Chacune de ces unités peut seulement venir «avant» ou «après» une autre : ni «au-dessus/en-dessous» ni «à gauche/à droite» – une propriété sans doute issue du langage oral qui se déploie dans la seule dimension temporelle : au-delà des différences structurelles propres qui peuvent distinguer l'oral de l'écrit, toute production écrite peut donner lieu à une production orale (en la lisant) et doit donc s'inscrire dans cette unidimensionnalité. À l'échelle d'un bloc de texte, ces notions se traduisent par la lecture des lignes de haut en bas et, à l'intérieur d'une ligne, des mots de gauche à droite. Mais entre blocs, l'évidence d'ordre s'estompe : les pages comme celle présentée à la figure 3.11 barrée d'une figure se lisent de gauche à droite en commençant par les deux colonnes au-dessus de la figure (blocs 8 puis 9) avant de poursuivre sous la figure (blocs 2 puis 3). Mais dans des pages sans figure pleine page, il faut lire tous les blocs de la colonne de gauche de haut en bas puis continuer en haut de la colonne de droite. Cette règle est assez simple à implémenter mais la figure 3.11b montre que l'OLR utilisé par la BnF ne l'applique pas. À cette règle simple s'ajoute la nécessité de séparer tous les éléments péritextes (voir sous-section 3.1.2) parfois intégrés à tort dans un des blocs de texte. Le meilleur modèle du texte d'une encyclopédie est un flot linéaire de texte entouré de satellites qui existent hors de lui (les éléments péritextes ne sont pas «à lire» à un moment particulier du texte).



(a) Une photo de la page

(b) L'ordre des blocs inféré par l'OLR pour cette page

Figure 3.11 – Dernière page du premier tome de *LGE*

La valeur principale de `soprano` réside cependant dans sa capacité à détecter les débuts d'article afin de séparer les entrées. C'est cette fonctionnalité qui a fait l'objet des améliorations les plus importantes au cours de cette thèse par rapport à la version obtenue à l'issue du projet DISCO-LGE. Une fois de plus, le problème abstrait est excessivement simple : les débuts d'articles correspondent aux vedettes, qui sont repérables typographiquement par leur indentation et leur casse haute. Malheureusement, la présence de scories dans les marges ainsi que la très légère inclinaison des colonnes gênent la mesure précise de l'indentation des lignes. Il faut des conditions très laxes, que vérifient parfois des éléments qui ne sont pas des débuts de paragraphe (des signatures de contributeur, des titres dans un tableau ou même une légende de figure intégrée à tort dans un bloc de texte par l'OLR et qui n'a pas été repérée et corrigée par `soprano`). Les majuscules ne permettent guère de gagner en précision : les erreurs de l'OCR peuvent les transformer arbitrairement en chiffres ou en signes de ponctuation, ce qui empêche également de pouvoir bénéficier de l'information capitale qu'elles

doivent constituer une suite croissante selon l'ordre alphabétique. À la place, il est nécessaire de procéder avec la plus grande tolérance, en vérifiant des ressemblances graphiques entre caractères pour ne pas rejeter des vedettes et sous-segmenter, c'est-à-dire laisser au sein du même fichier plusieurs articles. Devant la difficulté à vérifier l'ordre alphabétique des entrées, une règle simple mais pragmatique a été ajoutée depuis la version issue du projet DISCO-LGE : `soprano` permet de préciser la vedette du premier et du dernier article et ne vérifie que l'appartenance des vedettes potentielles à cet intervalle pour prendre la décision de créer un nouvel article. Avec cette régle le nombre précédent de 229 475 articles tombe à 134 820, nombre d'articles retenu pour les présents travaux. Un rapide échantillonnage montre qu'il reste des problèmes de sous-segmentation (des articles groupés au sein du même fichier), vraisemblablement plus que de sur-segmentation (un article éclaté en plusieurs fichiers). Les auteurs de *LGE* en mentionnent 200 000 quand ils écrivent l'article ENCYCLOPÉDIE (La Grande Encyclopédie, T15, p.1014), mais la fin du projet est encore assez lointaine alors et quand on tient compte des difficultés financières qu'a connu le projet à sa toute fin (Jacquet-Pfau, 2015, p. 90), il paraît possible que cette prévision n'ait pas été tout à fait atteinte. Le nombre réel d'articles dans *LGE* demeure donc inconnu, mais pourrait être de l'ordre de 150 000 à 200 000 articles, soit environ du double au triple de celui de l'*EDdA*. Il faut pourtant poursuivre, car la quête d'une qualité parfaite est sans fin et des erreurs subsistent toujours, même dans les sources les plus fiables. Ainsi la version de l'*EDdA* de l'ARTFL contient aussi au moins un cas de sous-segmentation, l'entrée pour MÉLER UN CHEVAL (L'Encyclopédie, T10, p.313) contient également un article de *Géographie ancienne*, MÉLÈS qui décrit une rivière d'Asie.

La mise en regard de ces deux textes produit un corpus d'étude de 209 010 articles pour un total de 78 722 979 tokens. Cependant, les techniques présentées ci-dessus ne constituent que la première étape pour accéder au texte. Son encodage a nécessité une étude séparée pour déterminer les éléments appropriés à la représentation de ses constituants et à sa structuration. Cette étude, qui n'a concerné que *LGE* puisque l'*EDdA* était déjà disponible en format XML, fait l'objet de la section suivante 3.2.

3.2 Encodage des fichiers

3.2.1 Des différences entre Encyclopédies et Dictionnaires

Les tomes obtenus de l'ARTFL sont encodés dans le format XML-TEI (voir la section 2.1). Pour *LGE*, puisque la version numérique a été créée à partir d'une ré-OCRisation de l'œuvre, il a été nécessaire d'encoder les articles. Il est naturel, à la suite de l'ARTFL de diffuser le texte de *LGE* encodé dans le format XML-TEI (voir la section 2.1 p.18). Mais ce choix ne suffit pas à définir directement un encodage car les recommandations publiées contiennent un très grand nombre de balises aux sémantiques précises et aux règles de composition complexes. Le module *dictionaries* identifié à la section 2.1.2 n'est ainsi pas utilisé par l'ARTFL dans son encodage de l'*EDdA*. Il apparaît naturel d'utiliser le même schéma d'encodage mais la sous-section 3.1.3 vient de montrer qu'il n'est pas sans défauts, c'est pourquoi une étude systématique a été menée pour choisir un schéma d'encodage pour les articles de *LGE*.

Ainsi, il y a par exemple plusieurs balises utilisées pour délimiter une unité structurelle de texte. La plus générale est `div` mais elle n'a été ajoutée qu'à partir de l'itération 0.6 de la version P5 des recommandations de la TEI. Historiquement il y avait une collection de variantes de cette balise intégrant une notion de profondeur fixe : `div1`, `div2`, etc. jusqu'à `div7`. Les bonnes pratiques recommandent d'utiliser soit seulement des `div` (sans numéro) à toutes les profondeurs soit seulement leurs équivalents numérotés mais de ne pas mélanger les deux. Les tomes de l'*EDdA* obtenus via l'*ARTFL* utilisent encore des balises numérotées `div1` etc. mais la profondeur importante de certains articles de *LGE* est un argument contre l'emploi d'un jeu de balises dont la profondeur est limitée car avant de disposer d'une version numérique il est impossible de savoir quel article a la structure la plus profonde et

si cette profondeur est bien inférieure à 7. Cette partie, en reprenant et étendant des travaux exposés dans Brenon (2024), explore les possibilités et les écueils mis en évidence page 22 à la section 2.1.3 pour définir un schéma d'encodage approprié à *LGE*.

Outils pour une exploration systématique

La version de la TEI utilisée pour cette étude (4.3.0, correspondant à la révision b4f72b1ff) contient 590 éléments [17], documentés sur le site web du consortium. Avec une moyenne de presque 80 éléments pouvant être inclus sous un élément donné (79.91), parcourir manuellement les pages de la documentation peut se révéler très long et difficile à faire de manière exhaustive à cause de l'explosion combinatoire du nombre de chemins possibles à partir d'un nœud à chaque étape supplémentaire.

Aussi, pour aller plus loin que les remarques préliminaires de la section 2.1.3 sur les difficultés prévisibles pour encoder une encyclopédie à l'aide du module *dictionaries*, le problème peut être avantageusement transformé en un graphe pour bénéficier des nombreux résultats existants sur ces objets. L'ensemble des éléments de la TEI est ainsi représenté par un graphe orienté dont les nœuds sont les éléments XML disponibles en TEI et dont les arêtes représentent la possibilité d'inclure un élément correspondant à leur nœud de destination directement sous un élément correspondant à leur nœud source. Il est alors possible d'utiliser des méthodes classiques et éprouvées comme l'algorithme de Dijkstra (1959) qui permet de calculer le plus court chemin entre deux nœuds d'un graphe et de savoir s'ils ne sont pas connectés pour tester de manière systématique toutes les manières dont les éléments peuvent être combinés sans risque d'oubli.

Il n'y a évidemment aucune garantie que le plus court chemin ait un sens en termes de schéma d'encodage mais l'algorithme fournit au moins un moyen efficace de vérifier s'il est seulement possible d'inclure un élément sous un autre et donne une longueur minimale nécessaire pour réaliser l'inclusion. Cette méthode trouve donc tout son intérêt pour obtenir des résultats négatifs. L'absence totale de chemin entre deux nœuds prouve l'impossibilité d'une inclusion entre eux ; une profondeur anormalement élevée est un indice sérieux qu'une inclusion ne devrait pas être possible et n'est qu'un accident émergeant des règles du schéma sans sens réel en termes d'encodage.

Le graphe du schéma XML-TEI est donc défini formellement de la manière suivante. Un nœud est créé pour chacun des 590 éléments présents dans les spécifications. Puis une arête est créée d'un nœud source `a` vers un nœud destination `b` si et seulement si le schéma stipule qu'un élément de la classe représentée par `b` peut être contenu directement dans un élément de celle représentée par `a`. De manière équivalente, il est possible de dire que les arêtes du graphe représentent la relation ”admet pour enfant direct”, c'est-à-dire que – toujours sous l'hypothèse d'une arête d'un nœud `a` vers un nœud `b` dans le graphe – l'arbre XML de la figure 3.12 est valide selon le schéma.

```
<a>
    <b/>
</a>
```

Figure 3.12 – Traduction sous forme d'arbre XML de l'existence d'une arête `a` → `b` sur le graphe

Le mot «élément» est ici employé dans le même sens que dans la documentation de la TEI pour désigner l'appareil conceptuel caractérisé par un nom de balise tel que `p` ou `div` pris pour la classe de tous les éléments ayant ce nom, et non pas une de ses instances particulières qui pourrait se trouver dans un document précis. La figure 3.13 en montrant le graphe obtenu par cette transformation

17. la section 2.1.2 p.20 fait état de 587 éléments : depuis la version 4.3.0, 6 éléments ont été retirés des spécifications, et 3 nouveaux ajoutés : `<eventName/>`, `<gender/>` et `<post/>`. Ces éléments sans lien avec les encyclopédies n'appartiennent pas au module *dictionaries* et ne remettent donc pas en cause les résultats présentés.

restreint au seul module *dictionaries* laisse imaginer la complexité de la totalité des spécifications du format XML-TEI.

Définitions Une des premières notions définies en théorie de graphe est celle de chemin entre deux nœuds, obtenue en considérant la fermeture transitive de la relation binaire représentée par les arêtes : formellement, on dit que deux nœuds `a` et `b` sont reliés par un chemin s'il y a une arête entre les deux ou bien s'il y a une arête de `a` à un troisième nœud `c` tel que `c` et `b` sont reliés par un chemin. Étant donnée la relation choisie pour définir le graphe de la TEI, cette notion de chemin traduit la possibilité d'inclure un élément sous un autre directement ou indirectement.

Les nœuds visités sur le graphe le long d'un chemin représentent les éléments XML intermédiaires qui sont nécessaires pour produire un arbre XML valide suivant le schéma TEI. Au vu de la sémantique descendante associée à ces arbres, il est pertinent d'appeler la longueur d'un tel chemin d'inclusion sa *profondeur*. La possibilité d'inclure un élément directement sous lui-même correspond à la présence de boucle sur le graphe (c'est-à-dire une arête d'un nœud vers lui-même) et trouve une illustration dans la TEI avec l'élément `<abbr/>` : une `<abbr/>` (abréviation) peut directement en contenir une autre.

En généralisant l'idée à des chemins d'inclusion de n'importe quelle longueur supérieure à 1 on obtient la notion de *cycle*. L'élément `<address/>` en fournit un exemple : bien qu'une `<address/>` ne puisse pas en contenir directement une autre, elle peut en revanche contenir un `<geogName/>` qui, à son tour, peut contenir une nouvelle `<address/>`. D'un point de vue de théorie des graphes, on dira que cet élément admet un cycle d'inclusion de longueur 2.

Applications Intuitivement, la pertinence du plus court chemin entre deux nœuds diminue quand la distance entre deux éléments dans le graphe augmente, mais elle peut s'avérer utile pour fournir des suggestions lorsque l'on découvre les modules de la TEI. En effet la documentation présente une vision locale des nœuds du graphe car la page d'un élément décrit les éléments admissibles immédiatement au-dessus ou en-dessous du premier mais ne peut donner que quelques exemples d'arbres XML qui le contiennent (la combinatoire rend impossible l'intégration d'une liste exhaustive d'exemples d'utilisation même sur seulement deux niveaux). Si ceux fournis ne correspondent pas à la structure que l'on souhaite encoder, il peut donc être difficile de comprendre comment un élément peut être placé sous un autre.

C'est dans ce type de situation que l'algorithme du plus court chemin permet de «découvrir» que pour inclure une `<pos/>` (pour représenter une «partie de discours» ou POS en anglais) dans une `<entry/>` représentant un article de dictionnaire, il faut passer par une balise comme `<form/>`, `<cit/>`, `<dictScrap/>` ou `<gramGrp/>` mais que cela ne peut pas se faire directement. En effet, le parcours exhaustif de tous les chemins reliant les deux nœuds correspondants sur le graphe ramène les chemins `entry-form-pos` et `entry-gramGrp-pos`. Il revient à la personne en charge de l'annotation de choisir le chemin le plus approprié en fonction de sa pertinence dans le contexte.

Un dernier exemple éclairant sur l'usage de cette méthode peut être trouvé en recherchant le plus court chemin d'inclusion d'un `<pos/>` sous le `<body/>` d'un document : il passe directement par un `<entryFree/>` (c'est un chemin de longueur 2, `body-entryFree-pos`) qui, à la différence d'`<entry/>` l'accepte directement sans nœud intermédiaire. Ce pourrait ne pas être le chemin attendu, selon que d'autres éléments grammaticaux dans l'article tels qu'un `<case/>` ou un `<gen/>` justifient le recours à un `<gramGrp/>`, mais en augmentant le périmètre de recherche jusqu'aux chemins de longueur 3, on trouve comme attendu le chemin passant par `<entry/>`, parmi d'autres. Pour qui découvrirait ces balises, une image commencerait à apparaître au vu de ces chemins : `<pos/>` n'est pas un élément très profond, il peut apparaître près de la «surface» des articles, dont les «points d'entrée» possibles depuis le reste du graphe sont les balises `<entry/>` et sa version

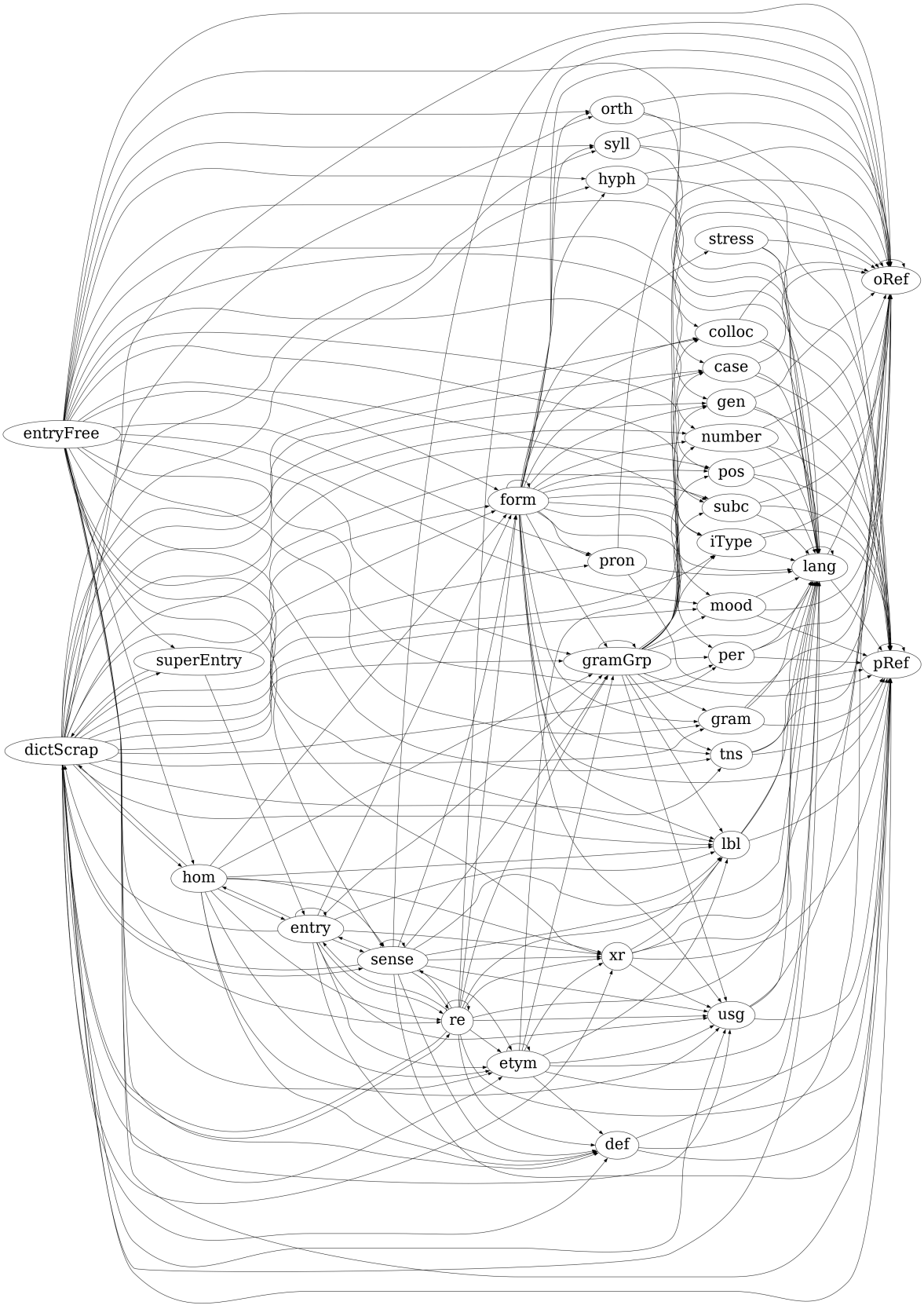
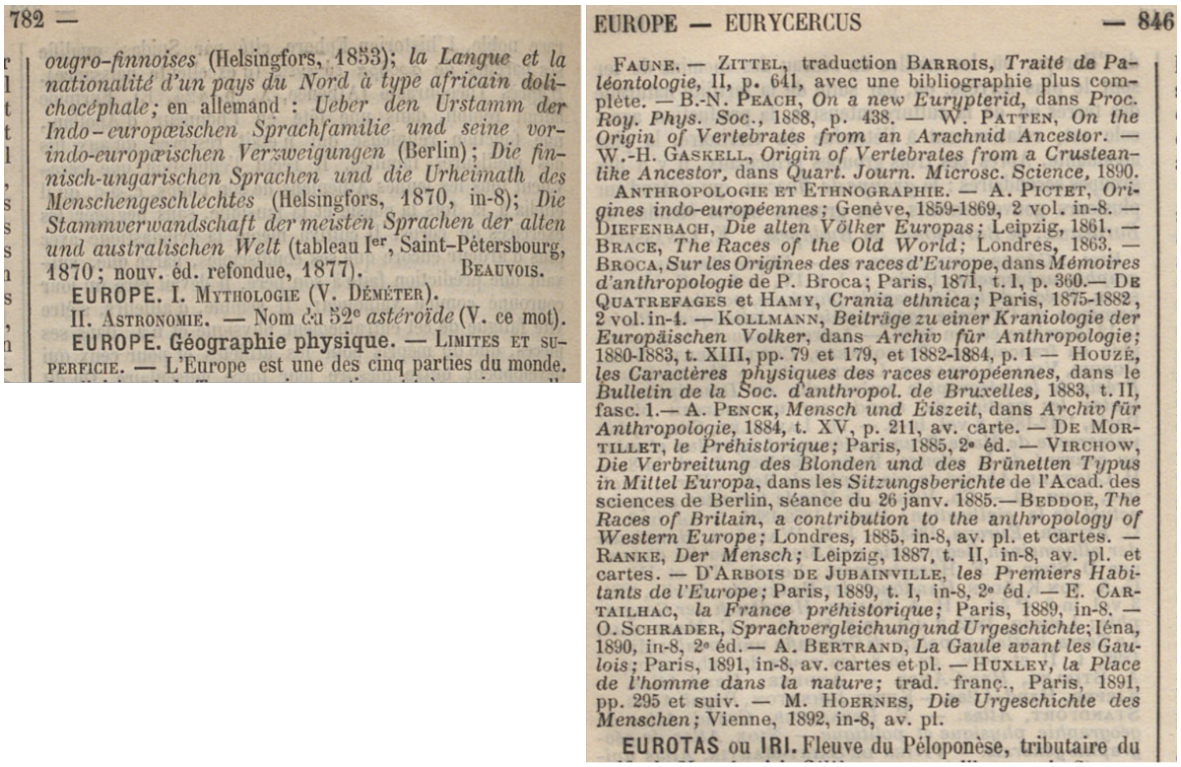


Figure 3.13 – Le sous-graphe du module *dictionaries*

moins contrainte `<entryFree/>`.

Limites du module *dictionaries*

Les encyclopédies se distinguent des dictionnaires par un certain nombre de traits comme la présence de désignants marqués typographiquement, le découpage en paragraphe rendu nécessaire par des articles parfois très longs – l'article EUROPE (La Grande Encyclopédie, T16, p.782) s'étend ainsi sur 64 pages comme l'illustre la figure 3.14 – et très structurés. Les outils d'exploration systématique de graphes mis en place ci-dessus permettent d'identifier plusieurs faiblesses dans le module *dictionaries* pour encoder les éléments caractéristiques des encyclopédies identifiés ci-dessus.


782 —

ougro-finnoises (Helsingfors, 1853); *la Langue et la nationalité d'un pays du Nord à type africain dolichocéphale;* en allemand : *Ueber den Urstamm der Indo-europæischen Sprachfamilie und seine vorindo-europæischen Verzweigungen* (Berlin); *Die finnisch-ungarischen Sprachen und die Urheimath des Menschengeschlechtes* (Helsingfors, 1870, in-8); *Die Stammverwandschaft der meisten Sprachen der alten und australischen Welt* (tableau Ier, Saint-Pétersbourg, 1870; nouv. éd. refondue, 1877). Beauvois.

EUROPE. I. Mythologie (V. Déméter).

II. Astronomie. — Nom du 52e *astéroïde* (V. ce mot).

EUROPE. Géographie physique. — Limites et superficie. — L'Europe est une des cinq parties du monde.

EUROPE — EURYCERCUS — 846

Faune. — Zittel, traduction Barrois, *Traité de Paléontologie*, II, p. 641, avec une bibliographie plus complète. — B.-N. Peach, *On a new Eurypterid*, dans *Proc. Roy. Phys. Soc.*, 1888, p. 438. — W. Patten, *On the Origin of Vertebrates from an Arachnid Ancestor.* — W.-H. Gaskell, *Origin of Vertebrates from a Crustean-like Ancestor*, dans *Quart. Journ. Microsc. Science*, 1890.

Anthropologie et Ethnographie. — A. Pictet, *Origines indo-européennes*; Genève, 1859-1869, 2 vol. in-8. — Diefenbach, *Die alten Völker Europas*; Leipzig, 1861. — Brace, *The Races of the Old World*; Londres, 1863. — Broca, *Sur les Origines des races d'Europe*, dans *Mémoires d'anthropologie* de P. Broca; Paris, 1871, t. I, p. 360.— De Quatrefages et Hamy, *Crania ethnica*; Paris, 1875-1882, 2 vol. in-4. — Kollmann, *Beiträge zu einer Kraniologie der Europäischen Volker*, dans *Archiv für Anthropologie*; 1880-1883, t. XIII, pp. 79 et 179, et 1882-1884, p. 1 — Houzé, *les Caractères physiques des races européennes*, dans le *Bulletin de la Soc. d'anthropol. de Bruxelles*, 1883, t. II, fasc. 1.— A. Penck, *Mensch und Eiszeit*, dans *Archiv für Anthropologie*, 1884, t. XV, p. 211, av. carte. — De Mortillet, *le Préhistorique*; Paris, 1885, 2e éd. — Virchow, *Die Verbreitung des Blonden und des Brünetten Typus in Mittel Europa*, dans les *Sitzungsberichte* de l'Acad. des sciences de Berlin, séance du 26 janv. 1885.— Beddoe, *The Races of Britain, a contribution to the anthropology of Western Europe*; Londres, 1885, in-8, av. pl. et cartes. — Ranke, *Der Mensch*; Leipzig, 1887, t. II, in-8, av. pl. et cartes. — D'Arbois de Jubainville, *les Premiers Habitants de l'Europe*; Paris, 1889, t. I, in-8, 2e éd. — E. Cartailhac, *la France préhistorique*; Paris, 1889, in-8. — O. Schrader, *Sprachvergleichung und Urgeschichte*; Iéna, 1890, in-8, 2e éd. — A. Bertrand, *La Gaule avant les Gaulois*; Paris, 1891, in-8, av. cartes et pl. — Huxley, *la Place de l'homme dans la nature*; trad. franç., Paris, 1891, pp. 295 et suiv. — M. Hoernes, *Die Urgeschichte des Menschen*; Vienne, 1892, in-8, av. pl.

EUROTAS ou IRI. Fleuve du Péloponèse, tributaire du

Figure 3.14 – Article EUROPE dans *LGE*, T16 s'étendant de la p.782 à la p.846

Désignants Introduits à la section 3.1.2, les désignants sont utilisés pour situer les articles dans un espace organisé de connaissances. Dans les œuvres du corpus, ils sont marqués formellement en apparaissant immédiatement après la vedette entre parenthèses et souvent abréviés. Comme cela a été montré à la figure 3.7, le même dispositif typographique est utilisé pour les entrées biographiques dans *LGE* afin de donner le prénom de la personne dont elles traitent. Cependant, tous les emplois ne se ramènent pas à l'un de ces deux cas d'usage. L'article BALAGRUS (La Grande Encyclopédie, T5, p.48) reproduit à la figure 3.15 et extrait de la même page que la figure 3.2 de la section 3.1.2 p.50 montre ainsi un exemple où le mot entre parenthèses n'est pas le prénom de la personne décrite par l'entrée mais une répétition du nom qui sert de vedette dans son alphabet grec d'origine : «BALAGRUS (Βάλαγρος)». Il existe potentiellement d'autres articles pour lequel le «désignant» typographique sert à donner encore une autre information.

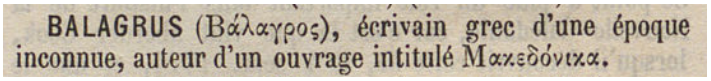

BALAGRUS (Βάλαγρος), écrivain grec d'une époque inconnue, auteur d'un ouvrage intitulé Μακεδόνικα.

Figure 3.15 – Article BALAGRUS dans *LGE*, T5, p.48

Pour représenter ces éléments (les vrais désignants et les autres qui prennent leur place), aucune solution parfaite n'émerge. Le module *dictionaries* fournit `<usg/>` dont la sémantique correspond bien à la fonction première des désignants et qui a été utilisée par exemple par Salgado et al. (2024, p. 139). En revanche, utiliser cette balise pour représenter des prénoms comme le fait *LGE* nécessiterait l'emploi des valeurs `comp` (complément typique de la vedette) ou `colloc` (collocation typique de la vedette) pour l'attribut `type` si l'on considère que le prénom d'un personnage historique représente seulement une information que l'on apprend à associer directement à son nom par un mécanisme mémoriel semi-volontaire. Cela supposerait tout de même de réussir à distinguer les différents cas de manière automatique (voir la section 3.2.2 plus loin, en particulier à partir de la page 74). En l'absence d'une liste exhaustive des domaines ou des prénoms apparaissant dans ses pages et compte tenu de la subtilité des sémantiques associées à certains objets (le fait par exemple que «Βάλαγρος» à la figure 3.15 ci-dessus ne soit pas un prénom mais l'orthographe originelle du nom en grec), cette opération semble pour le moins délicate à réaliser de manière satisfaisante et sans erreurs. Il serait également possible de ne pas chercher à distinguer les cas et d'utiliser systématiquement la valeur `hint` pour cet attribut mais en ce cas, autant employer directement une balise traduisant simplement le choix typographique plutôt qu'une sémantique précise. Hors du module, l'élément `<domain/>`, malgré son nom, ne peut se trouver que dans l'en-tête d'un document et couvre seulement le contexte social d'un texte et pas son sujet. Le nom d'`<interp/>` peut aussi prêter à confusion puisqu'il ne sert pas à étiqueter un groupe de mots comme une interprétation possible donnant un contexte (ce qu'il serait défendable d'argumenter pour un désignant qui, placé en tête d'article, oriente l'esprit de la personne qui le lit vers un sujet particulier). Au contraire, la documentation montre clairement qu'il s'agit d'un outil d'annotation pour ajouter du texte qui n'est pas présent dans le document original mais qui est le résultat d'une analyse effectuée dans le cadre de l'annotation.

Structures imbriquées Les encyclopédies se caractérisent également par des développements parfois longs et très structurés avec des sections et sous-sections imbriquées jusqu'à des profondeurs importantes. *LGE* n'échappe pas à la règle, le discours se structurant sur au moins 5 niveaux par exemple dans l'article EUROPE. Certains passages contiennent même un niveau supplémentaire sous la forme d'énumérations, comme l'illustre la figure 3.16.

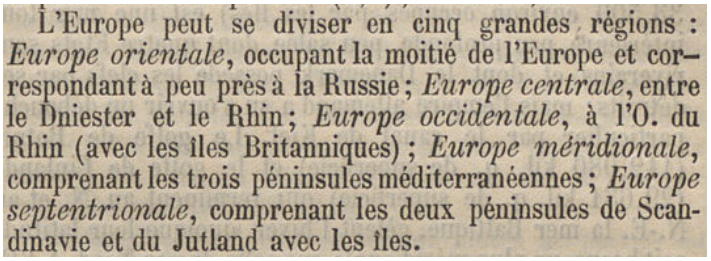
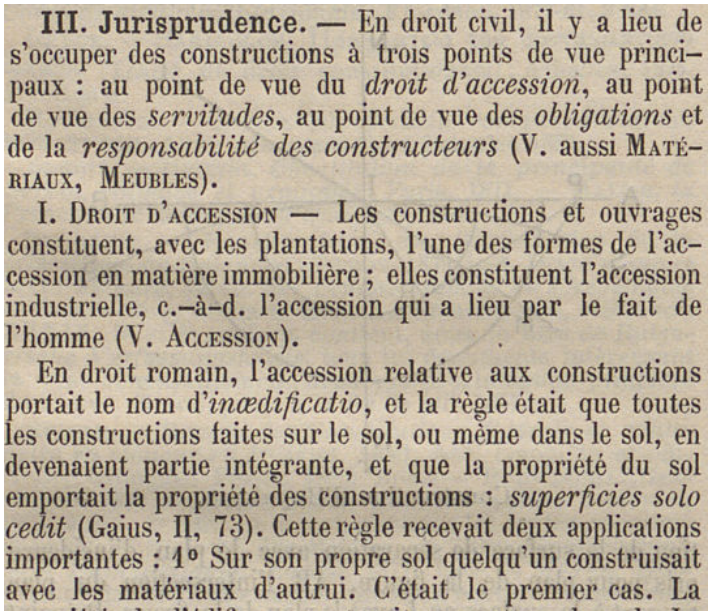

L'Europe peut se diviser en cinq grandes régions : *Europe orientale*, occupant la moitié de l'Europe et correspondant à peu près à la Russie ; *Europe centrale*, entre le Dniester et le Rhin ; *Europe occidentale*, à l'O. du Rhin (avec les îles Britanniques) ; *Europe méridionale*, comprenant les trois péninsules méditerranéennes ; *Europe septentrionale*, comprenant les deux péninsules de Scandinavie et du Jutland avec les îles.

Figure 3.16 – Énumération dans l'article EUROPE, *LGE*, T16, p.783

La figure 3.17 extraite de l'article CONSTRUCTION (La Grande Encyclopédie, T12, p.732) montre ainsi le début de la partie «III. Jurisprudence» de l'article, dans lequel s'ouvre un «I. Droit d'accession» à l'intérieur duquel on trouve un «1°». Ce dernier, bien que la typographie ne le marque que faiblement, sans paragraphe séparé, sans titre et survenant au milieu d'une phrase après des « :» constitue pourtant bien une unité structurelle du texte à la portée extra-phrastique puisqu'on voit sur la figure 3.17a cette phrase se terminer, et que le «2°» visible sur la figure 3.17b n'apparaît que plus bas sur la même page après plusieurs autres phrases. Cette question de la typographie est d'ailleurs importante puisque les paragraphes peuvent autant servir à marquer une sous-division du texte qu'à simplement aérer un passage long ou à marquer la succession de plusieurs idées comme le fait le 3ème paragraphe visible sur la figure 3.17a ou même celui de la figure 3.19. Même les listes comme celles visible à la figure 3.16 bien qu'elles ne soient pas écrites verticalement avec un élément

par ligne (pour des raisons évidentes de compacité de l'information requise par les encyclopédies) pourraient tout à fait être encodées à l'aide de la balise TEI `<list/>` dont l'attribut `@rend` inclut la valeur `inline` pour rendre compte de ce choix de présentation.

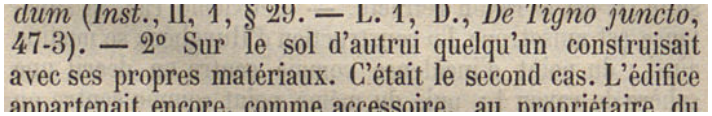


III. Jurisprudence. — En droit civil, il y a lieu de s'occuper des constructions à trois points de vue principaux : au point de vue du *droit d'accession*, au point de vue des *servitudes*, au point de vue des *obligations* et de la *responsabilité des constructeurs* (V. aussi Matériaux, Meubles).

I. Droit d'accession — Les constructions et ouvrages constituent, avec les plantations, l'une des formes de l'accession en matière immobilière ; elles constituent l'accession industrielle, c.-à-d. l'accession qui a lieu par le fait de l'homme (V. Accession).

En droit romain, l'accession relative aux constructions portait le nom d'*inædificatio*, et la règle était que toutes les constructions faites sur le sol, ou même dans le sol, en devenaient partie intégrante, et que la propriété du sol emportait la propriété des constructions : *superficies solo cedit* (Gaius, II, 73). Cette règle recevait deux applications importantes : 1° Sur son propre sol quelqu'un construisait avec les matériaux d'autrui. C'était le premier cas. La

(a) Début de la section III. Jurisprudence



dum (*Inst.*, II, 1, § 29. — L. 1, D., *De Tigno juncto*, 47-3). — 2° Sur le sol d'autrui quelqu'un construisait avec ses propres matériaux. C'était le second cas. L'édifice appartenait encore, comme accessoire, au propriétaire du

(b) Fin du 1° et début du 2° quelques lignes plus bas

Figure 3.17 – Extrait du début de la section III. Jurisprudence de l'article CONSTRUCTION dans *LGE*, T12, p.734

Or, malgré plusieurs exemples de connexions entre le module *dictionaries* et le reste de la XML-TEI, en particulier vers le module *core* (l'élément `<ref/>` en est un exemple), le module *dictionaries* apparaît dans une certaine mesure isolé d'éléments structurels importants tels que `<head/>` ou `<div/>`. En effet, en calculant par un parcours systématique tous les chemins vers ce dernier depuis les éléments `<entry/>` ou `<sense/>` dont la longueur est inférieure ou égale à 5, tous les chemins trouvés (au nombre de 8 943 et 38 649 respectivement en excluant les cycles pour éviter les doublons) contiennent soit un élément `<floatingText/>` soit un élément `<app/>`. Le premier comme son nom le suggère est utilisé pour encoder du texte qui n'a pas sa place dans le flot normal du document, par exemple dans le contexte de récits enchâssés. Les deux exemples fournis dans la documentation en ligne montrent un élément `<body/>` directement sous le `<floatingText/>`, séparant de manière nette son contenu de ce qu'il y a à l'extérieur de la balise. La finalité du deuxième, bien que son nom – abréviation d'«apparatus» – soit moins clair, est d'agréger ensemble plusieurs versions d'un même extrait par exemple pour dénoter plusieurs lectures possibles d'un groupe de mots obscur d'un manuscrit ou quand le processus d'encodage tente de rassembler une version unique d'une œuvre à partir de plusieurs sources qui diffèrent sur un passage. Dans les deux cas, il semble clair que ces éléments n'ont pas à être présents systématiquement dans l'encodage d'un article d'encyclopédique simplement pour pouvoir structurer ses sections et sous-sections. Ainsi, malgré une connectivité interne plutôt élevée, le module *dictionaries* ne parvient pas non plus à fournir de quoi représenter des structures imbriquées récursivement avec `<div/>`.

Les structures enchâssées mises en évidence guident la recherche d'éléments imbricables pour les

représenter. Plus précisément, elles imposent plusieurs contraintes sur les éléments XML à utiliser : il s'agit de trouver une paire d'éléments, le premier – représentant tout découpage en section et sous-section – devant pouvoir contenir lui-même et le deuxième élément, sur lequel la seule contrainte est d'avoir une sémantique compatible avec la représentation d'un titre de section. De plus, le premier élément doit aussi pouvoir contenir plusieurs éléments `<p/>`, la balise de référence pour encoder les paragraphes d'après la documentation XML-TEI.

En filtrant le contenu du module pour ne retenir que les éléments qui vérifient les conditions ci-dessus et peuvent être inclus sous un élément `<entry/>` il ne reste aucun candidat. Relâcher cette contrainte additionnelle en remplaçant `<entry/>` par `<entryFree/>`, un élément de sémantique similaire mais bien plus flexible pour s'adapter à des structures de dictionnaire inhabituelles n'apporte aucune amélioration à ce résultat. Ce manque de candidat impose d'alléger encore les critères de recherche sur les éléments utilisables. Il se pourrait ainsi qu'un élément intermédiaire soit nécessaire entre le niveau le plus haut de l'article (l'élément `<entry/>` dans le scénario initial) et l'élément encodant les sections et sous-sections. De même, l'imbrication récursive aussi pourrait nécessiter un élément «compagnon» pour fonctionner (l'élément représentant une sous-section s'insérant non plus directement sous celui pour une section mais sous un nouvel élément intermédiaire). En termes de théorie des graphes, il s'agit de relâcher la contrainte sur l'existence d'une boucle pour considérer aussi les cycles d'une certaine longueur assez faible (il s'agit quand même de représenter une inclusion assez directe).

En fixant leur longueur à 3, c'est-à-dire en acceptant un seul élément intermédiaire dans le chemin d'inclusion entre l'élément qui représentera les sections et lui-même représentant une sous-section, seuls 21 éléments sont possibles. Aucun d'entre eux n'apparaît cependant comme une bonne solution évidente : tous les chemins pour inclure `<p/>` depuis n'importe quel élément du module *dictionaries* contiennent une `<figure/>` (dont la sémantique montre clairement son inadéquation : une sous-section n'est pas une figure à l'intérieur du texte), ou un `<stage/>` (réservé au théâtre) ou bien un `<state/>` (traduisant une qualité temporaire d'une personne ou d'un endroit, une fois de plus sans lien avec le contexte d'utilisation requis). Les chemins d'inclusion vers `<head/>` ou `<title/>` sont aussi décevants. Et, là encore, le remplacement de `<entry/>` par `<entryFree/>` n'apporte aucune amélioration. Sans constituer une preuve absolue qu'aucun des éléments envisagés ne pourrait convenir, il n'en demeure pas moins vrai qu'aucune solution satisfaisante n'apparaît dans le module, ce qui incite à poursuivre les recherches ailleurs.

Les contraintes sont relâchées encore un peu pour inclure les éléments hors du module *dictionaries* pour représenter les sections et sous-sections, sous les mêmes conditions que précédemment, en particulier que l'ensemble représentant l'article puisse apparaître sous un élément `<entry/>`. Cette nouvelle recherche ne rapporte que 3 candidats : `<figure/>`, `<metamark/>` et `<note/>`.

Le premier ne convient évidemment pas, comme cela a déjà été remarqué précédemment. Le but de `<metamark/>` est de transcrire les marques d'éditions qui apparaissent sur un exemplaire d'une source primaire pour altérer le flot normal du texte et suggérer une lecture alternative (suppression, insertion, réordonnancement, toute note laissée à cet effet par un humain sur un livre physique). Cela n'a pas de rapport avec une section d'un article d'encyclopédie. Le premier élément qui pourrait ressembler à un candidat sérieux est `<note/>`. Il est fait pour contenir du texte, dans le but d'expliquer quelque chose et semble assez général (il n'est pas rattaché à un genre spécifique ni à la présence d'un objet particulier sur la page). Sa sémantique demeure toutefois un peu éloignée des contraintes : sa documentation le décrit comme un commentaire additionnel («*additional comment*»), ce qui semble convenir mais qui apparaît hors du flot de texte principal («*out of the main textual stream*») alors que les développements des articles sont précisément la matière des textes encyclopédiques, pas des remarques en marge ou en bas de page.

Ainsi, le module *dictionaries* semble équipé avec un élément satisfaisant mais difficile à utiliser pour

les désignants (<usg/>), mais il n'inclut aucun élément pour les titres de sections et ne permet pas non plus d'avoir des structures imbriquées récursivement avec une sémantique satisfaisante, même encodées par des éléments qui auraient été choisis hors de ce module.

Une notion de sens mal définie

Mais au-delà des seules contraintes typographiques, d'autres problèmes apparaissent en s'intéressant à la sémantique des entrées. Il peut être difficile de définir des mots sans fournir d'entrées tautologiques (Haiman, 1980, p. 330), mais pour certains mots, et en particulier les noms propres, la notion de sens paraît encore plus insaisissable : il est possible d'écrire des sommes d'information, même sur plusieurs tomes, sans pour autant avoir «défini» un sujet (au sens où tout ce texte se rattache à l'objet mais ne le caractérise pas ni ne fixe ses limites). Il y a ainsi dans les encyclopédies en général et dans *LGE* en particulier des articles historiques et des biographies. Si la notion de «sens» peut sembler mal à-propos pour le récit d'événements historiques, il reste possible d'argumenter qu'il y a bien mise en relation du concept qui les rassemble avec le nom sous lequel l'événement est connu. Mais quand il s'agit du récit de la vie d'une personne, de décrire son rôle dans des événements historiques et ses relations à d'autre personnes la notion de sens convient encore moins. Une entrée telle que SANJO SANETOMI (La Grande Encyclopédie, T29, p.441) dont le début est visible à la figure 3.18 ne constitue en aucun cas une *définition*.

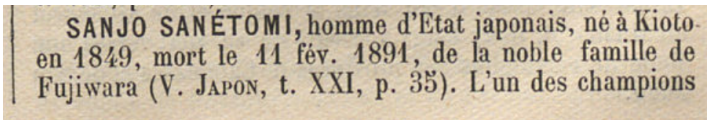
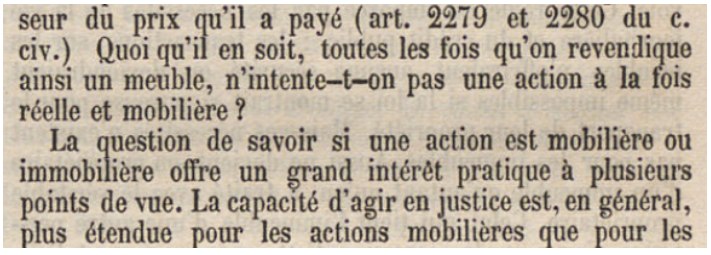

SANJO SANÉTOMI, homme d'Etat japonais, né à Kioto en 1849, mort le 11 fév. 1891, de la noble famille de Fujiwara (V. Japon, t. XXI, p. 35). L'un des champions

Figure 3.18 – Début de l'article SANJO SANETOMI dans *LGE*, T29, p.441

De plus, les encyclopédies, du fait de leur héritage philosophique des *Lumières*, ne sont pas uniquement des espaces d'assertivité mais intègrent aussi une composante dialectique intrinsèque. Certains articles posent les bases requises pour comprendre la complexité d'un problème et susciter à leur lecture des réflexions sur le sujet sans apporter de réponse définitive. Cela peut aller jusqu'à l'emploi de points d'interrogation comme dans l'article ACTION (La Grande Encyclopédie, T1, p.489) dont la figure 3.19 reproduit un extrait.


seur du prix qu'il a payé (art. 2279 et 2280 du c. civ.) Quoi qu'il en soit, toutes les fois qu'on revendique ainsi un meuble, n'intente-t-on pas une action à la fois réelle et mobilière ?

La question de savoir si une action est mobilière ou immobilière offre un grand intérêt pratique à plusieurs points de vue. La capacité d'agir en justice est, en général, plus étendue pour les actions mobilières que pour les

Figure 3.19 – Extrait de l'article ACTION dans *LGE*, T1, p.495

Dans cet extrait l'auteur invente une situation hypothétique pour montrer à quel point la démarcation entre deux sous-catégories d'actions juridiques qu'on pourrait croire mutuellement exclusives est en fait ténue. La motivation profonde du passage est de transmettre l'idée que le terme est difficile à définir sans ambiguïtés et nécessite une part d'interprétation. L'encoder dans un élément <sense/>, ou pire un élément <def/> serait véritablement un contrensens. Leur emploi pour encoder des articles encyclopédiques en général et ceux de *LGE* en particulier n'est donc pas approprié.

Trois caractéristiques fondamentales des encyclopédies au moins rendent donc le module *dictionaries* impropre à les encoder : la présence de désignants, les longs développements très structurés et le fait que la notion de «sens» y joue un rôle moins central que dans les dictionnaires. Même un

encodage composite utilisant des éléments hors de ce module sous une balise `<entry/>` ne permet pas de remplir les conditions requises.

3.2.2 Proposition d'un schéma d'encodage

Pour ces raisons, il faut chercher ailleurs de quoi encoder les articles de *LGE*. La quête n'est pas longue puisque le module *core* fournit évidemment ce qu'il faut à travers les balises `<div/>`, `<head/>` et `<p/>` qui ont en plus l'avantage d'avoir une charge sémantique moindre que `<sense/>` et `<def/>`. C'est pourquoi l'encodage retenu pour représenter *LGE* se contente d'utiliser ces balises.

Pour illustrer ce choix sur un cas concret, il est appliqué à l'article CATHÈTE reproduit à la figure 3.20. Par souci de concision et de cohérence avec les remarques ci-dessus, les exemples se limiteront à ce qui se passe au niveau de l'article, c'est-à-dire à partir de la balise immédiatement sous l'élément `<body/>` du document. Tout ce qui est lié aux métadonnées se trouve comme il convient dans un fichier TEI sous l'élément `<teiHeader/>` qui a tout ce qu'il faut pour les représenter.

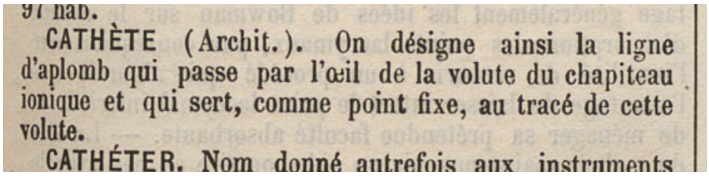

97 hab.
CATHÈTE (Archit.). On désigne ainsi la ligne d'aplomb qui passe par l'œil de la volute du chapiteau ionique et qui sert, comme point fixe, au tracé de cette volute.
CATHÉTER. Nom donné autrefois aux instruments

Figure 3.20 – Article CATHÈTE dans *LGE*, T9, p.849

Le schéma complet

En restant à l'intérieur du module *core* pour la structure, tous les éléments utiles sont disponibles et le schéma d'encodage retenu paraphrase pour l'essentiel la documentation officielle. Chaque article est représenté par une balise `<div/>`. Il est conseillé de définir dessus un attribut `xml:id` avec un identifiant unique pour l'entrée à l'échelle du corpus complet d'étude afin de pouvoir ensuite gérer les renvois et manipuler les articles sans ambiguïté dans les différents outils d'exploitation du corpus. Le système d'identification des articles retenu dans le cadre du projet consiste à simplement les numéroter séquentiellement par œuvre et par tome, et à concaténer les trois composantes – en base 10 pour celles qui sont numériques – séparées par le caractère `_` (tiret bas).

```
<div xml:id="LGE_9_2567"></div>
```

Figure 3.21 – L'élément conteneur `div` pour l'article CATHÈTE

À l'intérieur de cet élément doit se trouver un `<head/>` contenant la vedette de l'article. Un élément `<hi/>` à l'intérieur permet le cas échéant de traduire le fait qu'une typographie particulière est utilisée pour la mettre en valeur (comme du gras ou des petites majuscules).

Aucun élément satisfaisant n'a été trouvé pour représenter les désignants. Le meilleur candidat est la balise `<usg/>` dans le module *dictionaries* mais elle ne peut pas apparaître sous un élément `<head/>` [18] d'après le schéma de la TEI. Tous les chemins d'inclusion de longueur inférieure ou égale à 3 contiennent des éléments inappropriés (`<cit/>`, `<figure/>`, `<castList/>` et `<nym/>`) et doivent donc être écartés. Les autres possibilités incluent `<term/>` (imprécis) et `<rs/>` («referring string» soit «chaîne de caractères faisant référence», une formulation très générale qui pourrait convenir – après tout les désignants font référence à un domaine de connaissance – mais tous les exemples fournis dans la documentation sont des entités nommées ou des chaînes de référence,

18. le désignant fait partie de l'en-tête de l'article et c'est pourquoi le choix privilégié exploré ici est de le place sous la balise `<head/>`, mais un `<usg/>` ne peut pas apparaître non-plus ni dans un `<p/>` ni même directement dans un `<div/>`

des contextes assez différents de l'usage souhaité pour représenter des objets abstraits apparaissant seulement une fois en début d'entrée). Le meilleur compromis est de s'en tenir à des caractéristiques objectives quand elles sont présentes (souvent abréviées, un élément `<abbr/>` peut être approprié). Pour les prénoms de personnes remplaçant dans *LGE* les désignants pour les notices biographiques, l'élément `<persName/>` convient. La figure 3.22 représente l'en-tête de l'article CATHÈTE en tenant compte de ces remarques.

```
<div xml:id="LGE_9_2567">
    <head>CATHÈTE (<abbr>Archit.</abbr>)</head>
</div>
```

Figure 3.22 – L'encodage de la vedette et du désignant de l'article CATHÈTE

Ensuite, chaque sens est encodé par un élément `<div/>` dont l'attribut `type` est défini à la valeur `sense` en référence à l'élément `<sense/>` du module *dictionaries* qui aurait été employé. Cette balise devrait être numérotée à partir de 0 en fonction de son rang parmi les différents sens couverts par l'article et présente même si la vedette n'en a qu'un seul comme c'est le cas pour l'article CATHÈTE à la figure 3.23.

```
<div xml:id="LGE_9_2567">
    <head>CATHÈTE (<abbr>Archit.</abbr>)</head>
    <div type="sense" n="0"></div>
</div>
```

Figure 3.23 – La structure vide permettant de représenter le seul sens du mot CATHÈTE

De plus, chaque ligne de l'article doit commencer par un élément `<lb/>`, y compris avant l'élément `<head/>` comme le présente la figure 3.24, une façon de souligner l'importance sémantique d'un retour chariot séparant deux articles dans la mise en page dense des enyclopédies où la place est rare. Marquer explicitement les débuts de ligne permet de reconstruire un facsimile mais a également l'avantage de mettre en avant le fait que la vedette, bien qu'isolée dans sa propre balise, se trouve sur la même ligne que le reste de l'entrée, une caractéristique que les encyclopédies partagent avec les dictionnaires.

Pour compléter la structure, les sections et sous-sections apparaissant dans le corps de l'article sont encodées de la façon habituelle avec des `<div/>` contenant des `<p/>` pour les paragraphes, qui peuvent éventuellement contenir des `<head/>` locaux pour représenter leurs titres.

```
<div xml:id="LGE_9_2567">
    <lb/><head>CATHÈTE (<abbr>Archit.</abbr>).</head>
    <div type="sense" n="0">
        <p>
            On désigne ainsi la ligne
            <lb/>d'aplomb qui passe par l'œil de la volute du chapiteau
            <lb/>ionique et qui sert, comme point fixe, au tracé de cette
            <lb/>volute.
        </p>
    </div>
</div>
```

Figure 3.24 – Un encodage complet de l'article CATHÈTE

Certains articles comme BOUMERANG (La Grande Encyclopédie, T7, p.704) comportent des figures avec des légendes (voir la figure 3.25) qui sont encodées de la manière standard avec respectivement les éléments `<figure/>` et `<figDesc/>` comme dans la figure 3.26.

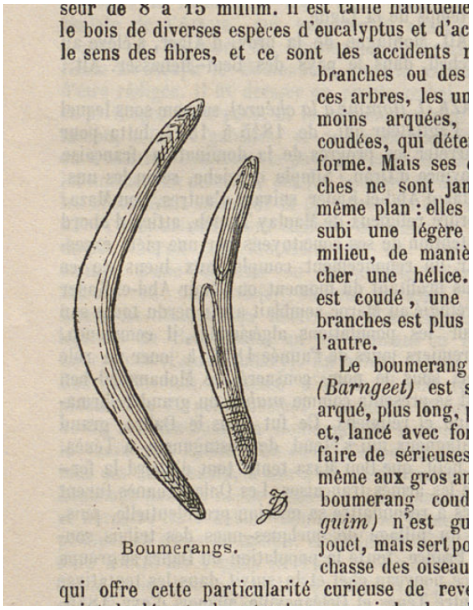


Figure 3.25 – Figure de l'article BOUMERANG dans *LGE*, T7, p.704

```
<figure>
    <graphic url="t7p725_1.png"/>
    <figDesc>Boumerangs.</figDesc>
</figure>
```

Figure 3.26 – Encodage de la figure dans l'article BOUMERANG et de sa légende

Un autre problème dû à l'abandon de la balise `<entry/>` pour représenter les articles est que l'élément `<xr/>` ne peut être utilisé dans aucun des éléments du module *core* du schéma d'encodage choisi. Or, cet élément est utile pour encoder les renvois, présents dans les encyclopédies comme dans les dictionnaires, dont un est visible à la figure 3.27 montrant l'article GELOCUS (La Grande Encyclopédie, T18, p.699). Il faut donc utiliser à la place l'élement `<ref/>`, disponible à l'intérieur d'un `<p/>`. Son attribut `target` doit avoir pour valeur le `xml:id` de l'article vers lequel il pointe, préfixé d'un dièse (`#`) comme l'illustre la figure 3.28. Cette solution a été préférée au fait d'introduire un élément `<dictScrap/>` pour pouvoir placer un `<xr/>`, parce que cela suggérerait qu'il s'agit d'un insert de contenu dictionnairique dans un contexte qui ne l'est pas, alors que la lourdeur d'encodage ajoutée est purement artificielle et ne provient que de l'impossibilité d'utiliser le module *dictionaries*.

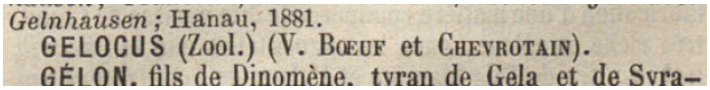


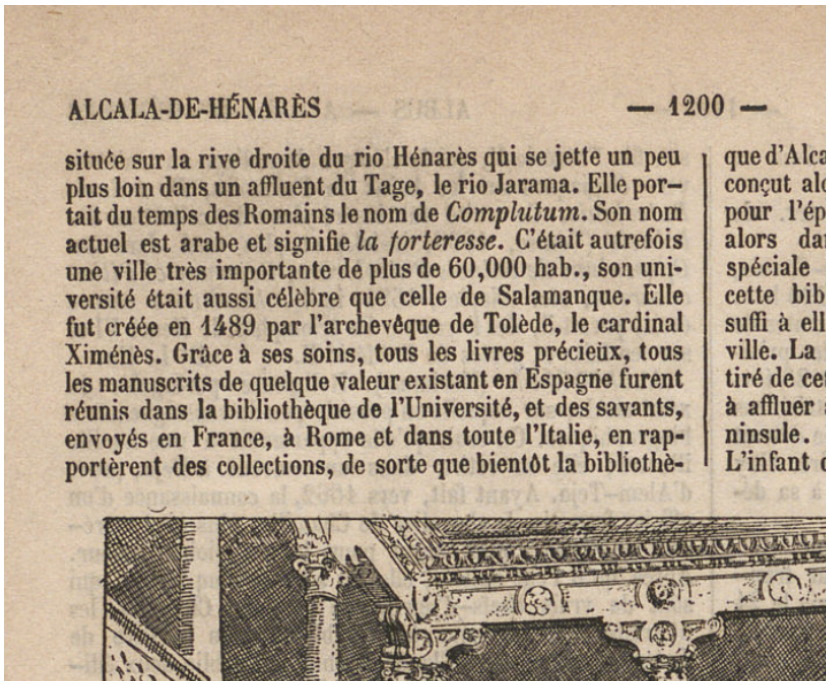
Figure 3.27 – Article GELOCUS dans *LGE*, T18, p.699

Les éléments péritextes évoqués précédemment à la section 3.1.2 et qui apparaissent sur les pages d'encyclopédies notamment pour apporter de l'information sur la position de la page (son rang et la plage de vedettes qu'elle couvre) peuvent être encodés par des balises `<fw/>` (pour «forme work»). Leurs attributs `place` et `type` devraient être renseignés pour les positionner sur la page et identifier leur fonction si celle-ci a pu être déterminée (ce qui n'est pas toujours le cas car ces éléments brefs et placés sur les bords des pages sont justement ceux qui subissent le plus souvent des dégâts quand le papier s'abîme et sont en conséquence mal lus par l'OCR).

```
<p>
    (V. <ref target="#LGE_7_247">Boeuf</ref> et
    <ref target="#LGE_10_3368">Chevrotain</ref>).
</p>
```

Figure 3.28 – Encodage des renvois dans l'article GELOCUS

Enfin, d'autres éléments TEI s'avèrent utiles pour représenter des «événements» qui surviennent au fil du texte comme le début d'une nouvelle colonne de texte ou d'une nouvelle page. La figure 3.29 montre ainsi le haut de la gauche de la dernière page du 1er tome de *LGE* (visible dans son ensemble à la figure 3.11a dans la section 3.1.3 p.61) qui comporte des éléments péritextes et constitue un début de page. Les éléments habituels (`<pb/>` pour le début de page, `<cb/>` pour le début de colonne) peuvent et doivent être utilisés comme le propose la figure 3.30.



ALCALA-DE-HÉNARÈS — 1200 —

située sur la rive droite du rio Hénarès qui se jette un peu plus loin dans un affluent du Tage, le rio Jarama. Elle portait du temps des Romains le nom de *Complutum*. Son nom actuel est arabe et signifie *la forteresse*. C'était autrefois une ville très importante de plus de 60,000 hab., son université était aussi célèbre que celle de Salamanque. Elle fut créée en 1489 par l'archevêque de Tolède, le cardinal Ximénès. Grâce à ses soins, tous les livres précieux, tous les manuscrits de quelque valeur existant en Espagne furent réunis dans la bibliothèque de l'Université, et des savants, envoyés en France, à Rome et dans toute l'Italie, en rapportèrent des collections, de sorte que bientôt la bibliothè-

que d'Alca
conçut al
pour l'ép
alors da
spéciale
cette bib
suffi à ell
ville. La
tiré de ce
à affluer
ninsule.
L'infant

Figure 3.29 – Extrait de l'article ALCALA-DE-HÉNARÈS dans *LGE*, T1, p.1200

```
<lb/>nerie générale de Madrid et du diocèse de Tolède. Elle est
<pb n="1247" />
<fw type="header">ALCALA-DE-HÉNARÈS</fw>
<fw type="pageNum">- 1200 -</fw>
<lb/>située sur la rive droite du rio Hénarès qui se jette un peu
<lb/>plus loin dans un affluent du Tage, le rio Jarama. Elle por¬
<lb/>tait du temps des Romains le nom de Complutum. Son nom
```

Figure 3.30 – Encodage d'un début de page dans l'article ALCALA-DE-HÉNARÈS

Les contraintes d'un traitement automatique

L'encodage dont il est ici question a été développé pour être appliqué par `soprano` pour encoder les articles extraits du flot brut des pages. Cette opération ne pouvait bien sûr qu'être automatisée devant le grand nombre d'objets à traiter : des 37 797 pages, `soprano` extrait 134 820 articles. Le standard XML-TEI est un outil très large permettant des usages très divers mais certains de ses éléments

comme `<unclear/>` ou `<factuality/>` permettent de représenter des informations sémantiques très abstraites (la deuxième touche à la notion de vérité même) nécessitant une compréhension très fine de l'intégralité du texte et sur laquelle même des experts humains pourraient être en désaccord. Il n'est bien sûr pas souhaitable que de telles considérations soient laissées à un processus automatique.

Pour cette raison, le choix d'un schéma d'encodage pour *LGE* a été principalement contraint par le besoin de ne prendre en compte que des caractéristiques objectives du texte pouvant être extraites sans intervention humaine. Des traits comme la position relative des blocs et l'usage d'une typographie particulière comme le gras ou les majuscules suffisent par exemple à repérer les vedettes d'articles.

Le cas des renvois est un peu particulier et pourrait apparaître comme un contre-exemple à ce principe. Pourtant, la pratique consistant à établir des liens entre articles est tellement fréquente (autant dans les dictionnaires que dans les encyclopédies) que le processus échappe dans une certaine mesure au reste du discours et prend une forme particulière et codifiée, entre parenthèses et après un token précis qui invite à effectuer soi-même la redirection lors de la lecture. Dans *LGE* la très large majorité des renvois a ainsi lieu entre parenthèses, et le renvoi est suggéré par le verbe «voir» raccourci à son initiale «V.» comme cela était déjà visible sur la figure 3.27. Quelques très rares exceptions conservent la forme «voir», «voy.» ou même «voyez» (avec ou sans majuscule) mais il s'agit dans la plupart des cas de renvois externes, vers d'autres œuvres que *LGE*. Lorsqu'il s'agit bien de renvois entre articles, c'est que le contexte particulier de ce renvoi exige une formulation plus explicite qui sort du «lien» codifié typographiquement pour revenir à une tournure plus construite : par exemple une reprise anaphorique pour éviter une répétition – «Voy. ce mot» à l'article COMBATTANT (La Grande Encyclopédie, T11, p.1153) – ou un renvoi multiple – «voir les mots Poste et Télégraphe» à l'article AFFRANCHISSEMENT (La Grande Encyclopédie, T1, p.702).

À ces rares exceptions près, il semble donc possible d'utiliser cette structure très codifiée pour détecter et encoder les renvois, bien que cela n'ait pas encore été implémenté dans `soprano`. Renseigner correctement l'attribut `target` de la balise `<ref/>` pourrait en revanche se révéler plus difficile et nécessiter de traiter les articles en plusieurs étapes, pour d'abord découvrir toutes les entrées existantes – et ce faisant associer des vedettes aux identifiants uniques d'articles – avant d'essayer de résoudre les mots après ces «V.» entre parenthèses. Puisque la segmentation des articles a lieu indépendamment pour les différents tomes et que les renvois ont une portée qui dépasse le tome où ils sont utilisés, il n'est pas possible pour `soprano` de simplement construire cette table associative en mémoire avant de produire les fichiers et il faudra donc envisager un traitement en plusieurs passes.

La résolution des renvois, si elle doit être implémentée un jour ne devrait donc pas remettre en question une caractéristique principale de l'encodeur utilisé. Si des lexicographes pourront reprocher à l'encodage proposé une trop grande superficialité, il a l'avantage de ne pas nécessiter le stockage de structures de données trop complexes en mémoire pour une durée longue. L'algorithme responsable de l'encodage dans `soprano` produit ainsi les articles dès qu'il peut sans attendre d'avoir une représentation complète de l'article en mémoire. Dès qu'un bloc est prêt, il est encodé et le XML-TEI correspondant est écrit dans le fichier de destination. Cela permet de conserver une utilisation de la mémoire vive raisonnable (de l'ordre de 3Go), ce qui permet de traiter plusieurs tomes simultanément sur la plupart des machines récentes. Ainsi même en prenant environ 4min par tome, le temps total requis pour segmenter et encoder la totalité de *LGE* sur une machine de 16Go de RAM peut rester voisin de 40min en lançant 3 processus en parallèle (au lieu des $31 \times 4 = 124min = 2h04$ totales nécessaires sans parallélisation).

État de l'implémentation actuelle

L'implémentation de référence pour ce schéma d'encodage est le programme soprano [19]. Ce logiciel a déjà servi à créer deux versions mais il n'implémente pas encore toutes les spécifications de l'encodage proposé précédemment. La figure 3.31 montre ainsi l'état actuel de l'article CATHÈTE à la fin de la phase d'encodage.

```
<div xml:id="LGE_9_2567">
    <lb/><head>CATHÈTE</head> (Archit.). On désigne ainsi la ligne
    <lb/>d'aplomb qui passe par l'œil de la volute du chapiteau
    <lb/>ionique et qui sert, comme point fixe, au tracé de cette
    <lb/>volute.
</div>
```

Figure 3.31 – Encodage actuel de l'article CATHÈTE produit par `soprano`

La détection des vedettes ne permet pas encore de reconnaître les désignants et c'est pourquoi ils apparaissent à l'extérieur de la balise `<head/>`. Aucun travail n'est non plus fait pour résoudre les abréviations (ici, par exemple, retrouver qu'«Archit.» correspond à «Architecture») et les encoder comme telles. Cela serait pourtant faisable grâce à la liste des abréviations présente au début du premier tome et qui a été encodée sous un format numérique dans le cadre du projet [20]. Il serait également intéressant de distinguer les noms de domaines des prénoms de personnes puisque l'ambiguïté existe dans *LGE* pour ce qui tient typographiquement la place de désignant.

De la même manière, puisque la détection des titres au début des sections n'est pas complète, aucune analyse de la structure n'est effectuée pour l'instant sur le texte à l'intérieur de chaque article qui est gardé sans structure, directement sous l'élément `<div/>` correspondant à l'entrée plutôt que dans des `<div/>` imbriquées. Les paragraphes ne sont pas encore identifiés et pour cette raison pas encodés non plus.

En revanche, `soprano` gère déjà correctement les figures et leurs légendes quand il y en a. L'encodeur garde la trace des lignes, pages et colonnes en cours et insère les éléments vides correspondants (`<lb/>`, `<pb/>` et `<cb/>`) lorsqu'une de ces structures s'ouvre. Les pages sont numérotées en fonction de leur rang dans le document numérique tel qu'il est disponible, qui diffère de celui inséré par les éditeurs. Il permet donc de savoir à quelle photo de page se rendre, pas de connaître le nombre écrit en haut de la page correspondante. La suite des numéros de page est en effet surprenamment difficile à suivre. D'abord la numérotation ne démarre pas à la première page, les premières pages avant et après le frontispice n'ont pas de numéro. De plus les cartes géographiques présentes dans l'œuvre sont imprimées sur des feuilles distinctes des cahiers en sextodecimo contenant le texte proprement dit avec lesquels elles sont simplement reliées. Elles ne comportent pas de numéro de page, et la numérotation se met donc en quelque sorte en pause sur les quelques pages qui les contiennent. Si elle n'incrémente donc pas simplement de un à chaque page en entrée du processus, il n'est pas non plus facile de la lire directement sur chaque page sans contexte : en tant qu'élément péritexte, les caractères sont souvent malmenés par l'OCR et ne ressemblent parfois pas à des entiers. Pour les obtenir, il faudrait donc pouvoir croiser plusieurs informations (les numéros lus sur toutes les pages pour s'assurer que la suite est croissante ou les marques péritextes des cahiers sextodécimo entre autres) pour espérer synthétiser une donnée assez fiable. Cela n'est pour l'instant pas implémenté.

19. https://gitlab.inria.fr/abrenon/soprano
20. Le fichier correspondant à ces données est distribué dans le dépôt https://gitlab.liris.cnrs.fr/geode/lge-meta.git

3.2.3 Le mythe du «texte brut»

Parler d'encodage se rapporte généralement – comme c'est le cas dans toute la sous-section précédente 3.2.2 – à des formats hautement structurés comme les différents dialectes du XML, le JSON ou bien le YAML entre autres. On oppose fréquemment ces formats à un autre censé représenter le contenu textuel sans mise en forme aucune, auquel on réfère le plus souvent par l'expression «texte brut», et qu'on enregistre par convention dans des fichiers portant l'extension `.txt` [21]. Le format texte étant employé pour certaines études de cette thèse (voir la section 3.3.1), il est important de mettre en lumière que, comme les autres formats cités précédemment, il peut en fait faire l'objet de choix décisifs requiérant des interprétations sémantiques radicales et l'adoption de conventions : une forme d'encodage.

Pour revenir à la distinction faite à la section 2.5.1 (p.36), parler de format renvoie au texte en tant que matière, que quantité indénombrable plutôt que comme une unité sémantique au sein d'un corpus d'étude. Rastier (1996, p. 18) refuse de réduire la notion de texte à sa seule représentation numérique – une «chaîne de caractères» – mais la définition qu'il propose à la page 19 s'avère tout de même éclairante sur cette composante qui représente sa «substance graphique». Il ne limite pas l'usage du mot «texte» aux seules productions écrites mais requiert tout de même qu'une «suite linguistique» soit «fixée sur un support» pour pouvoir parler de texte. Une version numérique constitue une trace qui permet de fixer un texte et de le diffuser (partage, copie, sauvegarde).

Le deuxième aspect de la définition de Rastier (1996, *ibid*) insiste sur l'existence d'un texte dans le cadre d'une pratique sociale. Quel que soit le support utilisé pour fixer un texte, c'est rarement la trace elle-même mais plutôt l'objet abstrait qu'elle dénote qui intéresse son lectorat ou les linguistes. Avant même de considérer quelque niveau d'analyse que ce soit sur le texte – ni sémantique, ni syntaxique ni même la lemmatisation – plusieurs particularités physiques de son support sont écartées pour considérer un texte. Le choix des propriétés du texte à ignorer et la manière dont cette abstraction a lieu dépend de la pratique sociale dans le cadre de laquelle il a été produite.

Ainsi on considérera la plupart du temps qu'un texte est «le même» bien qu'il soit prononcé par deux personnes différentes (avec une voix différente), mais dans le cadre d'une œuvre où ces voix auraient été associées à des personnages différents, le changement de voix pourrait alors revêtir une importance. À l'écrit, la mise en page d'un texte ou la fonte utilisée pour l'imprimer est généralement ignorée, même si certaines formes littéraires jouent précisément sur ces paramètres. En termes physiques, et de manière tout à fait analogue avec la réduction décrite à la section 3.1.3 du format XML-ALTO à un flot linéaire de texte, on peut voir cette mise de côté de certaines propriétés comme une projection visant à réduire les dimensions d'un texte pour obtenir un objet fondamentalement unidimensionnel. En termes mathématiques, il s'agit plutôt de classes d'équivalence : la pratique sociale associée à un texte définit une relation selon laquelle deux traces représentent «le même» texte. Derrière l'expression «texte brut», il y a la croyance qu'un «fichier texte» pourrait constituer une trace textuelle canonique du texte qu'il représente [22]. Cette section se propose de montrer que cette croyance est infondée.

Lignes et paragraphes

La notion d'encodage dans le contexte de format textuel désigne plutôt d'habitude celui des caractères : l'association à un glyphe d'un code symbolique et numérique, et d'une séquence d'octets pour

21. on trouve aussi du «text brut» sous d'autres formes, par exemple dans une cellule d'un fichier tableur, ou dans une valeur au sein d'un fichier JSON ; le terme «fichier» (textuel) sera employé par souci de simplicité mais les remarques faites dans cette section s'appliquent aussi à ces autres méthodes de stockage de texte

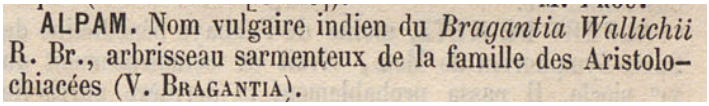
22. et donc qu'on pourrait assimiler un texte à cette trace particulière, de la même façon qu'on peut parfois noter par abus «2» la classe d'équivalence de tous les nombres congrus à 2 pour un certain modulo $p > 2$, en identifiant cette classe à l'entier 2 qui en fait partie et constitue un représentant fort commode

le représenter. Ainsi le concept de la lettre «A», première lettre de l'alphabet latin telle qu'on la trace en majuscule, reçoit le code arbitraire 65, et est représenté par une séquence d'un unique octet, `0x41` – représentation électronique de la valeur 65, notée en base hexadécimale par convention. Il existe différents encodages, incompatibles entre eux (UTF-8, ISO-8859-1, EUC-JP…), utilisés couramment sur diverses plateformes mais ce n'est pas de cette distinction dont il s'agit ici. En effet, il est aisé de préciser l'encodage d'un fichier qu'on distribue (ils sont tous nommés et spécifiés très précisément), et normalement un seul format est utilisé au sein d'une même plateforme de traitement. Mais même en supposant une convention d'encodage universellement acceptée, des problèmes pratiques demeurent.

La plupart des formats d'encodage étendent en effet le format ASCII (c'est-à-dire que tout fichier ASCII est par exemple aussi un fichier UTF-8 ou ISO-8859-1 valide), regroupant la plupart des signes typographiques communs, les chiffres et les lettres de l'alphabet latin sans diacritiques : tous les glyphes sont donc représentés par la même séquence d'octets quel que soit l'encodage utilisé (et en l'occurrence, toutes ces «séquences» ne comportent en réalité qu'un seul octet, ce pourquoi on identifiera dans ce qui suit le code d'un caractère et l'octet le représentant). Or, ce sous-ensemble commun représenté par ASCII est déjà suffisant pour requérir des choix cruciaux dans la représentation des fichiers texte. Le jeu de caractères ASCII contient en effet plusieurs façons de séparer les mots : l'espace (`' '`, code `0x20`) mais également la tabulation (`'\t'`, code `0x09`) et la nouvelle ligne (`'\n'`, code `0x0a`). En pratique suivant le système d'exploitation les lignes dans un fichier peuvent être terminées par la séquence `"\n"` (un seul `'\n'`), `"\r"` (un seul caractère de code `0x0d`) ou `"\r\n"`, ce pourquoi il sera question ici de «fin de ligne» pour désigner cette séparation sémantique sans préoccupation pour sa réalisation exacte (qui constituerait un niveau de choix supplémentaire dans ce qui suit mais sera ignoré car fixé sur la plateforme sur lequel le traitement a lieu).

Le texte lorsqu'il est affiché pour un humain (comme c'est le cas de ces lignes dans le PDF ou la version papier que vous êtes en train de lire) est réparti en lignes d'une largeur raisonnable pour aérer le texte et coïncider avec la taille des formats de papier usuels. Il est possible dans un fichier texte de former au contraire des lignes arbitrairement longues, mais ce n'est pas sans présenter des inconvénients techniques (difficultés pour lire et se déplacer dans le fichier, lenteur potentielle suivant l'éditeur de texte utilisé). En pratique, des caractères séparateurs seront utilisés pour isoler des unités indépendantes : phrases, paragraphes, sections… le niveau auquel se fera cette séparation et la manière dont elle sera réalisée constituent un premier ensemble de «paramètres d'encodage textuel» qui sont autant de degrés de liberté pour déterminer la façon de représenter un texte sous forme numérique. Un premier choix peut consister par exemple à réserver les fins de ligne à la séparation des paragraphes, et à conserver tout le texte d'un paragraphe sur une même ligne (ce qui n'est pas sans poser d'autres difficultés, voir ci-dessous). Un autre choix peut consister à créer des lignes d'une largeur maximale fixe, comme c'est le cas à la figure 3.32b. Dans ce deuxième cas, la largeur à utiliser constitue un deuxième paramètre, le nombre de «Caractères Par Ligne» (CPL), souvent lié au matériel sur lequel ces fichiers sont traités plus qu'à une intention de la personne qui les édite : les largeurs de lignes utilisées historiquement sont de l'ordre de 80 CPL, mais la figure 3.32b montre une sorte de facsimile textuel de l'article tel qu'il apparaît sur la page imprimée de *LGE* (visible juste au-dessus à la figure 3.32a) dont les colonnes de texte font 50 caractères sans compter la ponctuation et les espaces fines d'après l'Avant-Propos (La Grande Encyclopédie, T1, p.XI). Quel que soit le nombre de CPL choisi, le contenu réel sauvé dans le fichier se voit encombré de caractères non significatifs, introduits pour des raisons techniques mais ignorés dès que cela sera possible à la lecture ou à l'analyse. Sur la figure 3.32c, il s'agit des caractères `'\n'` (nouvelle ligne) présents dans la suite d'octets qui représente l'article, surlignés en mauve.

Le choix précédent a ensuite des répercussions sur la séparation entre paragraphes : en effet, si les fins de lignes ne leur sont pas réservées, il faut trouver une façon de désambiguïser un nouveau paragraphe et une ligne qui commencerait simplement par un mot en majuscule. En pratique dans


ALPAM. Nom vulgaire indien du *Bragantia Wallichii* R. Br., arbrisseau sarmenteux de la famille des Aristolochiacées (V. Bragantia).

(a) Photo de l'article

```
ALPAIYI. Nom vulgaire indien du Bragantia Wallichii
R. Br., arbrisseau sarmenteux de la famille des Aristolo-
chiacées (V. Bragantia).
```

(b) Texte extrait par `soprano`

```
00000000  41 4c 50 41 49 59 49 2e  20 4e 6f 6d 20 76 75 6c  |ALPAIYI. Nom vul|
00000010  67 61 69 72 65 20 69 6e  64 69 65 6e 20 64 75 20  |gaire indien du |
00000020  42 72 61 67 61 6e 74 69  61 20 57 61 6c 6c 69 63  |Bragantia Wallic|
00000030  68 69 69 0a 52 2e 20 42  72 2e 2c 20 61 72 62 72  |hii.R. Br., arbr|
00000040  69 73 73 65 61 75 20 73  61 72 6d 65 6e 74 65 75  |isseau sarmenteu|
00000050  78 20 64 65 20 6c 61 20  66 61 6d 69 6c 6c 65 20  |x de la famille |
00000060  64 65 73 20 41 72 69 73  74 6f 6c 6f 2d 0a 63 68  |des Aristolo-.ch|
00000070  69 61 63 c3 a9 65 73 20  28 56 2e 20 42 72 61 67  |iac..es (V. Brag|
00000080  61 6e 74 69 61 29 2e 0a                           |antia)..|
```

(c) Contenu du texte détaillé par octet

Figure 3.32 – Article ALPAM dans *LGE*, T2, p.437

l'*EDdA* comme dans *LGE*, les paragraphes ne sont pas séparés par des lignes vides mais sont indentés d'une largeur d'environ 2 caractères. Un choix «facsimile» existe donc encore pour ce deuxième paramètre : reproduire l'indentation, mais comme pour la largeur des lignes, cette décision peut donner lieu à plusieurs implémentations en pratique. La bataille des conventions d'indentation est un lieu commun des langages de programmation : on peut trouver l'emploi de la tabulation (`'\t'`), ou d'un nombre fixe d'espaces (généralement 2, 4 voire 8). Rester au plus près de la mise en page dans le corpus exigerait sans doute le choix d'une tabulation ou de deux espaces.

Récupérer un texte linéaire

En fonction des choix de paramètres d'encodage textuel décrits ci-dessus, un même texte peut donc donner lieu à différentes traces textuelles. Récupérer un texte linéaire à partir d'une de ses traces n'est jamais évident et demande toujours un travail, qu'un humain réalise à la lecture sans y penser mais qui nécessite des règles précises basées sur la connaissance du choix d'encodage textuel pour des algorithmes qui tenteraient d'y accéder. Ainsi, le lectorat voit naturellement une espace entre «*Wallichii*» et «R.» sur la figure 3.32b (alors qu'il n'y en a pas, il n'y a qu'une fin de ligne), et ignore en revanche le tiret et la fin de ligne pour former le mot «Aristolochiacées» à la ligne suivante alors que cette séquence n'est pas présente telle quelle dans le fichier.

Ce problème du recollement des lignes au sein d'un paragraphe est bien plus difficile qu'il n'y paraît. En première approche, pour joindre deux lignes il suffit d'ajouter une espace quand le dernier caractère de la première n'est pas un tiret (`'-'`) et de simplement concaténer la première privée de son `'-'` final à la deuxième dans le cas contraire. Mais en pratique, l'élimination du tiret dépend du contexte : il fait partie de certains mots composés comme «peut-être» ou «pis-aller» et permet de rattacher les enclitiques au mot qui les précède comme dans «celui-là» ou «puis-je», ce qui interdit sa suppression. Des ressources lexicales peuvent aider dans le premier cas mais avec des risques de faux-positifs (que faire sans analyse sémantique fine d'un hypothétique nom propre «Pisaller» ne tenant pas sur le reste d'une ligne et qui aurait donné lieu à la séquence «Pis-`'\n'`aller» dans un fichier ?). Dans le second cas, la grande variété d'éléments combinables n'incite de toute façon pas à opter pour une simple vérification d'appartenance à une liste préétablie d'exceptions connues. Par conséquent, une analyse fine des dynamiques au niveau au moins morphosyntaxique sinon syntaxique semble donc requise pour déterminer avec certitude le sort d'un tiret survenant juste avant

une fin de ligne.

Loin de constituer une base commune pour représenter un texte, le format `.txt` représente donc en fait un impensé de l'encodage. Dans le cadre des présents travaux, les paramètres d'encodage textuel utilisés pour représenter les articles correspondent à ce qui est présenté à la figure 3.32b. Les fichiers utilisent l'encodage UTF-8 et le caractère `'\n'` (nouvelle ligne) pour représenter les fins de ligne. Des fins de ligne sont insérées là où il y en avait dans l'édition de *LGE* dont est extraite l'article, de manière à conserver la largeur originelle du texte. Les paragraphes sont en revanche séparés entre eux par deux fins de ligne (c'est-à-dire, visuellement dans le fichier, par une ligne vide). Les indentations ne sont pas marquées. Ce choix n'est pas nécessairement meilleur qu'un autre : il retarde le recollement des mots divisés entre plusieurs lignes pour ne pas introduire d'erreur définitive à cette étape (suite à la remarque du paragraphe précédent), mais rend de ce fait l'accès au texte un peu plus complexe.

À l'issue de cette section, l'ensemble des paramètres clefs pour représenter les textes du corpus ont été identifiés et discutés. En confrontant les besoins d'encodage définis par l'observation des pages de *LGE* avec les structures définies par le schéma d'encodage *TEI*, les limites du module *dictionaries* pour représenter une encyclopédie telle que *LGE* ont été mises en évidence. Un encodage alternatif utilisant des éléments plus généraux de la *TEI* – tout comme l'avait fait l'ARTFL dans sa version de l'*EDdA*, mais en opérant des choix légèrement différents – a été illustré sur quelques articles et une portion de cet encodage est implémentée dans l'outil `soprano`, permettant de générer automatiquement des articles au format XML-TEI à partir des pages en ALTO livrées par la BnF. Cet outil permet également de produire des fichiers en format texte, dont il a été montré que le contenu faisait également l'objet de choix précis qui ont été explicités pour cette version.

3.3 Des textes au corpus

Après avoir décrit les œuvres puis les manières de les représenter par des formats numériques ainsi que les choix opérés sur cette question, il est temps pour conclure ce chapitre de montrer comment ces étapes préparatoires convergent. Elles permettent de constituer un corpus qui structure les données et facilite leur emploi avec différents outils pour les besoins de cette thèse. Cette section sera également l'occasion de faire une remarque méthodologique sur l'obtention et la maintenance du corpus et des métadonnées qu'il contient.

3.3.1 Différents formats pour différents usages

Deux formats de départ différents sont utilisés pour représenter les articles du corpus en amont de tout traitement. Le premier est un format XML-TEI utilisant les éléments principaux du schéma (module *core*, par opposition au module *dictionaries*, pour les raisons présentées à la section 3.2.1 p.62). Le choix de balises diffère légèrement entre l'*EDdA* (pour laquelle l'ARTFL utilise des éléments `<div1/>` pour les articles dont les métadonnées sont stockées dans des éléments `<index/>`) et *LGE* (pour laquelle les rares métadonnées accessibles sont stockées directement comme attribut de l'élément `<div/>` choisi pour représenter les articles). Le second est un format textuel utilisant les choix d'encodage décrits à la sous-section 3.2.3 (p.79).

Le second peut s'obtenir du premier en extrayant le contenu de certaines balises précises dans l'encodage des fichiers. C'est de cette façon qu'a été obtenue la version texte des articles de l'*EDdA*. Pour *LGE* en revanche, la version textuelle peut être extraite directement à partir des pages en XML-ALTO, indépendamment de la version XML-TEI, les deux correspondant à des modes de sortie de `soprano` (la génération des métadonnées décrite à la section 3.3.2 constitue également un mode séparé).

Format textuel

Le format textuel est en pratique le plus utilisé dans les différentes études conduites au cours de cette thèse. La plupart des outils de TAL attendent en entrée un flot d'unités indépendantes de texte (avec toutes les réserves émises à ce sujet à la section 3.2.3 p.77) débarrassées des contingences liées à leur représentation (retours à la ligne au milieu d'une phrase, divisions au mileu des mots coupés par un retour à la ligne...). Pour analyser les phrases en syntaxe et morphosyntaxe, un outil doit voir au moins toute une phrase d'un coup (il peut en voir plusieurs au sein d'un même paragraphe mais donnera une analyse erronée s'il ne reçoit qu'une phrase partielle); pour classer un article, un modèle doit avoir accès à tout le texte qu'il contient.

Puisque le choix fait pour représenter les articles dans des fichiers texte consiste à garder la mise en page des lignes telles qu'elles apparaissent dans les encyclopédies du corpus, il faut d'abord «nettoyer» le texte avant de pouvoir appeler dessus une chaîne de classification, un analyseur syntaxique ou un annotateur d'entités nommées. Pour éviter les redondances de traitement et s'assurer que tous les outils ont reçu la même entrée, la phase d'extraction a été factorisée, et une version linéarisée des articles est produite et stockée dans des fichiers texte utilisant d'autres paramètres d'encodage textuel (voir p.77 et seq.).

Le format textuel rempli donc deux rôles dans ces travaux. Il y a d'abord un format «patrimonial» dont le but est de fournir une représentation relativement fidèle des articles et de leur mise en page (l'exception majeure étant les paragraphes, qui sont séparés par des lignes vides pour limiter les ambiguïtés). De ce format patrimonial est tiré un format texte «linéaire» dont le but est d'être consommable directement par des outils de TAL. Ce deuxième format se caractérise par les paramètres d'encodage textuels suivants : les fichiers utilisent le même encodage UTF-8 que pour le texte patrimonial, les fins de lignes typographiques sont ignorées et les mots divisés sont recollés (à l'aide d'une heuristique simpliste par rapport aux remarques faites à la section 3.2.3 en particulier p.79, ce point est discuté dans la sous-section 3.3.3). Enfin, la séparation des paragraphes par des lignes vides, bien qu'inutile en soi suite au recollement, est conservée pour favoriser la lisibilité de ces fichiers par des humains à des fins de contrôle de la qualité des données.

Les outils de classification automatique utilisés tout au long du chapitre 4 prennent ainsi en entrée du texte linéaire et produisent pour chaque texte une valeur symbolique, le nom d'une classe dont relève le texte d'après le modèle. Cette information est stockée dans des fichiers de métadonnées tabulaires (voir la sous-section 3.3.2). L'analyse syntaxique en dépendances universelles (UD) est effectuée avec la librairie python Stanza [23] (Qi et al., 2020), qui retourne les articles annotés au format CoNLL-U. Cette annotation comprend une étiquette morphosyntaxique pour les tokens identifiés, exprimée avec les parties de discours propres aux UD, les UPOS [24].

Format XML-TEI

Le format XML-TEI (voir la section 2.1.2), au contraire du format texte discuté ci-dessus permet de conserver plus d'information sur les pages tout en distinguant le rôle de chaque objet et sa relation aux autres. Les éléments péritextes par exemple, présentés à la section 3.1.2 p.47, introduiraient beaucoup d'ambiguïté dans des fichiers texte, à moins d'en complexifier encore l'encodage. L'option `-k` (`--keep` en version longue) de `soprano` permet de choisir quel type d'éléments conserver et permet de produire des fichiers ne contenant que le texte des articles. Par défaut tous les éléments sont conservés (en particulier les changements de page, les légendes d'images, etc.), ce qui prend tout son sens avec le format XML-TEI.

Il constitue donc une version plus riche, cruciale pour la diffusion des données et leur réutilisation

23. https://stanfordnlp.github.io/stanza/
24. https://universaldependencies.org/docs/u/pos/

dans le cadre d'autres études. C'est ce qui a permis aux présents travaux de bénéficier des résultats des efforts des projets ARTFL et ENCCRE. Le projet GEODE comporte une dimension patrimoniale et met à disposition le texte de *LGE* au format XML-TEI. Il est à souhaiter que ces données connaissent le même destin que celles de l'*EDdA*. Certains thèmes de recherche un peu trop éloignés de la problématique initiale et qui n'ont pas pu être suivis au long de cette thèse pourraient en effet intéresser de nombreux spécialistes des HN. Le rapport entre géographie et colonisation suggéré à la Préface (La Grande Encyclopédie, T1, p.I) – voir la section 3.1.2 p.54, ou le traitement des notices biographiques de femmes et d'hommes par exemple mériteraient sans doute une exploration soigneuse.

Les logiciels de textométrie utilisés dans les présents travaux prennent en entrée des fichiers au format XML-TEI. Le logiciel TXM (Heiden et al., 2010) permet ainsi d'explorer directement le corpus à partir de la version XML-TEI des deux œuvres et de faire des mesures de nombre de tokens des œuvres ou de leur partitions (notamment par domaine de connaissance, voir la section 5.1.1 à partir de la page 134). L'extension intégrant l'étiqueteur TreeTagger [25] permet même d'accéder à des étiquettes morphosyntaxiques pour les tokens. En pratique, cette extension n'a pas été utilisée dans cette étude puisque l'annotation a été faite avec Stanza sur les fichiers texte, et la sortie en CoNLL-U a été extraite pour produire une version XML-TEI intégrant en plus les informations morphosyntaxiques en UPOS. En ce qui concerne le Lexicoscope (Kraif, 2016) qui utilise un format dual, conservant du contenu CoNLL-U dans un environnement de balises XML-TEI ordinaires (des paragraphes `<p/>` groupant des éléments `<s/>` dont chacun contient l'annotation CoNLL-U de la phrase qu'elle représente plutôt que du texte directement lisible pour des humains), les fichiers XML-TEI ont été entièrement synthétisés à partir des sorties de Stanza.

3.3.2 Principes de structuration

Un corpus n'est pas constitué que de données, quelle que soit la qualité du format qui les représente. Une part importante de l'effort d'organisation de cette thèse a été employée à mettre en regard des données des métadonnées qui soient pertinentes pour les analyses et pratiques à utiliser techniquement.

Des désignants à des ensembles de domaines

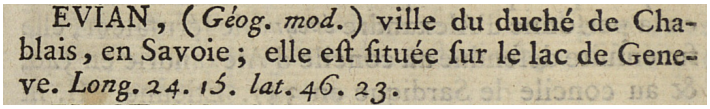
La notion de domaine de connaissance introduite en détail à la section 3.1.2 (p.49) joue un rôle central dans cette thèse. Disposer d'un ensemble de valeurs à associer aux articles pour représenter leur appartenance à un domaine a demandé un travail important de normalisation. Le chapitre 4 décrit en détail les techniques utilisées pour identifier le domaine de chaque article mais il s'agit ici, en amont de ces opérations, de déterminer les valeurs à utiliser pour la classification (c'est-à-dire le codomaine des fonctions implémentées par les différents classifieurs).

D'Alembert avait prédit que non seulement une encyclopédie ne serait jamais complète mais que son organisation même pourrait être renouvelée par l'expérience de son lectorat. Dans son *Discours préliminaire* (L'Encyclopédie, T1, p.xv), il offre du contexte au «Systême» (voir la figure 3.4 p.52) si rigide en apparence en affirmant que «la forme de l'arbre encyclopédique dépendra du point de vûe où l'on se mettra pour envisager l'univers littéraire. On peut donc imaginer autant de systèmes différens de la connoissance humaine, que de Mappemondes de différentes projections [...]».

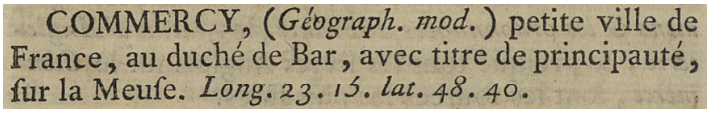
Il n'est pas surprenant pour un projet de l'envergure de l'*EDdA* et s'étendant sur autant d'années (21 au total) que les désignants qui matérialisent les domaines de connaissance au sein des articles comptent plus de 7 000 formes uniques différentes. La plupart sont dues à des variations sur les abréviations, l'orthographe ou la ponctuation utilisée. Une très grande partie de ces formes uniques n'ont que très peu d'occurrences, et un système automatique n'a pas de compréhension sémantique des

25. https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/

regroupements qu'il pourrait être légitime de faire. Utiliser ces désignants tels quels n'aurait que peu de sens pour constituer des classes dans l'optique d'études contrastives. Il faut donc déjà normaliser les formes variées prises par les désignants, ce qui a fait l'objet de travaux à la fois par l'ARTFL et l'ENCCRE. Dans l'exemple de l'article EVIAN présenté à la figure 3.33a, le désignant «Géog. mod.» se normalise en «Géographie moderne» ce qui permet de l'identifier à celui de COMMERCY à la figure 3.33b qui est «Géograph. mod.» et diffère donc en toute rigueur.


EVIAN, (*Géog. mod.*) ville du duché de Chablais, en Savoie; elle eſt ſituée ſur le lac de Geneve. *Long.* 24. 15. *lat.* 46. 23.

(a) Article EVIAN dans l'*EDdA*, T6, p.146


COMMERCY, (*Géograph. mod.*) petite ville de France, au duché de Bar, avec titre de principauté, ſur la Meuſe. *Long.* 23. 15. *lat.* 48. 40.

(b) Article COMMERCY dans l'*EDdA*, T3, p.700

Figure 3.33 – Deux entrées de «Géographie moderne» dans l'*EDdA*

L'ARTFL a réduit ce nombre à 2 620 classes normalisées (Horton et al., 2009, para. 3) qui correspondent dans l'encodage XML-TEI de l'œuvre à l'attribut `normclass`. À partir des mêmes données, l'ENCCRE a créé manuellement une liste de 2 160 «désignants explicités» , prenant également en compte les tournures évoquées précédemment à la section 3.1.2 («en termes de», etc., voir p.51). L'ENCCRE a ensuite cherché à identifier les thématiques proches et rassemblé ces classes en 312 domaines, puis sur un deuxième niveau en 44 «ensembles de domaines». Une altération minime de ces ensembles de domaines a été faite dans le cadre du projet GEODE pour rassembler au sein d'une seule classe *Métiers* les différentes branches de professions qui dans le jeu de l'ENCCRE ont chacune un ensemble de domaines à part. Ce choix définit un ensemble de 38 *domaines regroupés* visibles à la figure 3.34 et utilisés dans la comparaison de méthodes de classification automatique conduite au chapitre 4.

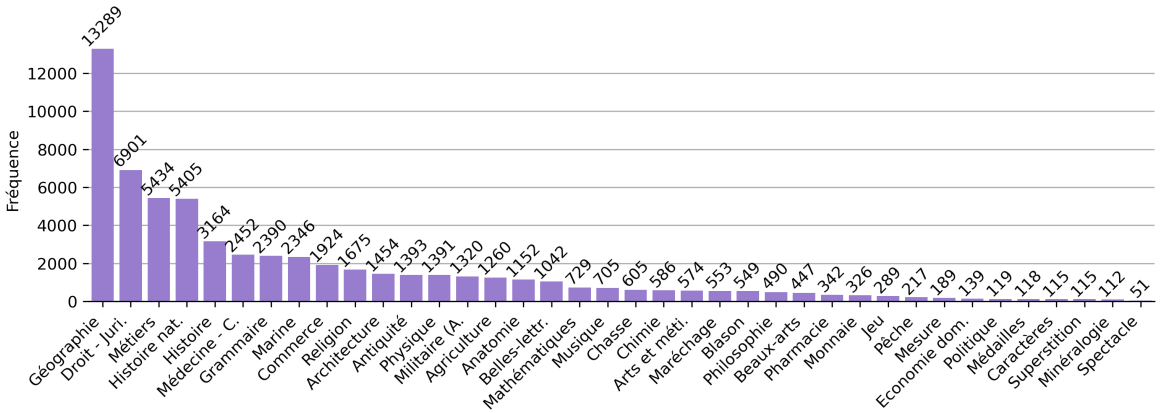


Figure 3.34 – Distribution des nombres d'articles par *domaine regroupé*

Dans ce nouveau système, les deux articles de la figure 3.33 se retrouvent classés en *Géographie*, le domaine regroupé qui inclut la *Géographie moderne* et comporte 13 289 articles. La figure 3.34 montre qu'il s'agit du groupe le plus fréquent parmi les articles pour lesquels une classification des auteurs est connue, et que le suivant, *Droit - Jurisprudence* compte quasiment deux fois moins d'articles (6 901).

Ce jeu des «domaines regroupés» posait encore quelques difficultés de dimensionnement pour les

études contrastives présentées au chapitre 5. De plus, la réduction du déséquilibre entre classes permet d'obtenir des prédictions de meilleures qualités sur les classes moins représentées. Pour cette raison un nouveau jeu de classes plus simple a été créé pour représenter les domaines de connaissance dans l'*EDdA*. Ce jeu ne comprend plus que 17 classes qui ont été construites manuellement par Katherine McDonough, historienne et membre du projet GEODE, en vue de rendre compte de manière concise des thématiques principales abordées dans l'*EDdA* et, au contraire des regroupements de classes opérés par l'ENCCRE, en assemblant des «branches» et des «sous-branches» à différents niveaux dans le «Systême figuré» (voir 3.4). Le jeu de classes, appelées par convention dans le projet les «superdomaines» (c'est le nom qui sera retenu dans le reste de cette thèse), a donc été conçu dès l'origine comme un regroupement non pas d'*ensemble de domaines* ou de *domaines regroupés* mais directement de désignants normalisés. Il est constitué des classes suivantes :

— Agriculture
— Beaux-arts
— Belles-lettres
— Chasse
— Commerce
— Droit Jurisprudence
— Géographie
— Histoire
— Histoire naturelle
— Médecine
— Métiers
— Militaire
— Musique
— Philosophie
— Physique
— Politique
— Religion

Structure des métadonnées

Le premier besoin pour attacher des informations additionnelles aux textes sans devoir dupliquer leur contenu est la détermination d'une information qui identifie de manière unique les fichiers, ce qui constitue une clef primaire dans le vocabulaire des bases de données. En distinguant les articles des propriétés qu'on souhaite pouvoir leur associer, c'est l'approche la plus modulaire puisqu'elle permet de croiser entre elles plusieurs métadonnées issues de traitements différents ou de sélectionner des sous-corpus, par opposition à une approche en «jeu de données» ou *datasets* dans laquelle toutes les informations requises, texte et métadonnées, seraient groupées ensemble de manière autonome en un objet unique.

Les présents travaux optent pour une clef primaire simple basée sur le rang des articles dans chaque tome des œuvres du corpus, suivant le choix de l'ARTFL (voir la section 3.1.2 p.48). Puisque le corpus d'étude comprend plusieurs encyclopédies, un code pour chaque œuvre (une des deux valeurs symboliques `EDdA` ou `LGE` [26]) vient compléter les numéros de tome et d'article pour former un triplet. Il avait également été envisagé d'utiliser à la place du rang un identifiant basé sur la vedette de l'article (complétée par un entier pour garantir unicité malgré les homonymies potentielles), de manière à obtenir un système plus résistant aux redécoupages alors que la segmentation de *LGE* était en cours d'amélioration (ainsi, un identifiant avait plus de chance de demeurer identique alors que le rang de tous les articles survenant dans un tome après un article où la segmentation avait été corrigée

26. les valeurs `Universalis` et `Wikipedia` étaient également possible pour le reste du corpus GEODE mais n'ont pas été utilisées en pratique

se retrouve modifié). Cette idée a été abandonnée suite à la difficulté d'obtenir une représentation unifiée d'un système à l'autre de caractères accentués [27].

Contrairement aux différentes étapes de traitement et d'annotation qui représentent chacune les articles dans un format différent, les métadonnées constituent des propriétés stables qui peuvent être associées aux textes en tout point de la chaîne (un auteur, un domaine de connaissance…). Elles sont stockées dans des fichiers au format TSV, conçus pour faciliter l'accès par des moyens programmatiques (voir la section 2.1). La structure des métadonnées relève donc déjà davantage du code que des données, ce pourquoi les colonnes sont nommées en anglais, *lingua franca* pour les codes sources. Ainsi, le tableau 3.1 représente les métadonnées minimales nécessaires pour référencer par exemple les deux articles de la figure 3.33, c'est-à-dire leur clefs primaires.

work	volume	article
EDdA	3	3184
EDdA	6	421

Tableau 3.1 – Clefs primaires des entrées COMMERCY et EVIAN dans l'*EDdA*

À cette clef primaire doivent s'ajouter des colonnes pour assurer la navigation et le retour au plein texte. La plus évidente est la vedette (un humain cherche un article à propos d'EVIAN, pas le 3 184ème article de l'*EDdA*). Pour outrepasser le problème d'homonymie des vedettes soulevé ci-dessus, une version normalisée et rendue unique de cette vedette est également stockée, bien qu'elle n'ait finalement pas été retenue pour constituer la clef primaire. Enfin, le numéro de la page auquel apparaît un article est stocké dans ces métadonnées de navigation pour retrouver l'article dans les pages numérisées. Avec ces informations supplémentaires, les deux articles précédents sont représentés par les métadonnées visibles au tableau 3.2.

work	volume	article	headword	name	page
EDdA	3	3184	COMMERCY	commercy-0	700
EDdA	6	421	EVIAN	evian-0	146

Tableau 3.2 – Métadonnées de navigation pour les entrées COMMERCY et EVIAN dans l'*EDdA*

Ensuite, d'autres dimensions supplémentaires peuvent être ajoutées (domaines, auteurs, taille des textes…). Leur stockage dans des fichiers distincts fournit le plus de flexibilité, permettant par exemple de mettre à jour le domaine prédit pour un article suite à une amélioration dans la chaîne de traitement de classification automatique, sans devoir par exemple regénérer tous les fichiers qui dépendent de traitements prenant seulement en compte les tailles. Cette approche permet de cerner au plus près les dépendances entre données au fil des différents traitements, au seul prix de la répétition des clefs primaires dans chaque fichier, bien moins contraignante que la répétition des textes entiers des articles eux-mêmes. Il peut arriver de devoir grouper plusieurs métadonnées pour une étude particulière, ce qui peut s'effectuer par une simple jointure sur la clef primaire des articles. Le résultat d'une telle jointure peut s'observer dans les sources du présent manuscrit [28], par exemple dans le fichier `data/corpus/metadata.tsv` où les métadonnées utiles à la génération

27. dans l'encodage UTF-8, utilisé de manière assez générale sur de nombreux systèmes modernes, différentes conventions de normalisation existent (on distingue les formes normales composées – NFC – et décomposées – NFD) et tous les systèmes n'optent pas pour la même normalisation. Ainsi, un 'é' (e accent aigu) peut être réalisé directement en NFC (il lui correspond le code U+00e9, représenté par la séquence d'octets "`0xc3 0xa9`"), ou bien «assemblé» en NFD en combinant un 'e' simple (de code U+0065 représenté par l'octet `0x65`) avec U+0301, un code représentant un accent aigu à combiner avec une autre lettre, et qui se représente lui même par la séquence "`0xcc 0x81`"

28. https://gitlab.liris.cnrs.fr/abrenon/manuscrit

des figures du manuscrit ont été groupées en un seul fichier pour réduire la place utilisée par le dépôt (les outils de versionnement sont surtout faits pour travailler sur de petits fichiers, limiter la taille du dépôt permet d'éviter la dégradation des performances).

Structure des dossiers et des fichiers

Un corpus, en tant qu'association structurée de données et de métadonnées rassemblées dans un but d'étude précis, est représenté numériquement par un dossier contenant à la fois les textes du corpus et leurs métadonnées. Les différents fichiers associés à chaque métadonnée sont stockés à la racine du corpus. Chaque format correspondant à un état des textes (une étape de traitement ou un type d'encodage) est stocké dans un sous-répertoire dédié du corpus (par exemple `Text/` ou `TEI/` pour les formats évoqués à la section 3.3.1 qui contiennent les textes en amont de tout traitement, respectivement aux formats textuels ou XML-TEI).

Le triplet de la clef primaire assure non seulement le lien entre les métadonnées mais se manifeste aussi dans chacune des arborescences correspondant à un état donné des textes. À l'intérieur de chaque sous-dossier, les fichiers sont en effet organisés par œuvre (dans un répertoire dont le nom est la valeur de l'attribut `work` de l'article). Chaque dossier d'œuvre (`EDdA/` ou `LGE/` donc) contient un répertoire par tome, nommé en préfixant la valeur de l'attribut `volume` d'un 'T' (t majuscule) : `T1/`, `T2/`… jusqu'à `T17/` dans `EDdA/` et `T31/` dans `LGE/`. À l'intérieur de chaque dossier de tome, chaque article est nommé par son rang, suffixé de l'extension idoine (`.txt`, `.xml`, `.conllu`…). Ainsi, il est extrêmement facile d'accéder aux contenus des articles en itérant sur leurs métadonnées, puisque pour un état donné leurs chemins relativement au répertoire correspondant à cet état s'obtiennent par une simple opération de concaténation.

3.3.3 Reproductibilité

Sans constituer une spécification précise, les deux sous-sections précédentes 3.3.1 et 3.3.2 montrent la logique qui a présidé aux choix techniques de conception du corpus utilisé dans le reste de la thèse. De même que la description des idées clefs d'un algorithme fournit une aide précieuse pour comprendre une de ses implémentations dans un langage en particulier, elles tentent de documenter les données diffusées à l'issue du projet pour aider à s'en emparer. Avant de refermer complètement ce long chapitre, quelques remarques sur la notion de reproductibilité donneront une perspective aux choix opérés.

Cette thèse n'est pas reproductible. Au-delà de ce seul constat hélas vrai pour la plus grande partie des thèses soutenues à ce jour, cette sous-section se propose de donner un sens à cette assertion dans le contexte disciplinaire qui lui est propre avant d'aller voir quels efforts ont tout de même été entrepris en direction de la reproductibilité, comment cette notion s'intègre à la méthodologie esquissée par la structure du corpus décrite dans le reste de cette section et quels progrès restent nécessaires.

De la reproductibilité en SHS ?

Le premier obstacle à la reproductibilité de ces travaux réside dans l'impossibilité de partager une moitié du corpus. Les pages scannées de *LGE* et toutes les données extraites à partir de ces pages à l'issue du projet DISCO-LGE sont bien sûr diffusées sur la plateforme Nakala [29] et placées sous licence libre Creative Commons By-Nc-Sa [30], ce qui signifie qu'elles peuvent être utilisées librement aux seules conditions de signaler leur origine, de ne pas en faire un usage commercial et de ne les re-partager – elles ou toute donnée nouvelle créée à partir d'elles – que sous les mêmes conditions. Cela

29. https://nakala.fr/10.34847/nkl.74eb1xfd
30. https://creativecommons.org/licenses/by-nc-sa/4.0/deed.fr

est naturellement compatible avec une diffusion pour d'autres travaux de recherche. En revanche, en ce qui concerne l'*EDdA*, les données utilisées ont été communiquées par l'ARTFL sous condition de non-divulgation et il est donc impossible de les repartager. L'ARTFL accepterait vraisemblablement de partager les tomes de l'*EDdA* en XML-TEI avec d'autres équipes qui en feraient la demande, mais il est impossible de comparer cette autre hypothétique version des données avec celle utilisée dans le cadre du projet GEODE : il se pourrait tout à fait que l'ARTFL ait fait des améliorations dans ses chaînes de traitements et qu'elles ne produisent plus exactement les mêmes données.

Si des spécialistes du XIX^{ème} siècle estiment surprenante la proportion d'articles de géographie dans *LGE* trouvée dans les présents travaux au vu de leur connaissance préalable de l'œuvre, il leur sera possible de reprendre les données et les outils utilisés pour la mesure afin de recalculer séparément le résultat ; éventuellement d'inspecter les états intermédiaires pour trouver la source de l'erreur ou confirmer le résultat. Si un doute semblable devait survenir sur les données de l'*EDdA*, il serait possible de réappliquer les traitements, mais pas de s'assurer que les données d'entrées seraient les mêmes. Si un résultat différent venait à être trouvé, il ne serait pas possible de déterminer l'origine de l'erreur.

Au-delà de ce seul exemple hypothétique, il est légitime de s'interroger sur la pertinence d'une démarche reproductible dans le cadre des HN ou, plus généralement, des Sciences Humaines et Sociales. Les problématiques de reproductibilité ont émergé dans des domaines où l'informatique occupe une place centrale, depuis la théorie des langages et la compilation où Thompson (1984) démontrait en 3 pages la difficulté à s'assurer de ce que faisait un programme – même en ayant accès à son code source (!) – jusqu'au Calcul Hautes-Performances (Courtès, 2021) où la reproductibilité joue un rôle clef pour contrôler la cohérence de logiciels déployés sur des plateformes variées et s'assurer de la validité des résultats qui y sont obtenus. Dans un contexte où le logiciel occupe une part croissante des travaux dans de nombreux domaines jusqu'à devenir un élément central de la notion de preuve, la reproductibilité apporte des garanties de confiance essentielles pour la recherche. Mais les HN ne constituent pas un domaine où les résultats s'obtiennent par application d'un simple calcul sur un ensemble d'entrées. Au contraire, elles passent par de nombreuses étapes exploratoires réalisées dans des logiciels dédiés. Ces interfaces (dans le cadre des présents travaux, TXM et le Lexicoscope sont des exemples évidents) sont conçues pour présenter les données mais requièrent des interactions humaines pour guider l'analyse. Il serait donc vain de vouloir automatiser de bout en bout une chaîne de traitement en HN jusqu'à la rendre reproductible et vérifiable de manière automatique sur une plateforme d'intégration continue ; toutefois, la reproductibilité peut constituer un atout précieux pour les HN et les présents travaux doivent reconnaître ses apports.

Mise en pratique dans le contexte de l'étude

Puisque la rediffusion de la totalité des données de cette thèse est inaccessible, il faut revenir aux définitions des concepts élémentaires en reproductibilité pour savoir ce qu'il est possible d'atteindre. La «reproductibilité» représente la possibilité de reproduire le plus exactement possible une expérience, en appliquant la «même» analyse aux «mêmes» données. Malheureusement, la difficulté à s'accorder sur le sens à donner à ces deux identités a pour conséquence une inhomogénéité de la terminologie (Plesser, 2017). Pour Rougier et al. (2017), «reproduire» une expérience signifie la reconstituer exactement, en utilisant exactement les mêmes outils et logiciels sur un jeu de données identique jusqu'au dernier octet, alors que «répliquer» tolère des données équivalentes. Cette définition est en contradiction avec celle des sciences expérimentales mais tend à s'imposer dans les sciences computationnelles.

Les deux concepts sont dans une relation ambivalente. Il est plus difficile techniquement et donc plus contraignant de concevoir une expérience en garantissant qu'elle sera reproductible, ce qui témoigne d'un contrôle plus fin sur la totalité de ses paramètres. Cependant une expérience réplicable

peut posséder une plus grande force de preuve : un résultat peut apparaître simplement à cause d'une erreur dans les données, d'une intéraction particulière entre code et données, ou tout simplement de vraies données tout à fait particulières et qui ne se généralisent pas. En appliquant les mêmes analyses sur de nouvelles données indépendantes, une seconde équipe qui réplique une expérience augmente la confiance que la communauté scientifique peut accorder à un résultat. Schloss (2018) distingue aussi la «robustesse» qui à l'inverse consiste à appliquer des analyses différentes aux mêmes données. Enfin, une expérience est «généralisable» quand il est possible de parvenir à des conclusions similaires en appliquant des analyses différentes de l'expérience initiale à d'autres données. Ces quatre critères, résumés au tableau 3.3 ne s'excluent pas mais sont idéalement à cumuler pour produire des recherches dignes de confiance.

Les seules expériences de cette thèse pouvant viser à la reproductibilité sont celles qui ne portent que sur *LGE*. Toutes celles qui intègrent l'*EDdA* ne peuvent par construction viser qu'à la réplicabilité. En pratique, étant donnée la part cruciale qu'a jouée l'*EDdA*, notamment dans l'entraînement des modèles de classification automatique (voir le chapitre 4 dans son ensemble) préalable aux analyses contrastives menées au chapitre 5, il faut viser à la réplicabilité sur l'ensemble des travaux. Il serait intéressant de tester la robustesse des résultats produits sur *LGE* en appliquant d'autres méthodes de classification aux articles de cette encyclopédie.

Tableau 3.3 – Quatre notions différentes de la reproductibilité

	mêmes données	données différentes
même analyse	Reproductible	Réplicable
analyse différente	Robuste	Généralisable

Mais la reproductibilité (au sens large à nouveau, hors de celui défini par le tableau 3.3) peut également apporter aux phases d'investigation des HN. La nature irrégulière des objets d'étude gêne le développement de traitements s'appliquant parfaitement à toutes les données. À l'échelle mésoscopique du présent corpus – la taille des encyclopédies est proverbialement grande mais les volumes de données en jeu restent très largement inférieurs aux mégadonnées générées par les collectes automatiques de dispositifs numériques – il faut considérer que tout ce qui peut arriver arrivera. En ce qui concerne les désignants par exemple, il suffit quasiment d'envisager l'existence d'une régularité et de tenter de la capturer dans un motif pour qu'il soit contredit par une donnée du corpus, qu'elle ait réellement été produite telle quelle par un choix ou une erreur humaine ou qu'elle émerge à cause des bruits générés par les imperfections des différentes étapes de traitement (vieillissement du papier et de l'encre, qualité de la numérisation, performances de l'OCR…). Ainsi, dans ce type d'étude il s'agit en permanence de placer un curseur entre ce qui est acceptable pour répondre à une question précise et ce qui est perfectible en vue d'études futures. Trouver cet équilibre engendre un va-et-vient continu entre traitement des données et analyses. Pour cette raison, les choix faits dans la représentation du corpus et en particulier dans la conception des métadonnées favorisent la plus grande flexibilité, en tâchant de minimiser l'effort nécessaire pour relancer les traitements affectés par un changement dans une donnée. Cette stratégie repose sur le constat empirique qu'il est vain d'espérer que la première version sera la bonne et qu'il vaut mieux considérer toutes les données en aval dans la chaîne de dépendance comme temporaires, susceptibles de mises à jour et donc jetables. La détermination d'un ensemble de domaines de connaissance présentée à la sous-section 3.3.2 p.82 d'une façon très linéaire résulte en réalité de ce type de négociation entre code et données. Naturellement, pour que ces allers et retours conduisent à un processus d'amélioration continue plutôt qu'à tourner en boucle, il est crucial de pouvoir contrôler finement ce qui change ou pas et de pouvoir s'assurer qu'un seul paramètre du traitement a été modifié à chaque nouvelle tentative pour mesurer son effet propre et pas par exemple celui d'un changement dans une librairie dont dépend le

programme utilisé. Cela permet de valider une amélioration et de garantir qu'il n'y a pas de régression dans la qualité des données. C'est ce type de garanties que peut apporter la reproductibilité des logiciels.

Si des «instantanés» de *LGE* à plusieurs étapes clefs des traitements sont distribués via Nakala pour faciliter la prise en main des données, toutes peuvent être regénérées localement (et donc vérifiées) en appliquant les mêmes traitements sur les pages OCRisées au format XML-ALTO, seules réelles entrées de toute la partie du corpus qui concerne cette œuvre. Le logiciel `soprano` [31] utilisé pour extraire les articles et les encoder en format texte ou en XML est distribué sous licence libre (BSD 3) et empaqueté pour le gestionnaire de paquets reproductible Guix [32]. Un autre dépôt [33] contient tous les outils utilisés pour les différentes analyses et la production des formats enrichissant les données (classification, annotation syntaxique, etc.). Le développement de ces outils a aussi donné lieu à plusieurs librairies qui sont bien sûr elles aussi distribuées sous licence libre (GNU GPL v3 pour `ghc-geode` [34] et `geopyck` [35]). Les outils spécifiques à l'*EDdA*, développés antérieurement au projet GEODE sont pour rappel (voir la section 3.1.3) aussi disponibles [36]. L'ensemble de ces logiciels est empaqueté pour Guix et ces paquets sont distribués dans un canal Guix spécifique au projet [37] afin de les regrouper et de faciliter leur installation. Enfin, les mêmes efforts de reproductibilité ont été appliqués à ce manuscrit lui-même. Ses sources sont ouvertes et disponibles [38] (sous licence GNU GPL v3). Le dépôt permet de recompiler une version PDF du texte que vous êtes en train de lire mais surtout, il contient les données utilisées pour tracer la plupart des tableaux et figures (celles utilisant un thème de couleur mauve ou une palette arc-en-ciel pour les domaines de connaissance). Il est ainsi possible de vérifier les résultats annoncés et d'étudier les codes utilisés pour les générer.

La longueur de ce chapitre reflète dans une large mesure le temps consacré dans cette thèse au traitement des données en vue de constituer le corpus d'étude. À partir des œuvres, dont la lecture et l'étude directe ont constitué un point de départ utile pour le reste des travaux, une réflexion a été développée sur les points communs structurels des articles dans le corpus et la notion clef de domaine de connaissance. Les besoins spécifiques à *LGE*, notamment en matière d'encodage, ont amené à identifier les différences intrinsèques qui séparent les encyclopédies des dictionnaires et à formuler des choix pour représenter les articles dans des fichiers numériques. Ces décisions ont été mises en perspective de l'utilisation des données qui est faite aux chapitres 4 et 5 et de la démarche reproductible qui sous-tend l'ensemble de cette thèse.

31. https://gitlab.inria.fr/abrenon/soprano
32. https://guix.gnu.org/
33. https://gitlab.liris.cnrs.fr/abrenon/outillage
34. https://gitlab.liris.cnrs.fr/geode/ghc-geode
35. https://gitlab.liris.cnrs.fr/abrenon/geopyck
36. https://gitlab.huma-num.fr/alicebrenon/eddaclinic
37. https://gitlab.liris.cnrs.fr/geode/geode-packages
38. https://gitlab.liris.cnrs.fr/abrenon/manuscrit

Chapitre 4

Classification automatique en domaines de connaissance

4.1 Comparaison de méthodes de classification automatique 92
4.1.1 Démarche 92
4.1.2 Choix des modèles et préparation des jeux de données 93
4.1.3 Évaluation des performances 94
4.1.4 Classification non supervisée 100
4.2 Relations entre *Géographie* et autres domaines de connaissance 102
4.2.1 Observations préliminaires des erreurs 103
4.2.2 Similarités lexicales 112
4.2.3 Application de la centralité à la structure du modèle 119
4.3 Annotation des articles 123
4.3.1 Choix d'un classifieur 123
4.3.2 Extension à *LGE* 126

À l'issue du chapitre 3, les textes du corpus sont disponibles dans une version XML-TEI ainsi que dans des fichiers texte qui suivent la convention décrite à la section 3.2.3 page 77. Mais si une structure a été choisie pour représenter le corpus et ses métadonnées, la valeur à associer à chaque article pour représenter son domaine de connaissance (voir la section 3.1.2 p.49) reste à déterminer pour une partie des articles de l'*EDdA* et la totalité de ceux de *LGE*. Le présent chapitre décrit les efforts menés sur ce versant en le considérant comme un problème de classification automatique. Il reprend dans une large mesure et étend des résultats déjà exposés dans Brenon et al. (2022).

La première étape consiste à comparer les résultats de différentes méthodes de classification, avant d'étudier à la section 4.2 les erreurs commises par l'une d'entre elles pour essayer de comprendre ce qu'elles peuvent révéler sur les ressemblances et les différences entre domaines de connaissance. Enfin, la dernière section 4.3 décrit concrètement l'annotation en domaines de connaissance du corpus entier.

S'agissant d'AA, la remarque faite p.16 à la fin de la section 1.3 s'applique tout autant dans ce chapitre qu'elle s'appliquait à la section 2.4 de l'état de l'art : devant la prédominance de l'emploi des noms anglophones pour les méthodes d'AA, même chez des locuteurs français, une exception locale est faite dans toute cette section à l'emploi exclusif de termes francophones qui est la règle dans le reste

de cette thèse. Ce choix se justifie par le souci de ne pas produire de formes traduites quelque peu artificielles qui n'auraient de toute façon pas cours hors de ces pages.

4.1 Comparaison de méthodes de classification automatique

Cette première section présente l'étude comparative menée pour choisir un modèle de classification, en commençant par exposer la démarche suivie avant de décrire son implémentation (sous-section 4.1.2). La sous-section (4.1.3) présente les résultats obtenus et discute leur signification tout en gardant à l'esprit le fait que les performances obtenues par les différentes méthodes présentées ne rejoignent pas nécessairement les raisons pour lesquelles les éditeurs d'encyclopédies assignent une classification aux articles. La sous-section (4.1.4) présente pour finir une brève expérience conduite sur des méthodes de classification non supervisées.

4.1.1 Démarche

Pour obtenir une classification complète des articles dans le système des domaines regroupés choisis à la section 3.3.2 (voir p. 82), il suffit pour les articles qui possèdent déjà un domaine de le convertir vers ce nouveau jeu d'étiquettes, et pour les autres de prédire une valeur à l'aide d'un modèle de classification automatique. Pour disposer d'un tel classifieur, il faut un ensemble d'articles possédant déjà un domaine regroupé assez vaste pour l'entraîner puis tester ses performances (pour évaluer la qualité de l'étiquetage obtenu). Ce deuxième cas se ramène donc au premier : il est nécessaire de commencer par appliquer la correspondance entre le domaine qui leur a été attribué par l'ENCCRE («ensemble de domaine») et les domaines regroupés. Puisque la seule différence entre les deux jeux d'étiquettes réside dans le regroupement des différentes branches professionnelles («métiers du bois», «métiers du papier», etc.) en une classe *Métiers* unique (voir p.83), la conversion est en réalité très simple puisqu'elle consiste à conserver la classe initiale partout sauf pour les métiers.

Toutefois, l'ENCCRE (dont proviennent les annotations en ensemble de domaines) et l'ARTFL (qui a fourni les fichiers encodés des articles, voir la section 3.1.1 p.46) segmentent différemment l'œuvre en textes ce qui complique cette conversion en empêchant d'établir une correspondance parfaite entre les deux corpus. En effet, en suivant la distinction faite à la section 3.1.2 (voir p.48) entre «article» et «entrée», l'ARTFL groupe ensemble certaines entrées mais en considère d'autres comme des textes individuels, au contraire de l'ENCCRE pour qui le niveau de référence est l'entrée et dont tous les textes correspondent à ce niveau de granularité (dans ses métadonnées, un article possède toujours au moins une entrée avec laquelle il coïncide si elle est unique). Par conséquent, seuls 69 531 articles ont pu être appairés entre ces deux sources sur les 74 190 au total présents dans la version de l'œuvre étudiée.

Pour cette raison certains articles sans désignant marqué typographiquement possèdent un domaine (inféré par l'ENCCRE d'expressions telles que «en termes de...» qui remplacent parfois les désignants – voir la section 3.1.2 p.51); à l'inverse, des articles pourtant pourvus d'un désignant mais qui n'ont pas pu être mis en correspondance avec leur équivalent chez l'ENCCRE peuvent se retrouver sans domaine et nécessiter ensuite une prédiction du modèle [1]. Dans les données de l'ENCCRE, 12 635 articles ne sont pas pourvus d'une classification explicite, mais l'application d'une classification experte basée sur le contenu des articles parvient à réduire ce nombre à 2 392. Le jeu de données de l'ARTFL qui n'a utilisé que les désignants explicites fournit quant à lui une classification pour 56 688 articles.

Cette approche est inspirée de l'étude conduite par Horton et al. (2009) qui avaient employé un

1. ce point est tout à fait contre-intuitif mais, s'il reste heureusement très marginal, il est important de le garder en tête pour comprendre pourquoi des prédictions ont pu être réalisées sur des articles qui comportent pourtant un ou plusieurs désignants, notamment dans la section 5.1.2 (p.150 et seq.)

modèle Naive Bayes Multinomial pour étendre la classification originale des auteurs. Bien que leur étude ne comporte pas d'évaluation formelle des performances de leur modèle, il offre une analyse détaillée d'un échantillon des résultats. Au vu de la vitesse à laquelle l'état de l'art en apprentissage automatique a bougé au cours de la décennie 2010-2020, il est intéressant d'actualiser leurs travaux pour y intégrer des modèles de classification récents, non seulement des méthodes d'AA classiques mais aussi des modèles d'AP.

4.1.2 Choix des modèles et préparation des jeux de données

Les présents travaux comparent des approches combinant différentes méthodes de vectorisation et de classification supervisée décrites aux sections 2.4.1 et 2.4.2 (voir pp.30 et seq.) ou issues de méthodes présentées dans ces sections. Ainsi, plutôt qu'un modèle *LSTM* (voir la section 2.4.2 p.32), c'est sa variante plus récente et bidirectionnelle *BiLSTM* qui a été testée. Cette étude vise à tester le plus de combinaisons possibles, mais toutes les combinaisons ne sont pas réalisables à cause de contraintes spécifiques.

Par exemple, la méthode *Naive Bayes* nécessite des vecteurs d'entrée sans valeurs négatives, ce qui impose l'emploi d'une représentation vectorielle *Bag of Words* ou *TF-IDF* mais empêche l'usage de plongements de mots. De plus, puisque les modèles de la famille de *BERT* intègrent à la fois la représentation vectorielle de leurs entrées et les couches de classification, permettant de travailler directement à partir du texte, ils n'ont pas pu être combinés à d'autres méthodes de vectorisation. En tenant compte de ces remarques, les combinaisons testées sont les suivantes :

1. Une vectorisation *BoW* et des algorithmes traditionnels d'AA (*Naive Bayes*, *Logistic Regression*, *Random forest*, *SVM* et *SGD*);
2. Une vectorisation utilisant un plongement de mots statique (*Doc2Vec*) et des algorithmes d'AA traditionnels (*Logistic regression*, *Random Forest*, *SVM* et *SGD*);
3. Une vectorisation utilisant un plongement de mots statique (*FastText[fr]*) et des algorithmes d'AP (*CNN* et *BiLSTM*);
4. Une approche tout-en-un utilisant des modèles de langue contextuels préentraînés (*BERT*, *CamemBERT*) dont les poids sont réajustés à la tâche dont il est question (*fine-tuning*).

Pour les algorithmes d'AA, la librairie Scikit-learn [2] et plus précisément *GridSearchCV* a été utilisée pour déterminer les hyperparamètres. Les résultats de la recherche avec *GridSearchCV* sont disponibles dans le dépôt git de l'étude [3]. Dans les expériences d'AP c'est la version française de *FastText* (Bojanowski et al., 2017) pour les plongements préentraînés qui a été utilisée pour vectoriser les entrées et les implémentations des *CNN*s et des *BiLSTM*s proviennent de la librairie Keras. Le modèle de *CNN* testé emploie une architecture classique avec une couche de plongement pour des vecteurs de dimension 300 (la taille des vecteurs de mots préentraînés par *FastText*), une couche de convolution avec la fonction d'activation ReLu, une couche de max-pooling et une couche softmax pour la sortie. Pour le modèle *BiLSTM*, plusieurs architectures ont été testées et celle retenue utilise une taille de vecteurs de 300 également, une couche *BiLSTM* (avec un taux de décrochage de 20%), une couche de max-pooling, deux couches denses avec la fonction d'activation ReLu avec un taux d'abandon de 50% entre chaque, et une couche softmax en sortie. Pour ajuster les modèles *BERT*, les données d'entraînement ont été groupées par lots de huit fichiers (à cause de contraintes de place en mémoire) pendant quatre époques – Devlin et al. (2019) recommandent entre deux et quatre époques pour ce type d'opération.

Deux jeux de données distincts sont utilisés pour entraîner et évaluer les modèles (respectivement dénommés dans ce qui suit *train* et *test*). Ces deux jeux de données (voir le tableau 4.1) contiennent les

2. https://scikit-learn.org/stable/
3. https://gitlab.liris.cnrs.fr/geode/EDdA-Classification/-/tree/master/reports

articles étiquetés avec l'ensemble de domaines regroupés déduit de l'étiquetage opéré par l'ENCCRE. Dans ce qui suit, traitant d'un problème de classification, ces domaines regroupés ne seront donc plus vus comme des objets épistémologiques mais comme autant d'étiquettes abstraites apposées sur des textes : des classes. Pour cette étude, une seule classe a été retenue par article. En réalité, 3 654 articles de l'*EDdA* (soit environ 5%) possèdent plusieurs classes. Quelques traitements sont appliqués avant de pouvoir débuter les expériences proprement dites. D'abord, les articles sans classe et plus courts que 15 mots sont écartés. Cela produit un jeu de 58 509 articles répartis sur une des 38 classes possibles. Ensuite, les désignants sont retirés des articles pour éviter que les modèles ne fondent leurs prédictions que sur ces éléments (ce qui donnerait des scores artificiellement élevés sur le jeu de validation mais les rendrait bien plus mauvais pour prolonger la classification sur les articles qui justement n'en sont pas pourvus). Enfin, la ponctuation et les mots vides – déterminants, prépositions, pronom etc. en reprenant la distinction de Lehmann & Martin-Berthet (2018, chap. 1, §1.2.2) – sont retirés des textes et les mots restants sont lemmatisés [4]. Cette dernière étape est contournée pour les modèles de la famille *BERT* (ligne 4. dans la liste précédente) puisque ces modèles peuvent recevoir en entrée des textes complets et la version des textes qui leur est présentée, pour l'apprentissage comme pour la prédiction est celle issue de l'étape précédente à savoir le texte purgé seulement de son éventuel désignant. Une proportion de 20% est réservée pour le jeu d'évaluation (*test*) et le reste devient le jeu d'entraînement (*train*).

Tableau 4.1 – Détail du nombre d'articles dans le corpus entier, le corpus prétraité et les jeux d'entraînement et de test.

Jeu de donnée	# articles
corpus entier	74 190
corpus prétraité	58 509
train (tout)	46 807
train (max 1 500)	27 381
train (max 500)	14 058
test	11 702

Du fait du déséquilibre entre les nombres d'articles des différentes classes, il apparaît essentiel d'évaluer l'impact de l'utilisation d'échantillons d'articles pour les classes surreprésentées. On peut faire l'hypothèse que cela va réduire les biais des modèles en leur faveur. Un paramètre supplémentaire est donc ajouté aux expériences pour comparer les résultats des modèles entraînés sur des jeux d'entraînement plus ou moins homogènes. Les classes sont limitées à 500 articles maximum, 1 500 articles maximum ou ne sont pas limitées du tout pour constituer 3 jeux d'entraînement différents. Ces 3 cas correspondent aux 3 lignes *train* dans le tableau ci-dessus (tableau 4.1), qui représentent donc des jeux d'entraînement de respectivement 14 058, 27 381 et 46 807 articles au total.

4.1.3 Évaluation des performances

Les résultats de classification sont évalués en mesurant la précision, le rappel et la F-mesure (moyenne harmonique de la précision et du rappel) de chaque méthode. Afin d'obtenir des résultats généraux pour toutes les classes, les F-mesures moyennes sur les 38 classes pondérées par leur nombre d'articles sont examinés. Le tableau 4.2 présente les F-mesures obtenues par les différentes méthodes.

4. les tokens sont identifiés avec la librairie SpaCy et la lemmatisation proprement dite est effectuée à l'aide du FrenchLefffLemmatiser (Sagot, 2010)

Tableau 4.2 – F-mesures moyennes pour les différents modèles entraînés sur un échantillon de 500 articles maximum par classe (1), 1500 maximum par classes (2) et tous les articles disponibles de chaque classe (3) appliqués sur le jeu de test.

Classification	Vectorisation	F-mesure (1)	(2)	(3)
Naive Bayes	BoW	0.63	0.71	0.70
	TF-IDF	0.74	0.69	0.44
Logistic Regression	BoW	0.74	0.77	0.79
	TF-IDF	**0.77**	0.79	**0.81**
	Doc2Vec	0.64	0.69	0.77
Random Forest	BoW	0.57	0.54	0.16
	TF-IDF	0.55	0.53	0.16
	Doc2Vec	0.63	0.66	0.60
SGD	BoW	0.70	0.73	0.75
	TF-IDF	**0.77**	**0.81**	**0.81**
	Doc2Vec	0.68	0.72	0.76
SVM	BoW	0.71	0.75	0.78
	TF-IDF	**0.77**	0.80	**0.81**
	Doc2Vec	0.68	0.74	0.78
CNN	FastText	0.65	0.72	0.74
Bi-LSTM	FastText	**0.69**	**0.79**	**0.80**
BERT Multilingue	-	**0.81**	**0.85**	**0.86**
CamemBERT	-	0.78	0.83	**0.86**

La méthode *Random Forest* obtient les plus mauvais résultats (entre 16% et 66%) et ce quelle que soit la méthode de vectorisation et l'échantillonnage utilisés. Les résultats de la méthode *Naive Bayes* vont de 44% à 74% et présentent une très forte dépendance à l'échantillonnage pour la vectorisation *TF-IDF*. Les méthodes *Logistic regression*, *SGD* et *SVM* obtiennent des résultats très semblables et les meilleurs sont atteints avec la vectorisation *TF-IDF* (environ 80%). De manière surprenante, le plongement de mot *Doc2Vec* produit des résultats légèrement inférieurs à la représentation *BoW*. Les scores augmentent avec la taille de l'échantillonnage, ce qui explique pourquoi cette approche a besoin d'un grand jeu de données pour bien fonctionner. Les approches d'AP utilisant des réseaux de neurones (*CNN* et *BiLSTM*) obtiennent des scores comparables (entre 65% et 80%) mais restent tout de même en dessous des meilleures méthodes d'AA traditionnelles associées à la vectorisation *TF-IDF*. Le rééquilibrage des classes a un effet négatif plus important que pour les méthodes traditionnelles à cause de la réduction du jeu de données qu'il occasionne. Les modèles de langue préentraînés tel que *BERT* Multilingue et *CamemBERT* obtiennent des résultats encore meilleurs de peu aux méthodes classiques d'AA, offrant des performances très comparables avec 86% pour la meilleure F-mesure. Les résultats de *BERT* Multilingue et *CamemBERT* sont très proches pour ce qui est de la moyenne globale mais diffèrent un peu sur chaque classe, en fonction de la taille d'échantillonnage.

La figure 4.1 montre les F-mesures obtenus pour *BERT* Multilingue (courbe bleue) et *CamemBERT* (courbe orange) pour chacune des classes (les courbes grises correspondent aux autres méthodes). Les classes sont triées de gauche à droite par ordre de prévalence dans l'échantillon (c'est-à-dire dans le même ordre que l'histogramme de la figure 3.34 à la page 83). Les deux méthodes obtiennent des scores similaires pour la plupart des classes et qui ne diffèrent que sur les classes avec très peu d'articles. Par exemple, la classe *Économie domestique* n'obtient que 44% avec *BERT* mais 61% avec *CamemBERT*. Pour la *Politique* au contraire, *BERT* parvient à atteindre 53% alors que *CamemBERT* plonge à 8%; de même sur la *Minéralogie* (70% contre 37%) et le *Spectacle* (62% contre 0%, *Camem-*

BERT faisant l'impasse complète sur cette classe). Mais comme les écarts les plus importants entre ces résultats sont sur des classes sous-représentées, l'impact sur la moyenne globale est très petit et se retrouve compensé par des écarts pourtant plus réduits mais sur des classes plus nombreuses. Cette figure met aussi en évidence le fait que la classe *Arts et métiers* pose des difficultés à toutes les méthodes sans exception : toutes obtiennent des scores inférieurs à 40%, vraisemblablement parce que ce domaine est souvent confondu avec les *Métiers*. L'ambiguïté du système de domaines et les ressemblances entre classes sont discutées plus en détail à la section 4.2. D'une manière générale, augmenter la taille des échantillons par classe (c'est-à-dire le nombre maximum d'articles pour chacune) et donc accroître l'écart entre les classes les moins peuplées et celles les plus peuplées améliore les résultats, mais détériore ceux obtenus sur les classes moins représentées. Cela suggère qu'il est préférable d'avoir plus de données même avec des classes déséquilibrées que moins de données avec des classes bien équilibrées si le but est d'améliorer les résultats pour un grand nombre de classes différentes.

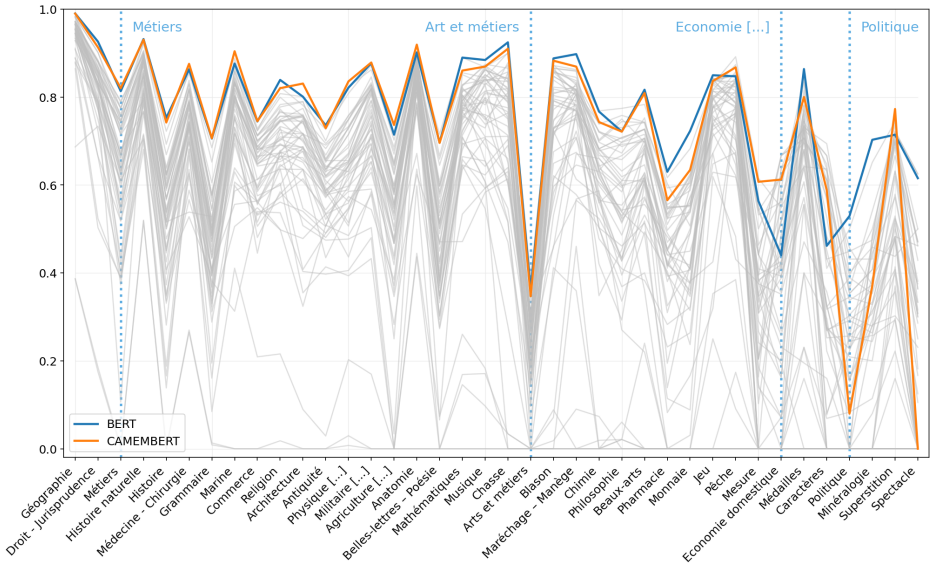


Figure 4.1 – F-scores obtenus par BERT Multilingual et CamemBERT sur chaque classe

Deux méthodes présentent un comportement différent : *Random Forest* et *Naive Bayes*, lorsqu'elles sont combinées avec la vectorisation *TF-IDF*. Dans le cas de *Naive Bayes*, la figure 4.2 montre clairement que la classification obtient de très mauvais résultats sans échantillonnage (courbe verte), et la stratégie d'échantillonnage a un impact positif sur les performances. Sans échantillonnage, les performances se dégradent très rapidement avec la diminution du nombre d'articles par classe et le modèle ne fonctionne que pour les celles qui sont les plus peuplées (à gauche de la figure). Le modèle entraîné avec des classes comprenant au plus 1 500 articles (courbe orange) fonctionne plutôt bien sur toutes les classes ayant plus de 500 articles. Enfin, seul le modèle avec le jeu d'entraînement le plus équilibré (avec 500 articles au plus par classe, courbe bleue) parvient à classer correctement la majorité des classes et ses performances ne chutent dramatiquement que sur les classes avec moins de 200 articles dans toute l'*EDdA* (de *Mesure* à *Spectacle*).

Parmi les méthodes traditionnelles d'AA, la vectorisation *TF-IDF* obtient quasi-systématiquement de meilleurs résultats que les *BoW* et *Doc2Vec*. Cela reste vrai aussi bien sur les F-mesures globales que sur les performances obtenues sur chaque classe. La figure 4.3 montre ainsi les scores obtenus par le classifieur *SGD* sur chacune des classes avec les trois différentes vectorisations testées et dans chaque cas sans échantillonnage. Les résultats sont très voisins mais le modèle *TF-IDF* (courbe orange) est légèrement au-dessus pour la plupart des classes.

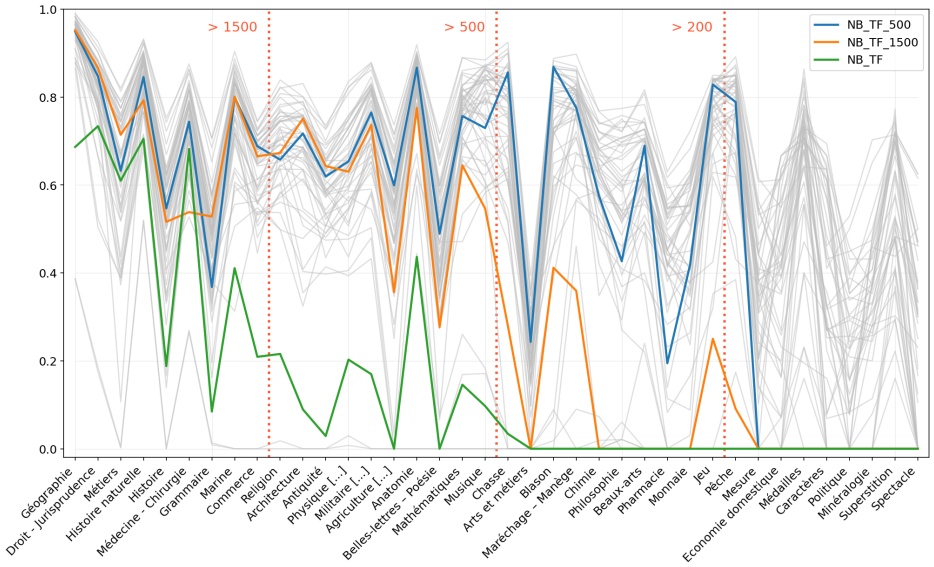


Figure 4.2 – F-mesures obtenues par *Naive Bayes + TF-IDF* sur chaque classe avec trois échantillonnages différents

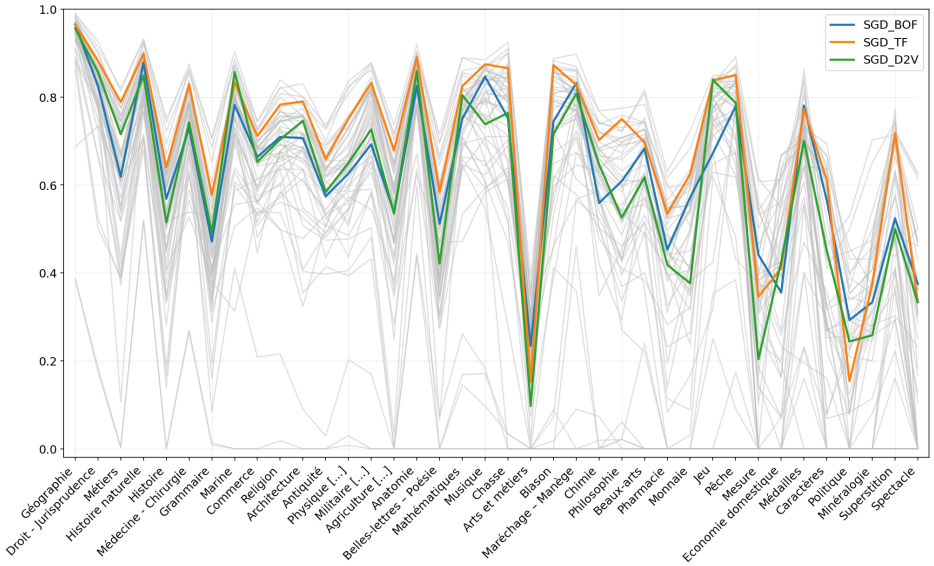


Figure 4.3 – F-mesures obtenues par *SGD* sur chaque classe avec trois vectorisations différentes et sans échantillonnage.

Le tableau 4.3 et la figure 4.4 présentent les différences de résultats obtenus en termes de F-mesure sur le jeu de test (sans échantillonnage) pour toutes les classes (triées par le nombre d'articles dans chacune) pour (1) *SGD+TF-IDF*, (2) *BiLSTM+FastText* et (3) *BERT* Multilingue. *BERT* obtient une F-mesure supérieure à 70% sur 31 des 38 classes au total que comporte le jeu des domaines regroupés (pour *SGD+TF-IDF* 25 classes, et *BiLSTM+FastText* seulement 19). Les performances du modèle *BERT* sont inférieures à 50% pour seulement 3 classes (5 avec *SGD+TF-IDF* et 10 avec *BiLSTM+FastText*). En général, les classes avec le plus d'articles (plus de 1 000) sont très bien reconnues par ces trois modèles. Pour les classes avec le moins d'articles (moins de 500), il y a une chute significative des performances. Il y a quelques exceptions comme *Pêche* et *Médailles*. Ces deux classes ont très peu d'articles dans le jeu d'entraînement (168 et 94 respectivement) mais sont étonnamment bien classées (à 85% et 86% par *BERT*).

Tableau 4.3 – F-mesures par classe obtenues sur le jeu de test par les combinaisons *SGD + TF-IDF* (1), *BiLSTM + FastText* (2) et *BERT* Multilingue (3).

Domaine regroupé (classe)	#	(1)	(2)	(3)	Domaine regroupé (classe)	#	(1)	(2)	(3)
Géographie	2 621	0.96	0.98	0.99	Chasse	116	0.87	0.87	0.92
Droit - Jurisp.	1 284	0.88	0.90	0.93	Arts et mét.	112	0.15	0.27	0.36
Métiers	1 051	0.79	0.76	0.81	Blason	108	0.87	0.86	0.89
Histoire naturelle	963	0.90	0.87	0.93	Maréchage	105	0.83	0.86	0.90
Histoire	616	0.64	0.64	0.75	Chimie	104	0.70	0.58	0.77
Médecine	455	0.83	0.80	0.86	Philosophie	94	0.75	0.49	0.72
Grammaire	452	0.58	0.54	0.71	Beaux-arts	86	0.70	0.62	0.82
Marine	415	0.83	0.86	0.88	Pharmacie	65	0.53	0.38	0.63
Commerce	376	0.71	0.69	0.74	Monnaie	63	0.63	0.50	0.72
Religion	328	0.78	0.77	0.84	Jeu	56	0.84	0.74	0.85
Architecture	278	0.79	0.74	0.80	Pêche	42	0.85	0.84	0.85
Antiquité	272	0.66	0.68	0.74	Mesure	37	0.35	0.10	0.56
Physique	265	0.75	0.76	0.82	Économie dom.	27	0.41	0.48	0.44
Militaire	258	0.83	0.82	0.88	Caractères	23	0.61	0.08	0.46
Agriculture	233	0.68	0.58	0.71	Médailles	23	0.77	0.70	0.86
Anatomie	215	0.89	0.84	0.90	Politique	23	0.15	0.22	0.53
Belles-lettres	206	0.58	0.41	0.70	Minéralogie	22	0.38	0.39	0.70
Mathématiques	140	0.82	0.85	0.89	Superstition	22	0.72	0.48	0.71
Musique	137	0.87	0.83	0.88	Spectacle	9	0.33	0.46	0.61

Au-delà de l'importance du nombre d'articles par classes, ces résultats soulignent la difficulté à distinguer entre certaines classes pour des raisons lexicales ou sémantiques comme cela était le cas pour *Arts et métiers* avec *Métiers*. Ce domaine, parmi les mieux représentés, semble attirer à lui des articles de nombreuses classes moins fréquentes mais proches sémantiquement. Cette hypothèse semble confirmée par la matrice de confusion de la figure 4.5.

Sur cette figure, une colonne pâle est en effet visible vers la droite. Cette colonne représente des articles de nombreux domaines pour lesquels le modèle *SGD+TF-IDF* a prédit à tort la classe *Métiers*. Les deux points plus colorés de cette colonne hors de la diagonale correspondent aux classes *Arts et métiers* et *Économie domestique*. Ils sont même les plus colorés des lignes correspondant à ces classes, ce qui signifie que la majorité des articles de ces deux classes ont été confondus avec des articles de *Métiers* par le modèle. De la même manière, *Mesure*, *Minéralogie*, *Pharmacie* et *Politique* ont été confondues avec *Commerce*, *Histoire naturelle*, *Médecine - Chirurgie* et *Droit - Jurisprudence* respectivement. Les ressemblances sémantiques entre ces classes illustrent les difficultés que rencontre un

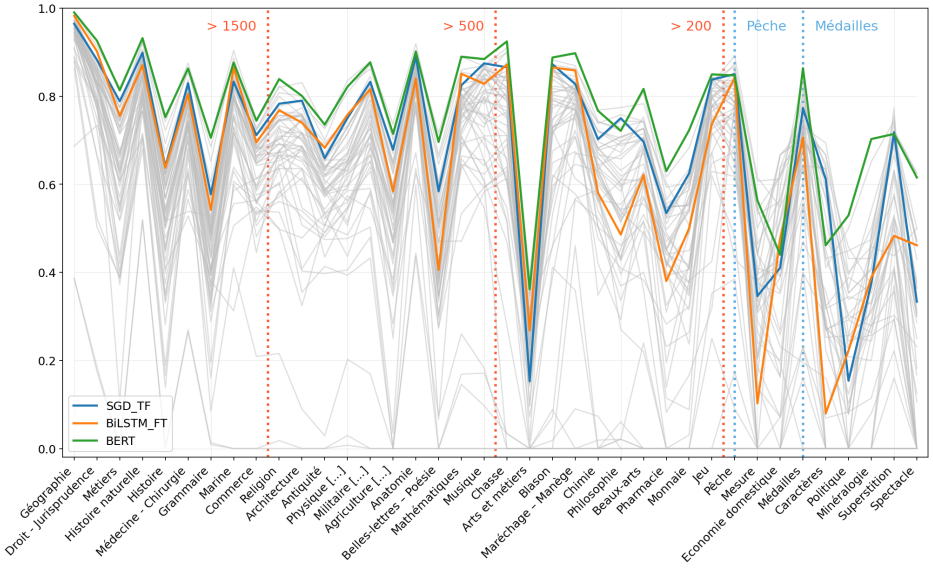


Figure 4.4 – F-mesures par classe obtenus sur le jeu de test par les combinaisons *SGD + TF-IDF*, *BiLSTM + FastText* and *BERT* Multilingue.

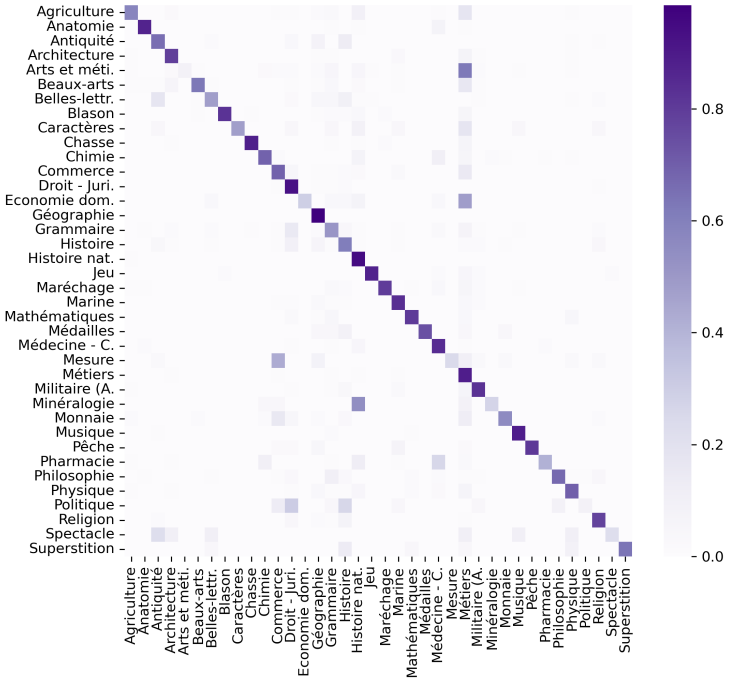


Figure 4.5 – Matrice de confusion matrix pour la combinaison *SGD+TF-IDF* sur le jeu de test.

modèle pour choisir la classe qui convient le mieux à un article donné. Les résultats confirment que là où il y une forte proximité sémantique le modèle tend à choisir la classe la mieux représentée dans le jeu de données, privilégiant ainsi les domaines regroupés qui contiennent le plus d'articles.

4.1.4 Classification non supervisée

En complément des méthodes de classification supervisées, des tentatives ont été faites pour utiliser de l'apprentissage non supervisé, permettant d'étudier autrement les relations entre les contenus des articles, les systèmes de classification et les prédictions générées automatiquement. Des techniques de *clustering* ont ainsi été employées pour grouper automatiquement les articles suivant leurs similarités, sans s'attacher aux étiquettes à appliquer à ces groupes (c'est-à-dire à une quelconque interprétation qu'il faudrait faire de ces groupes). Cette «similarité» est basée sur un calcul de distance entre les représentations vectorielles des articles. Pour cette expérience, la méthode des *K-Means* a été testée avec une vectorisation *TF-IDF*. La première expérience consiste à entraîner un modèle de *clustering* pour construire 38 classes (le nombre de domaines regroupés choisis précédemment). Des résultats décevants amènent à rechercher automatiquement le meilleur nombre de clusters en utilisant la méthode Silhouette (Shahapure & Nicholas, 2020). Les résultats suggèrent 36 comme le nombre optimal de clusters, mais la répétition de l'expérience précédente avec ce nombre produit des résultats similaires.

La figure 4.6 montre une carte de chaleur des distributions normalisées des nombres de clusters trouvés automatiquement avec la méthode *K-Means* en compartimentant les articles en 36 clusters. La plupart des clusters comportent un nombre important d'articles de plusieurs domaines et le cluster n°0 regroupe des articles de quasiment tous les domaines. De plus, pour beaucoup de clusters, la proportion est élevée et, de la même manière, beaucoup de domaines se retrouvent étalés sur plusieurs clusters. Cela est particulièrement vrai pour la *Géographie*, l'*Histoire naturelle* et les *Arts et métiers*. Bien que les résultats soient complexes, l'analyse des clusters demeure utile : par exemple le cluster n°22 regroupe des articles étiquetés *Belles-lettres - Poésie*, *Histoire*, *Médailles*, *Religion* et *Superstition*. Ces catégories possèdent des similarités thématiques qui dénotent leur classification initiale dans l'*EDdA* ainsi que celle choisie en termes de domaine regroupé (elles contiennent fréquemment du contenu à propos d'histoires réelles mais aussi fictionnelles et mythologiques). De manière semblable, le cluster n°7 comporte des articles des domaines *Médecine - Chirurgie*, *Anatomie*, *Physique - [Sciences physico-mathématiques]* et *Pharmacie*. Il semble possible de percevoir une thématique autour du corps et de la médecine dans cet ensemble de classes. Enfin, le cluster n°30 en rassemblant des articles de *Commerce* et de *Mesure* suggère une thématique autour des nombres et d'unités. Quelques clusters restent particulièrement homogènes et donc proches des classes comme le cluster n°10 composé à 96.46% d'articles des classes *Géographie* ou le n°34, à 99.05% de l'*Histoire naturelle*.

À l'inverse, la figure 4.7 montre la distribution de domaine regroupés par cluster. Comme sur la figure 4.6, le cluster n°0 ressort particulièrement en rassemblant plus de 14 000 articles de quasiment toutes les classes, alors que la très large majorité des autres comporte moins de 2 000 articles. Cet écart considérable entre ce cluster et les autres illustre toute la difficulté qu'il y a à trouver automatiquement une catégorisation satisfaisante des articles du jeu de données. Cette nouvelle figure montre encore l'hétérogénité des classes de quasiment tous les clusters (exceptés les cas discutés plus haut) et par là-même la difficulté à donner un sens aux clusters proposés par *K-Means*, sinon en terme de thématiques pour quelques uns d'entre eux.

De plus, sur les 38 classes, seules 15 apparaissent comme la classe principale d'un des clusters (voir le tableau 4.4). Parmi les clusters, 9 apparaissent relativement homogènes avec une classe principale «pure» dans ces clusters à plus de 90% (cela est le cas pour les clusters n°1, 6, 10, 13, 15, 19, 31, 33 et 34). Mais pour ces 9 clusters, seules 3 classes apparaissent comme la classe principale, et la *Géographie*

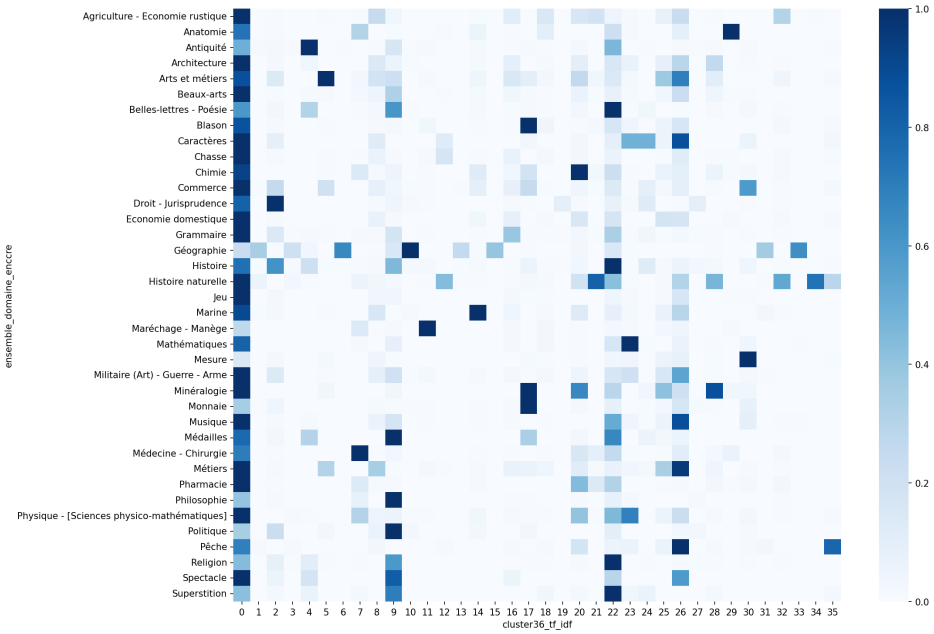


Figure 4.6 – Distributions normalisées des nombres de clusters par classe.

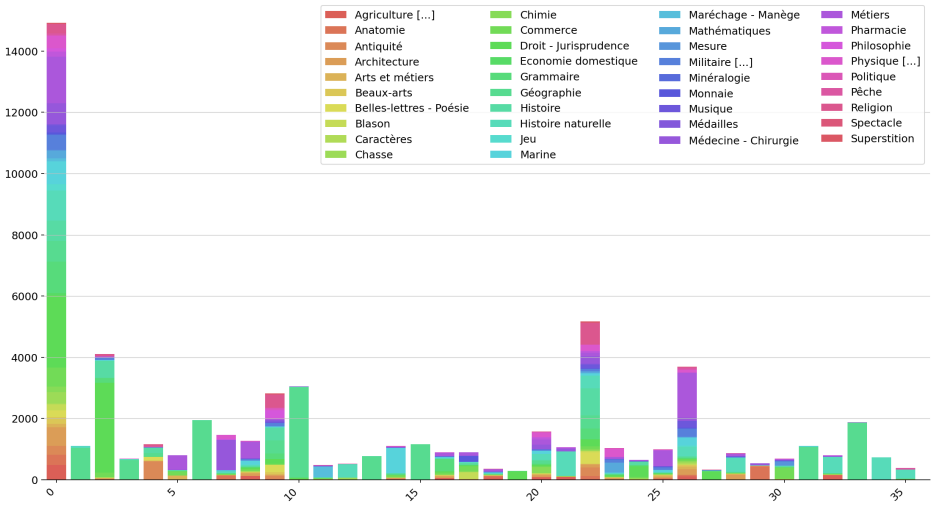


Figure 4.7 – Nombre d'articles de chaque classe par cluster

est la classe principale de 7 d'entre eux. Cela empêche d'interpréter la pureté importante des clusters comme la correspondance de certains d'entre eux avec des classes existantes, une hypothèse qui était pourtant à priori intéressante au vu de la proximité entre le nombre de classes (38) et de clusters (36). De plus, la classe *Géographie* est présente dans 31 clusters différents, c'est-à-dire la quasi-totalité d'entre eux, ce qui contredit une interprétation des 7 clusters comme autant de types d'articles de *Géographie* possibles.

Tableau 4.4 – Part de la classe la plus représentée dans chaque cluster.

Cluster n°	Classe principale	Part	Cluster n°	Classe principale	Part
0	Droit - Jurisprudence	15.85 %	18	Métiers	20.51 %
1	Géographie	91.95 %	19	Droit - Jurisprudence	95.50 %
2	Droit - Jurisprudence	71.54 %	20	Chimie	12.33 %
3	Géographie	89.03 %	21	Histoire naturelle	74.48 %
4	Antiquité	51.38 %	22	Histoire	16.99 %
5	Métiers	58.78 %	23	Mathématiques	30.95 %
6	Géographie	99.28 %	24	Droit - Jurisprudence	52.67 %
7	Médecine - Chirurgie	66.10 %	25	Métiers	51.76 %
8	Métiers	42.17 %	26	Métiers	39.85 %
9	Géographie	15.25 %	27	Droit - Jurisprudence	82.77 %
10	Géographie	96.46 %	28	Histoire naturelle	52.82 %
11	Maréchage - Manège	70.80 %	29	Anatomie	82.93 %
12	Histoire naturelle	83.24 %	30	Commerce	52.82 %
13	Géographie	96.53 %	31	Géographie	95.50 %
14	Marine	74.93 %	32	Histoire naturelle	65.09 %
15	Géographie	100.00 %	33	Géographie	99.47 %
16	Grammaire	44.65 %	34	Histoire naturelle	99.05 %
17	Blason	24.06 %	35	Histoire naturelle	71.61 %

Il est difficile de tirer des conclusions des résultats de cette expérience de *clustering* et, avant d'aller plus loin, davantage d'analyses qualitatives sont nécessaires pour pouvoir décider si l'apprentissage non supervisé peut être utile comme une méthode complémentaire pour classer automatiquement les articles de l'*EDdA* et à terme de *LGE*. Le *clustering* est, comme le *topic modeling* par LDA, une manière d'organiser le contenu du corpus sans présupposer un ensemble de classes qui serait *la* manière correcte de structurer la connaissance. Le choix fait précédemment de reprendre le nombre de domaines regroupés utilisés pour la classification supervisée pour le nombre de clusters biaisait nécessairement la découverte de clusters, mais le fait que l'heuristique Silhouette ait trouvé un nombre proche de clusters (36) demeure troublant. Il pourrait être intéressant dans de futurs travaux de tester de manière systématique un bien plus grand nombre de possibilités en partant d'un très petit nombre de clusters et en augmentant leur nombre tout en évaluant manuellement la pertinence du système trouvé à chaque étape. Dans le contexte de cette thèse, ces méthodes n'ont pas permis d'avancées significatives et ne sont pas retenues dans le reste des analyses.

4.2 Relations entre *Géographie* et autres domaines de connaissance

Après avoir tâché de tirer les meilleures performances des différents modèles de classification à la section précédente (4.1), cette section s'intéresse au contraire aux erreurs qui subsistent dans leurs prédictions. Faisant en effet l'hypothèse que les erreurs commises par le modèle reflètent dans une certaine mesure les proximités qui existent entre les domaines de connaissance elle se propose ainsi de comprendre les relations qui lient la *Géographie* aux autres domaines de connaissance.

Deux tendances gouvernent le choix d'un modèle dont étudier les erreurs. Il faut qu'elles soient assez nombreuses pour fournir matière à observation et donc ne pas prendre le meilleur, mais à l'inverse il est vraisemblable que les prédictions d'un classifieur obtenant de trop mauvais résultats recèlent davantage de bruit que de vérités profondes sur les contenus des articles. Pour ces deux raisons, c'est le modèle *SGD+TF-IDF* entraîné sur tous les articles disponibles (colonne n°3 du tableau 4.2 p.95) qui a été retenu pour cette étude et sera utilisé tout au long de cette section. Ce modèle parvient à classer correctement 9 630 des 11 702 articles, soit 82%. Ce sont les 2 072 autres, les erreurs, qui intéressent cette étude.

4.2.1 Observations préliminaires des erreurs

Pour comprendre les facteurs amenant à ces erreurs il faut identifier clairement les paramètres en jeu, représentés schématiquement par la figure 4.8.

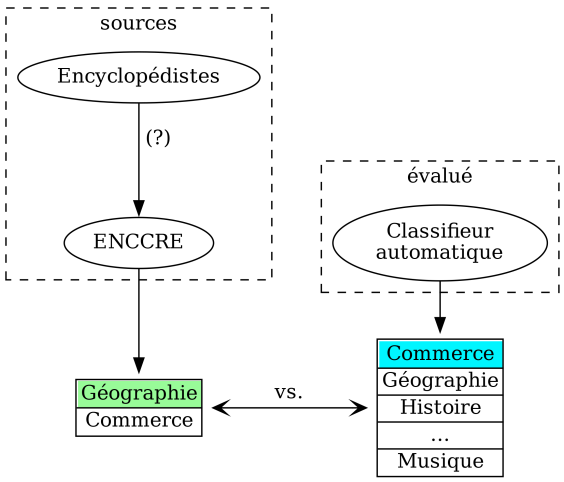


Figure 4.8 – Les paramètres en jeu lors de la classification automatique d'un article

Premièrement trois entités différentes (les ellipses sur la figure 4.8) associent des classes aux articles. La première et la seule qui puisse constituer une vérité indiscutable est le choix des éditeurs de l'*EDdA* au travers de l'emploi de désignants. L'ENCCRE les normalise et, en leur absence, émet parfois sa propre classification basée sur des indices présents dans le texte (notamment les *marques de domaine* décrites à la section 3.1.2 p.49). À la différence de ces deux sources, le modèle est une troisième origine de classes pour les articles qui, elle, peut – et doit – être remise en cause puisqu'il s'agit d'accorder ses sorties aux deux sources de vérité terrain utilisées pour l'entraîner puis l'évaluer. Les erreurs étudiées dans cette partie correspondent aux configurations possibles de différences entre ces trois origines des classes, sur un ou plusieurs domaines.

Après la question des origines, un deuxième paramètre s'ajoute dans l'équation. Tous les travaux de la section 4.1 considèrent en effet un problème de classification monoclasse : pour les 5% d'articles assignés à plusieurs domaines par les encyclopédistes (voir la section 4.1.2 p.93), une seule classe a été retenue pour entraîner le modèle (correspondant à la cellule en vert sur la figure). Or d'une part il ne semble pas y avoir d'ordre d'importance entre les désignants présents en tête d'un article, et d'autre part les combinaisons de classes apparaissant ensemble au sein de ces 5% possèdent une certaine cohérence. Il y a ainsi plusieurs articles relevant à la fois de l'*Histoire Naturelle* et de la *Botanique*[5] mais aucun relevant à la fois de la *Pharmacie* et du *Spectacle*. Malgré le choix d'une classe parmi

5. l'article BROMELIA (L'Encyclopédie, T2, p.434) présenté à la figure 3.1a p.48 dans la description des parties d'un article d'encyclopédie en offre un exemple

plusieurs pour simplifier la tâche du classifieur automatique, les combinaisons de classe existantes sont également une source d'information intéressante.

Ce problème posé par le choix d'une classe unique à partir d'un ensemble de classes existe non seulement du côté de la vérité terrain mais également du côté de la prédiction. Techniquement, un classifieur fait ses prédictions en attribuant à chaque classe du jeu d'étiquettes une probabilité d'être correcte et en retenant la classe qui obtient la plus élevée (en bleu sur la figure 4.8). Or, pour plus de la moitié des erreurs du modèle – 1 082, soit 9% du total – la classe qu'il fallait prédire a tout de même obtenu la deuxième plus haute vraisemblance, et pour les 990 autres, elle apparaît souvent, même avec un rang plus élevé ($3^{ème}$ ou $4^{ème}$). Ainsi, même quand il se trompe, le modèle ne donne pas des réponses entièrement absurdes et il y a donc un intérêt à considérer ses erreurs pour comprendre les similarités et les différences entre les groupes de domaines.

À la lumière des remarques ci-dessus, et pour mieux comprendre les erreurs du modèle, il peut donc être intéressant de confronter aux deux sources de vérités possibles que sont les encyclopédistes et l'ENCCRE non seulement la prédiction retenue par le modèle (celle de plus haute probabilité) mais aussi les suivantes, tout en tenant compte du fait que certains articles relèvent en réalité de plusieurs domaines à la fois. Cela permettrait d'apporter de l'information sur la nature intriquée de l'organisation de la connaissance dans l'*EDdA* et permettrait d'«aller au-delà du système de classification originel des éditeurs pour commencer à tracer les contours des pratiques discursives à plusieurs niveaux qui contribuent à la riche texture dialogique de l'*EDdA*» (Roe et al., 2016, p. 10).

En un sens, cette étude conduite sur la seule *EDdA* s'inscrit entièrement dans la perspective de d'Alembert citée à la section 3.3.2 en considérant la classification automatique comme une «projection» utile pour visualiser les articles : ceux dont la classe prédite correspond à la classe qui leur avait été assignée par les éditeurs et ceux pour lesquels elle ne correspond pas qui sont tout aussi intéressants. Mais il y a une tension indépassable entre la tâche d'évaluer une classification générée automatiquement et l'accès à une compréhension qu'il n'y a pas forcément une classification unique et «correcte» pour un article donné.

Les erreurs autour de la *Géographie*

Un point de départ simple pour cette étude consiste à étudier les faux négatifs et les faux positifs des articles de *Géographie*. Sur les 2 621 articles de cette classe présents dans le jeu de test, 191 ont été mal classés par le modèle.

Faux négatifs Seuls 39 articles étiquetés *Géographie* ont été classés à un autre domaine que *Géographie* par le modèle. Ils constituent donc des faux négatifs du modèle *SGD+TF-IDF* pour cette classe, dont la répartition est visible à la figure 4.9. En tout, 12 domaines autres que *Géographie* sont prédits parmi les 39 faux négatifs, avec une forte prévalence de l'*Histoire* et de l'*Antiquité*. Ces faux négatifs pour la classe *Géographie* peuvent se regrouper en trois types.

Le premier, celui des «faux multiclasses» correspond au cas des articles ayant un seul désignant et pour lesquels le modèle n'a attribué que la deuxième ou troisième plus haute probabilité à la classe qu'il fallait prédire. Ce type de faux négatif correspond par exemple à l'article INDOUSTAN (L'Encyclopédie, T8, p.686). Il ne possède que le désignant «Géog.» et fait donc partie du domaine ENCCRE *Géographie*. La meilleure prédiction pour cet article est *Histoire* (avec 50% de certitude), et le domaine qu'il fallait prédire, *Géographie* vient ensuite avec 47,6%. En regardant son contenu on observe, après 5 lignes qui situent l'«Indoustan» dans l'espace, un paragraphe plus long, de 12 lignes au total, qui trace un historique de la région. Si le modèle a donc bel et bien échoué à reproduire la logique éditoriale de l'*EDdA*, sa prédiction est tout à fait satisfaisante du point de vue de la compréhension du texte qu'on pourrait attendre d'un lectorat humain et il n'aurait donc pas été surprenant

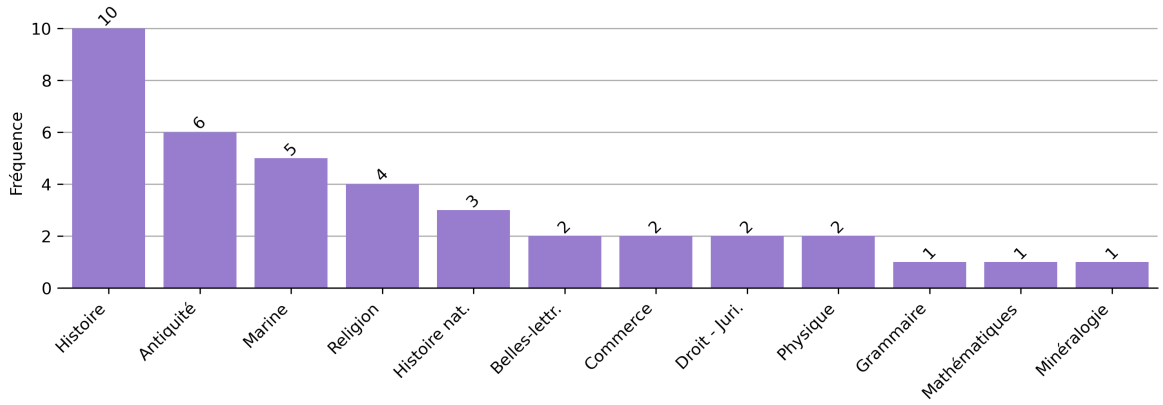


Figure 4.9 – Articles étiquetés *Géographie* par l'ENCCRE mais pas par le modèle (faux négatifs)

que l'article soit pourvu d'un désignant pour l'*Histoire* en plus de celui pour la *Géographie*. Les articles de toponymes de l'*EDdA* comportent très fréquemment une multitude de contenus qui ne sont pas nécessairement «géographiques», et cela est particulièrement le cas sous la plume de Jaucourt qui signe cet article. La prévalence de l'*Histoire* sur la figure 4.9 est un premier signe de sa proximité avec la *Géographie*.

Le deuxième type de faux négatifs, celui des «désaccords» inclut des articles auxquels les auteurs de l'*EDdA* n'avaient à l'origine pas assigné de domaine de connaissance et que l'ENCCRE a classés à *Géographie* mais pour lesquels *Géographie* ne fait pas partie des prédictions les plus probables du modèle. L'article de Diderot sur les îles des AÇORES (L'Encyclopédie, T1, p.110) présente d'après le modèle une très faible probabilité de relever de la *Géographie* (la classe est prédite avec seulement 2.7% de confiance). Les classes les plus probables d'après le modèle sont *Commerce* (45%), suivie de *Métiers* (43%) et *Arts et métiers* (8.6%). Et en effet, en allant lire l'entrée *AÇORES* elle-même, on constate qu'après une seule ligne situant l'archipel à la fois très grossièrement («de l'Amérique») et trop précisément (par des coordonnées en longitude et latitude mais sans donner aucune relation à d'autres terres que le lectorat serait susceptible de connaître), la majeure partie de l'article est consacrée à décrire en détail son intérêt pour le commerce, comme un point de passage sur des routes commerciales, et en énumérant longuement toutes les marchandises qui en proviennent ou y transitent. Sur 1 597 articles *Commerce*, seuls 3 possèdent aussi la classe *Géographie* selon les sources de domaines (colonne de gauche sur la figure 4.8). Ce type de faux négatifs révèle en fait les liens qui existent entre certains contenus, ici la *Géographie* et le *Commerce*. Alors qu'il n'apparaît pas explicitement dans le «Système figuré» de l'*EDdA* (voir figure 3.4 p.52), ce lien est pourtant assez fréquent sous la plume de Diderot, qui malgré la rigueur assumée de son style dans les articles de *Géographie* (Laramée, 2017, p. 168) intègre volontiers des «considérations sur le commerce local» (Vigier et al., 2022, p. 69).

La dernière sorte de faux négatifs, à l'inverse du premier type (et qu'on pourra donc naturellement nommer le type des «vrais multiclasses»), regroupe un ensemble d'articles qui possédaient plusieurs désignants, devenus plusieurs ensembles de domaines après les travaux de l'ENCCRE et dont le *domaine regroupé* unique choisi pour l'étude était *Géographie*. Par exemple GOLGOTHA relevait initialement de la *Géographie* et de la *Théologie* d'après les éditeurs, correspondant aux ensembles de domaines *Géographie* et *Religion* pour l'ENCCRE, mais la seule classe *Géographie* a été retenue comme vérité pour évaluer le modèle. Les deux classes les plus probables d'après celui-ci sont *Religion* (60%) et *Histoire* (31%)). De manière surpenante, *Géographie* n'obtient que 3%. Le nom «Golgotha» réfère à l'endroit où le messie de la religion chrétienne a été crucifié et l'article, fidèle au reste de la ligne éditoriale de l'*EDdA*, consiste en une comparaison des écrits des premiers chrétiens sur l'endroit. Il

se termine en rejetant les théories les plus imaginatives sur l'origine du nom. L'auteur, l'abbé Mallet, tranche en faveur de la suggestion de Saint Jérôme selon laquelle le Golgotha aurait été nommé ainsi (*calvaire* en Hébreu) car les crânes étaient abandonnés sur ce lieu d'exécution traditionnel. Sur cet exemple il semble que l'erreur du modèle révèle surtout en fait les limites d'un choix automatique fait lors de la préparation du corpus d'une classe parmi plusieurs. Ce dernier type des «vrais multi-classes» pourrait certainement intéresser une étude séparée sur les différences entre contributeurs dans leurs approches à la *Géographie* et aux autres domaines de connaissance. Pour l'entrée GOLGOTHA, il est ainsi intéressant de noter que son auteur, l'abbé Mallet est un théologien qui a rédigé de nombreux articles (plus de 2 000 en tout), mais seulement 7 en *Géographie*. Pour d'autres articles, l'approche propre à un autre contributeur, le chevalier de Jaucourt, est développée à la section 5.2 à partir de la page 157.

Faux positifs Pour 152 articles, le modèle a prédit *Géographie* comme la classe la plus probable alors qu'ils ne relevaient de ce domaine ni d'après ses auteurs ni d'après l'ENCCRE. L'article ROCHER (L'Encyclopédie, T14, p.313) visible à la figure 4.10, classé en *Géographie* par le modèle alors que son désignant le situe en *Grammaire*, constitue un exemple de faux positif. Extrêmement bref, il se contente d'établir l'équivalence sémantique entre les mots *rocher*, *roc* et *roche* et contient un renvoi vers l'entrée *ROC*.

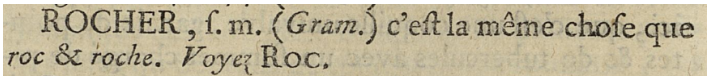

ROCHER, ſ. m. (*Gram.*) c'eſt la même choſe que *roc* & *roche*. *Voyez* ROC.

Figure 4.10 – Article ROCHER, l'*EDdA*, T14, p.313

Si l'on tient compte des prétraitements appliqués avant classification, en particulier l'élimination des mots vides, le modèle a même reçu comme entrée pour cet article :

rocher s. m. chose roc roche roc.

L'abondance de termes liés aux propriétés physiques du sol oriente vraisemblablement le modèle vers la *Géographie* comme classification la plus probable. Ici encore dans le contexte des faux positifs, le modèle révèle son utilité pour mettre en lumière des contenus thématiques qui resteraient cachés en s'en tenant à la seule classification des auteurs, ici du vocabulaire disciplinaire qui pourrait relever de la géographie. En général, les probabilités de prédiction pour un article donné décroissent très vite pour atteindre 0 après quelques classes (ce qui traduit la certitude d'après le modèle que l'article ne relève pas des classes suivantes) : il est remarquable que pour ROCHER le modèle attribue une vraisemblance non nulle à 16 classes différentes, traduisant l'ambiguïté d'un article très pauvre en éléments distinctifs et difficile à situer parmi les domaines de connaissance. Après *Géographie*, le modèle propose ainsi dans l'ordre la *Grammaire* (17%, ce qui correspondrait au premier type de faux négatifs des «faux multiclasses» identifiés plus haut si l'on conduisait pour la *Grammaire* l'étude conduite ici pour la *Géographie*), *Marine* (14%), et *Loi* (13%).

Les faux positifs se répartissent entre 23 domaines différents, soit presque deux fois plus que pour les faux négatifs. Les domaines impliqués dans les faux négatifs sont d'ailleurs tous inclus dans ces 23 domaines à l'exception de la *Minéralogie*, à l'origine d'un seul faux négatif. Leur distribution, visible à la figure 4.11, révèle une forte prépondérance de l'*Histoire* et de l'*Antiquité*, les deux classes les plus confondues avec la *Géographie*. On retrouve également les classes *Marine*, *Religion* et, dans une moindre mesure, *Histoire naturelle* dont seulement 3 articles ont été pris pour de la *Géographie* sur les 152 faux positifs – soit 2% – alors que les 3 articles de *Géographie* prédits comme appartenant à cette classe représentaient 7.7% des 39 faux négatifs et classaient l'*Histoire naturelle* en 5ème position des domaines les plus prédits à la place de la *Géographie* par le modèle. Plusieurs domaines présents mais assez rares parmi les faux négatifs (dans les 50% les moins représentés) se classent parmi les

faux positifs les plus fréquents (dans les 33% les mieux représentés) : *Droit - Jurisprudence* qui est le 3ème domaine causant le plus de faux positifs pour le modèle, *Physique*, *Grammaire* et *Belles-lettres*. Deux domaines totalement absents des faux négatifs, *Agriculture* et *Architecture*, sont plutôt bien représentés parmi les faux positifs puisqu'ils sont respectivement les 10ème et 11ème domaines les plus fréquents sur 23 domaines au total.

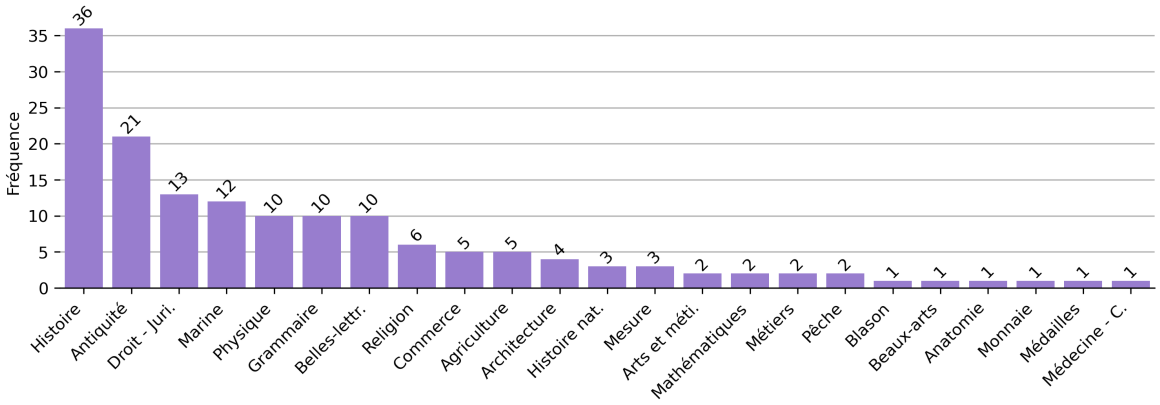


Figure 4.11 – Articles étiquetés *Géographie* par le modèle mais pas par l'ENCCRE (faux positifs)

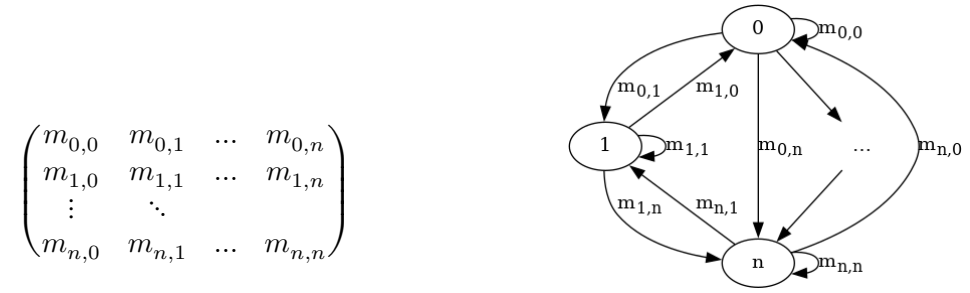
Les différences de distribution entre faux positifs et faux négatifs suggèrent une apparition fréquente de contenus géographiques en dehors du seul domaine de la *Géographie*, davantage que l'inverse. Ces contenus nourrissent les articles de nombreux autres domaines, montrant la Géographie comme une science transverse, nécessaire dans l'articulation du discours des autres sciences.

Graphe des Plus Proches Voisins

L'étude préliminaire qui précède révèle déjà des traits intéressants de la *Géographie* mais serait trop fastidieuse à conduire sur chacun des 38 *domaines regroupés* et, surtout, ignorerait des dynamiques complexes impliquant plus de deux domaines à la fois. Pour approfondir l'exploitation des erreurs du modèle, il faut des outils permettant d'adopter une approche d'ensemble.

Les matrices et les graphes sont des représentations complémentaires et équivalentes d'un même objet. À tout graphe dirigé fini et sans arêtes multiples correspond une matrice d'adjacence pondérée. Réciproquement toute matrice carrée dont les coefficients peuvent représenter les poids d'un graphe peut être vue comme sa matrice d'adjacence pondérée (ou matrice des poids). En effet, dans un graphe de n nœuds n'admettant qu'une seule arête entre deux nœuds donnés, il peut y avoir au plus n^2 arêtes. Si le graphe est complet ou si les coefficients qui pondèrent ses arêtes peuvent prendre une valeur particulière pour représenter une absence d'arête (typiquement 0), alors ces arêtes peuvent être stockées dans une matrice carrée de taille n : sa matrice des poids. Inversement, une matrice M de taille n permet de construire un graphe en créant n nœuds puis en ajoutant une arête pondérée du coefficient $M_{i,j}$ du nœud i vers le nœud j (de même que pour la traduction du graphe en matrice, en n'ajoutant éventuellement aucune arête si ce coefficient a une valeur particulière, typiquement 0) pour tous les coefficients de la matrice M. Cette équivalence est représentée par la figure 4.12.

Il faut remarquer ici la part d'arbitraire qui existe dans les deux opérations de conversion puisqu'elles reposent sur un choix implicite pour faire correspondre le sens d'une arête sur le graphe avec l'ordre entre colonne et ligne dans la matrice. Il aurait tout à fait été possible de stocker en $M_{j,i}$ le poids de l'arête de i vers j ou, de manière équivalente de construire à partir de $M_{i,j}$ l'arête de j vers i dans le graphe. La seule chose qui importe est d'adopter des conventions compatibles (ce qui est le cas de celles choisies) pour les deux transformations `matrice`→`graphe` et `graphe`→`matrice` afin que



(a) Les coefficients dans la matrice

(b) Les poids des arêtes du graphe

Figure 4.12 – Représentation schématique de l'équivalence entre matrice et graphe

leur composée revienne à la matrice ou au graphe de départ, et pas à la matrice transposée ou au graphe opposé (le même graphe en renversant le sens des arêtes).

Ces deux concepts représentent donc d'une façon complémentaire un même type d'objet : les systèmes dynamiques discrets et linéaires (SDDL). Ceux-ci comprennent par exemple les flux sur un réseau, une étape de transformation d'un modèle physique simplifié ou dans le cas présent la prédiction des classes d'un ensemble d'articles. Cela est clair en revenant à la définition d'une matrice de confusion : ses coefficients $m_{i,j}$ représentent la proportion d'objets de la classe i prédits par le modèle comme appartenant à la classe j. Dans l'évaluation d'un modèle sur un jeu de test, l'état initial est caractérisé par une certaine distribution d'éléments de chaque classe, qui constitue donc un premier vecteur. Après la prédiction du modèle, il y a une deuxième distribution qui correspond à la répartition des classes prédites et peut se représenter par un deuxième vecteur de même dimension que le premier (le nombre de classes possibles), et qui est aussi proche du premier que les performances du modèle sont bonnes. Le rapport entre ces deux vecteurs est la matrice de confusion du modèle, c'est-à-dire que le nombre d'objets dans une classe prédite (son cardinal) sera une combinaison linéaire des cardinaux des différentes classes de départ, pondérés par les coefficients de la matrice de confusion dans la colonne correspondant à la classe prédite. S'il est vrai que ce système ne possède que deux états (le modèle donne toujours la même prédiction sur une entrée donnée, il n'y a pas de sens à lui faire effectuer une nouvelle prédiction), il n'en constitue pas moins un SDDL et il est possible d'utiliser les techniques employées habituellement sur ces objets.

La dualité entre matrices et graphes permet d'explorer les SDDL en adoptant alternativement les points de vue d'un graphe ou d'une matrice. Lorsque le nombre de nœuds et d'arêtes devient trop élevé dans un graphe pour pouvoir exploiter sa représentation graphique (voire parfois même pour pouvoir simplement tracer une telle représentation en deux dimensions), l'exploitation de sa matrice des poids, qui dans tous les cas donne «seulement» un tableau bidimensionnel permet de prendre le relai. À l'inverse, il peut être compliqué d'obtenir une intuition de l'évolution d'un système à partir de sa seule matrice d'adjacence pondérée alors qu'un graphe traduit facilement une idée de dynamique.

Dans le cadre de cette étude, c'est donc à partir de la matrice de confusion normalisée, présentée à la figure 4.5 page 99 et notée C dans ce qui suit, que sera étudié le système correspondant aux prédictions du modèle *SGD+TF-IDF*. Suite à la remarque précédente, la convention choisie pour le sens des arêtes signifie que l'analyse est orientée vers ses faux négatifs (les lignes de C). La représentation de la matrice de confusion du modèle sous forme de graphe revient donc à le voir «déplacer» les articles du *domaine regroupé* auquel ils appartiennent vers le domaine qu'il considère le plus probable (cette visualisation comme un flux permet également de comprendre en quoi la situation est celle d'un SDDL puisque les analyses de flux en sont des exemples typiques). La convention contraire (basée sur les colonnes et donc les faux positifs) aurait signifié que les articles d'un domaine prédit par le modèle auraient été issus des vrais domaines pointés par les arêtes partant du nœud correspondant).

Elle semble moins intuitive, ce qui justifie la convention adoptée.

La matrice de confusion d'un modèle parfait est la matrice identité : des 1 sur la diagonale principale (du coin supérieur gauche au coin inférieur droit) et des 0 partout ailleurs, signifiant que pour chaque classe donnée tous les articles de cette classe sont correctement identifiés par le modèle («laissés au même endroit»). Avec l'équivalence discutée tout au long de cette partie, le graphe correspondant à un tel modèle parfait serait formé d'autant de nœuds qu'il y a de classes, tous isolés, et ne possédant chacun qu'une unique arête vers eux-mêmes (une boucle) pondérée de la valeur 1. Mais le graphe d'un vrai modèle comme celui étudié est bien plus difficile à représenter, comme le suggèrent les nombreuses cases colorées hors de la diagonale sur la figure 4.5. Il est d'ailleurs difficile de distinguer précisément quelles cases sont nulles ou pas à cause des différences d'ordre de grandeur dans les coefficients. En reprenant les données numériques, 484 cellules sont non nulles sur les 1 444 cellules au total dans la matrice (38×38). Ce «bruit de fond» sur la matrice correspondrait donc à un graphe extrêmement dense, avec $\frac{1}{3}$ des arêtes d'un graphe complet ($\frac{484}{1444} \simeq 0.335$). Il est donc difficile de se donner une intuition de la dynamique sous-jacente à cette matrice de confusion.

Pour dépasser cette difficulté, on simplifie le système considéré pour ne considérer que les erreurs les plus significatives du modèle. Cela revient à considérer la notion de «Plus Proche Voisin» (PPV au centre du problème des *K-Nearest Neighbors* (les «K plus proches voisins») sur le graphe des faux négatifs de sa matrice de confusion. La résolution de ce problème a des applications dans de nombreux domaines, des réseaux (Guo et al., 2023) à la vision par ordinateur (Bewley & Upcroft, 2013) où il met en jeu des nombres de points tellement grands qu'il faut des structures de données particulières pour accélérer la recherche comme les arbres de dimensions K (Friedman et al., 1977) ou trouver des heuristiques pour approcher la solution plus rapidement. Si la notion existe déjà, elle se trouve appliquée ici dans le cadre le plus simple possible au point de la rendre à peine reconnaissable : il ne s'agit de trouver qu'un unique voisin pour chaque nœud du graphe ($K = 1$) et il n'y a que 38 «points» (les classes des *domaines regroupés*). De plus, là où la plupart des points considérés dans les problèmes de *KNN* traités dans la littérature sont multidimensionnels et contiennent de grandes quantités de données parfois combinées de manière complexe, les seules grandeurs à comparer dans cette instance du problème sont des nombres flottants (les scores du modèle pour chaque paire de classe, entre 0 et 1) ce qui permet d'effectuer tous les calculs et de comparer les résultats pour obtenir *le* plus proche voisin de manière certaine. Ce cas particulier du problème se distingue aussi par le fait qu'il s'agit habituellement de trouver une distance minimale alors qu'il faut ici garder la classe dont le score dans la matrice de confusion est le plus grand.

Appliquée au problème de l'exploitation du graphe des faux négatifs de la matrice de confusion du modèle, la notion de PPV permet de considérer une matrice et un graphe représentant les faux négatifs les plus importants de chaque classe. Pour souligner la parenté avec le problème plus général, et pour pouvoir ensuite porter la notion à d'autres objets que des matrices et des graphes issus des mesures de performance d'un classifieur automatique, ces deux objets seront nommés respectivement MPPV (Matrice des Plus Proches Voisins) et GPPV (Graphe des Plus Proches Voisins) du modèle. La MPPV est la matrice d'adjacence pondérée du GPPV. Puisqu'il est possible de résoudre le problème de PPV complètement et de manière naïve dans le cas présent, il est en pratique plus aisé de commencer par décrire la MPPV à partir de la matrice de confusion du modèle, puis de la considérer comme une matrice d'adjacence pour générer le GPPV correspondant. Pour obtenir la MPPV à partir de la matrice de confusion, il suffit de mettre à 0 toutes les cellules diagonales (celles qui d'après la remarque ci-dessus encodent les prédictions exactes du modèle qui ne contiennent donc pas d'information sur ses erreurs) puis de calculer la valeur maximale sur chaque ligne et, pour une ligne donnée, rendre également nulles toutes les cellules sauf celle(s) où le maximum a été atteint. Ici encore, la simplification a lieu par ligne puisque ce sont les lignes qui régissent le graphe : ce choix correspond à ne garder que l'arête la (ou les arêtes les) plus forte(s) qui quitte(nt) chaque nœud, c'est-à-dire le domaine qui attire le plus de faux négatifs du domaine de départ. En appliquant

cette procédure à C, on obtient la MPPV pour le modèle *SGD+TF-IDF* présentée à la figure 4.13.

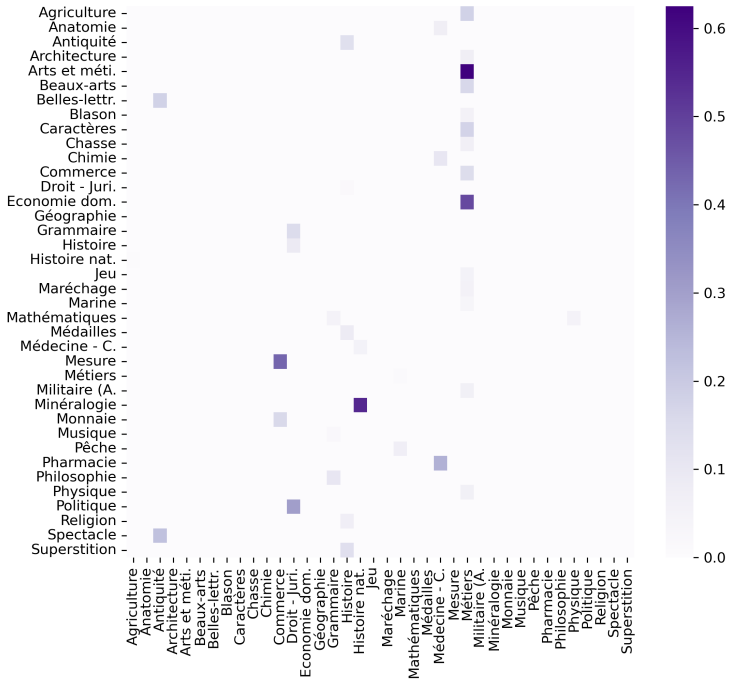


Figure 4.13 – Matrice des Plus Proches Voisins pour la méthode SGD+TF-IDF sur le jeu de test.

À partir de cette matrice, il est aisé de construire le GPPV correspondant en appliquant la procédure utilisée plus haut pour prouver le sens retour dans l'équivalence entre graphe et matrice, à savoir construire le graphe à partir des coefficients de la matrice. Le résultat de cette opération appliquée à la MPPV de la figure 4.13 est le GPPV du modèle *SGD+TF-IDF* visible à la figure 4.14. Ses arêtes reprennent le schéma de couleur utilisé dans ce manuscrit pour les matrices de confusion : le mauve y est d'autant plus prononcé que le taux de faux négatifs y est important. Ainsi, le cas des classes *Minéralogie* et *Arts et métiers*, très mal reconnues, ressort nettement sur le graphe : plus de la moitié de leurs articles sont incorrectement prédits par le modèle (en faveur des classes *Histoire Naturelle* et *Métiers* respectivement).

La particularité la plus visible est sans doute la très forte attractivité de la classe *Métiers*, traduisant la densité relativement élevée de la colonne correspondant à cette classe sur la MPPV de la figure 4.13. De manière cohérente avec le résultat des analyses du début de cette section, la *Géographie* pointe sur l'*Histoire* car ce domaine est le plus fort attracteur de ses faux négatifs. En revanche, il apparaît que les relations ne sont évidemment pas symétriques car si la plupart des faux positifs de *Géographie* viennent de l'*Histoire*, le modèle «envoie» encore plus d'article de ce domaine vers *Droit - Jurisprudence*. Entre ces deux classes par contre, il y a une réciprocité, car l'*Histoire* est aussi le domaine qui attire le plus de faux négatifs de *Droit - Jurisprudence*. Une autre configuration intéressante est celle de la classe *Mathématiques* dont partent deux arêtes [6] (c'est le seul nœud dans ce cas) : les classes *Physique* et *Grammaire* attirent autant de faux négatifs de la classe *Mathématiques*. Sans cet équilibre, le graphe serait partagé en deux composantes indépendantes : la partie supérieure de la figure 4.14 rassemble plutôt des domaines liés à un certain savoir-faire et la partie inférieure des domaines liés à la connaissance et à la réflexion. Il n'y a pas d'explication évidente à cette division qui soit liée à la méthodologie employée : elle n'est probablement pas stricte et certains liens

6. ce fait peut sembler déconcertant mais correspond aux pluriels entre parenthèses proposés un peu plus haut : en effet, vu la méthode de construction, si la valeur maximale est atteinte par plusieurs arêtes, il n'y a aucune raison d'en préférer une et elles sont toutes gardées

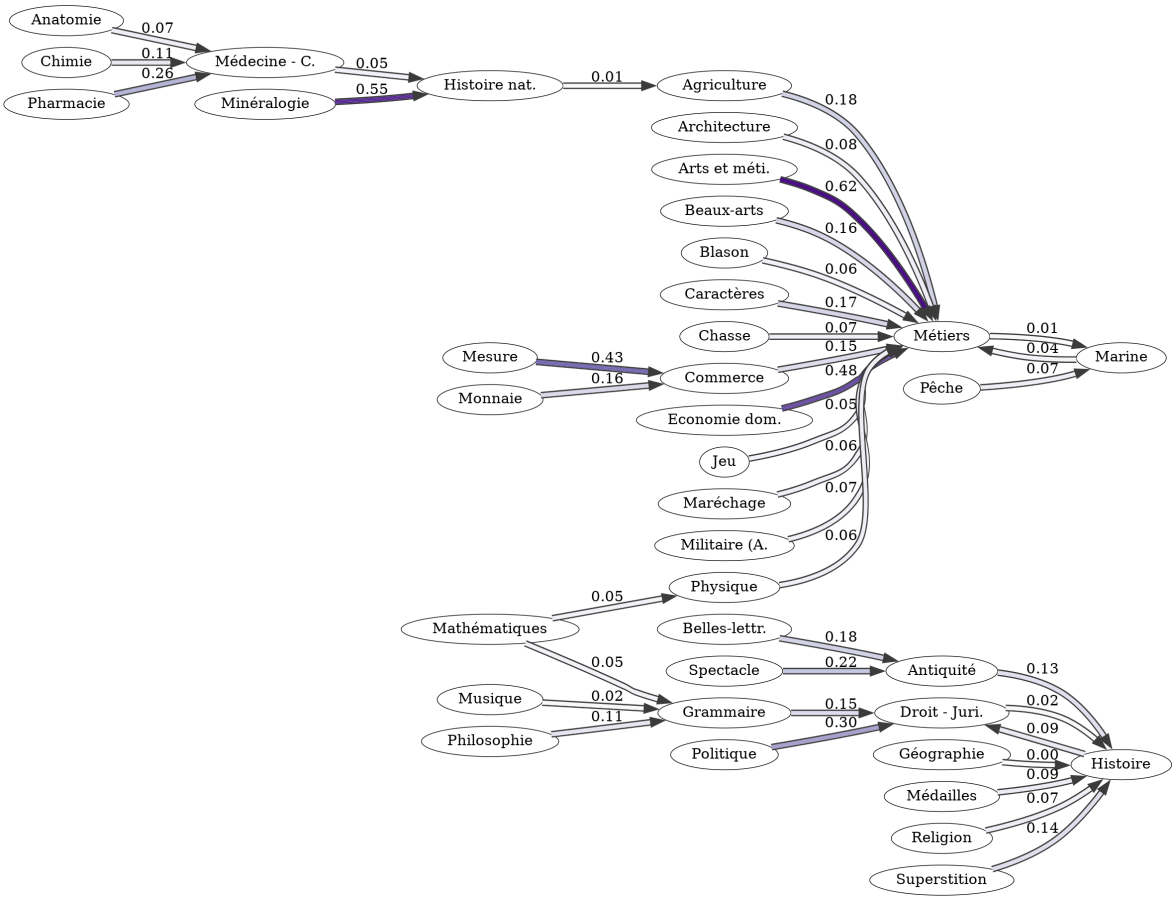


Figure 4.14 – Graphe des Plus Proches Voisins pour la méthode SGD+TF-IDF sur le jeu de test.

s'expliqueraient sans doute davantage par des ressemblances de thématiques et de vocabulaire que par une démarche similaire (on peut penser par exemple au lien entre *Histoire Naturelle* et *Agriculture*), mais il n'en demeure pas moins intéressant de remarquer que la *Physique* et la *Chimie* font toutes deux partie de la composante supérieure (ce qui est assez surprenant avec une perspective contemporaine très théorique de ces deux disciplines). La *Géographie*, quant à elle, se situe bien dans la composante inférieure ce qui semble traduire – si l'on accepte l'interprétation précédentes des deux groupes – un domaine qui compile des informations plus qu'il ne décrit des procédés.

En représentant les dynamiques des faux négatifs dans les prédictions d'un modèle, le GPPV constitue donc déjà une ressource visuelle utile pour identifier les proximités existant entre les classes. La section suivante poursuit l'exploration des proximités entre domaines en considérant les PPV d'autres graphes.

4.2.2 Similarités lexicales

Il est difficile de déterminer les raisons qui amènent le modèle *SGD-TF-IDF* à commettre les erreurs étudiées dans la section précédente (4.2.1). Si l'étude de cas menée à partir de la page 104 sur les erreurs autour de la classe *Géographie* a mis en évidence des passages consacrés à différentes disciplines, elle ne dit pas ce qui permet de les identifier. La présente section s'intéresse aux ressemblances en termes de lexique entre les *domaines regroupés* comme un facteur explicatif possible. Le but est de déterminer si la structure mise en évidence précédemment sur le GPPV (voir figure 4.14) peut être reproduite ou au moins approchée à partir de propriétés lexicales simples des articles.

Le modèle *SGD+TF-IDF* étudié, comme tout modèle d'AA, base ses prédictions de classification pour un article sur un ensemble de traits des tokens de cet article défini empiriquement lors de l'apprentissage mais inconnu pour qui l'utilise. Cette raison contribue à ce que les modèles d'apprentissage automatique soient parfois appelés «boîtes noires» et constitue un argument en faveur de la démarche d'Intelligence Artificielle «Explicable» («XAI» ou «eXplainable Artificial Intelligence» en anglais), visant à comprendre à quoi s'intéressent les modèles. Cette préoccupation peut avoir au moins deux buts : d'une part améliorer les performances des modèles (en préselectionnant les propriétés à observer pour s'assurer que tous utilisent celles du meilleur modèle plutôt que de laisser chaque modèle s'en approcher individuellement) et d'autre part transférer leurs «découvertes» statistiques pour apprendre des choses sur les traits qui caractérisent un genre ou, dans le cas présent, un domaine.

En plus des propriétés elles-mêmes (lemme ou forme, avec ou sans POS), la façon de les combiner peut également avoir un impact sur les performances du modèle. Ainsi, la notion de n-gramme propose de remplacer les tokens en entrée par des séquences brèves, faisant l'hypothèse que des fenêtres de mots contigus peuvent lever certaines ambiguïtés lexicales et jouer un rôle décisif dans les prédictions d'un modèle. On peut rapprocher cette idée de l'importance accordée aux collocations par l'école britannique (voir la section 2.5.2 p.38). Des travaux comme Tan et al. (2002) ou Mladenić & Grobelnik (2003) étudient comment sélectionner et combiner les propriétés apportant le plus d'information sur les textes afin de mieux les classer. La construction des n-grammes elle-même fait l'objet de recherches pour augmenter son efficacité (García et al., 2021). La longueur des n-grammes joue un rôle important : si les performances s'accroissent en utilisant des n-grammes par rapport à de simples tokens (c'est-à-dire quand $n > 1$ puisque les tokens peuvent être vus comme des «séquences de 1 token» donc des 1-grammes), elles décroissent en revanche quand n devient trop grand. Si toutes les études menées ne trouvent pas la même longueur seuil exactement, toutes semblent s'accorder sur sa faible valeur (inférieure à 5). Fürnkranz (1998) et à sa suite Tan et al. (2002, p. 531) estiment que les performances se dégradent au-delà de 3. Il est intéressant de constater que ces valeurs sont du même ordre de grandeur que les fenêtres utilisées généralement pour les recherches de cooccurrences dans les outils de textométrie.

Pour ces raisons, dans ce qui suit les articles seront représentés tour à tour par des séquences de leurs tokens de longueur 1, 2 puis 3 pour explorer à quel point les similarités lexicales des n-grammes obtenus ressemblent aux erreurs du modèle. D'autres paramètres s'ajoutent à n pour contrôler quelles séquences vont être considérées : le nombre de n-grammes associés à chaque classe (k dans ce qui suit) ainsi que la métrique utilisée pour effectuer les comparaisons.

Premièrement, soit $\mathcal{L}$ le lexique de l'*EDdA*, c'est-à-dire l'ensemble de tous les mots qui apparaissent dans les articles. Pour tout entier $n \in \mathbb{N}$, un n-gramme est un élément de l'ensemble $\mathcal{L}^n$, que l'on peut aussi écrire informellement comme le produit Cartésien répété $\mathcal{L} \times \mathcal{L} \times ... \mathcal{L}$, et ce n fois : en effet, un n-gramme a n composantes, chacune à valeur dans $\mathcal{L}$ ce qui correspond à la notion ensembliste de produit Cartésien, ou en logique à la notion de conjonction : avoir un n-gramme, c'est avoir n mots du lexique à la fois. Cette étude considère donc suite à la remarque ci-dessus les 1-grammes, les 2-grammes et les 3-grammes. Pour simplifier les calculs et éviter les disparités causées par les différences entre les classes en termes de longueur et de nombre d'articles, elle ne retient que les n-grammes les plus fréquents pour chaque classe, une même quantité pour toutes les classes, ce qui constitue le deuxième paramètre de l'expérience présenté plus haut, appelé k et qui prendra successivement les valeurs 10, 50 et 100 (c'est-à-dire, que chaque classe sera tour à tour représentée par ses 10 n-grammes les plus fréquents, ses 50 plus fréquents, puis ses 100 les plus fréquents). Utiliser plusieurs valeurs pour l'expérience permet d'observer l'impact de k et de contrôler dans quelle mesure cette simplification distord les résultats.

Pour représenter les classes par une combinaison de n-grammes, on considère maintenant l'espace $\mathbb{R}^{|\mathcal{L}|^n}$ (l'espace vectoriel sur le corps $\mathbb{R}$ de dimension la taille du lexique $\mathcal{L}$ à la puissance n) – une dimension certes grande devant les espaces vectoriels manipulés communément en géométrie élémentaire mais encore finie. Une base naturelle de cet espace est la famille canonique de vecteurs caractéristiques $\{e_W, W \in \mathcal{L}^n\}$ correspondant à chaque n-gramme et dont les composantes sont toutes nulles sauf une, celle qui correspond au n-gramme représenté par le vecteur considéré. Tant qu'il s'agit de compter des occurrences de chaque n-gramme, les coefficients ne pourront qu'être entiers (c'est-à-dire à valeur dans $\mathbb{N}$), mais faire le choix de $\mathbb{R}$ et donc de se placer dans un espace vectoriel plutôt qu'un simple module permet d'effectuer ensuite des divisions sur les coefficients, utiles pour normaliser les vecteurs obtenus. Ce formalisme revient à faire des *sac de n-grammes*, d'une manière similaire à laquelle *BoW* créait des *sac de mots* à partir des tokens.

Pour un n-gramme t donné, on a besoin de compter son nombre total d'occurrences parmi tous les articles de la classe c, noté $|t|_{c,n}$, c'est-à-dire qu'on considère la fonction $t \to |t|_{c,n}$ de l'espace $\mathcal{L}^n \to \mathbb{N}$. Par exemple, $|(\text{"à"}, \text{"plusieurs"}, \text{"choses"})|_{\text{Géographie},3}$ représente le nombre d'occurrence du trigramme ("à", "plusieurs", "choses") dans l'ensemble des articles classés à *Géographie* par le modèle. Comme chaque classe c sera représentée par l'ensemble de ses n-grammes les plus fréquents, ce qui sera noté $\mathcal{T}_{c,n,k}$, on pourra alors lui associer un vecteur $V_{c,n,k}$ dans l'espace vectoriel $\mathbb{R}^{|\mathcal{L}|^n}$ défini ci-dessus. Les composantes de ce vecteur sont le compte d'occurrences dans la classe avec la fonction précédente $|...|_{c,n}$ pour chacun de ses k n-grammes les plus fréquents et 0 pour tous les autres n-grammes possibles. Plus formellement, cela signifie qu'on peut écrire la combinaison linéaire :

$$V_{c,n,k} = \sum_{W \in \mathcal{T}_{c,n,k}} |W|_{c,n} \cdot e_W$$

Représenter ainsi les classes par des vecteurs donne la possibilité d'exprimer des mesures de similarité à l'aide d'outils comme le produit scalaire et les notions dérivées de norme et de distance. On définit ainsi le produit scalaire sur les vecteurs des classes comme la somme des produits des composantes des vecteurs qu'ils ont en commun :

$$\langle V_{c_i,n,k}, V_{c_j,n,k} \rangle = \sum_{W \in \mathcal{T}_{c_i,n,k} \cap \mathcal{T}_{c_j,n,k}} |W|_{c_i,n} \times |W|_{c_j,n}$$

Et, par extension, la norme d'un vecteur de classe est définie par la racine carrée de son produit scalaire avec lui-même (c'est-à-dire la racine carrée de la somme des carrés de ses coefficients, soit l'équivalent de la norme euclidienne manipulée couramment sur les espaces à 2 ou 3 dimensions, mais généralisée à l'espace de dimension $|\mathcal{L}|^n$ considéré) :

$$\|V\| = \sqrt{\langle V, V \rangle}$$

Avec ces définitions, il devient possible de considérer deux métriques. La première, qu'on appellera la similarité *cardinale* consiste à simplement compter la proportion de n-grammes que deux classes données ont en commun. En supposant que c_i et c_j sont deux classes, leur similarité *cardinale* $\langle c_i, c_j \rangle_{n,k,card}$ peut s'exprimer ainsi (la division par k sert à normaliser le nombre obtenu pour le ramener entre 0 et 1 – deux ensembles de k éléments ne peuvent pas avoir plus de k éléments en commun – pour la rendre comparable indifféremment de k) :

$$\langle c_i, c_j \rangle_{n,k,card} = \frac{\|\mathcal{T}_{c_i,n,k} \cap \mathcal{T}_{c_j,n,k}\|}{k} \tag{4.1}$$

La seconde, qu'on appellera la similarité *scalaire*, s'obtient en calculant leur produit scalaire, normalisé en le divisant par le produit de leurs normes. Toujours avec la même notation, cette autre métrique peut s'écrire :

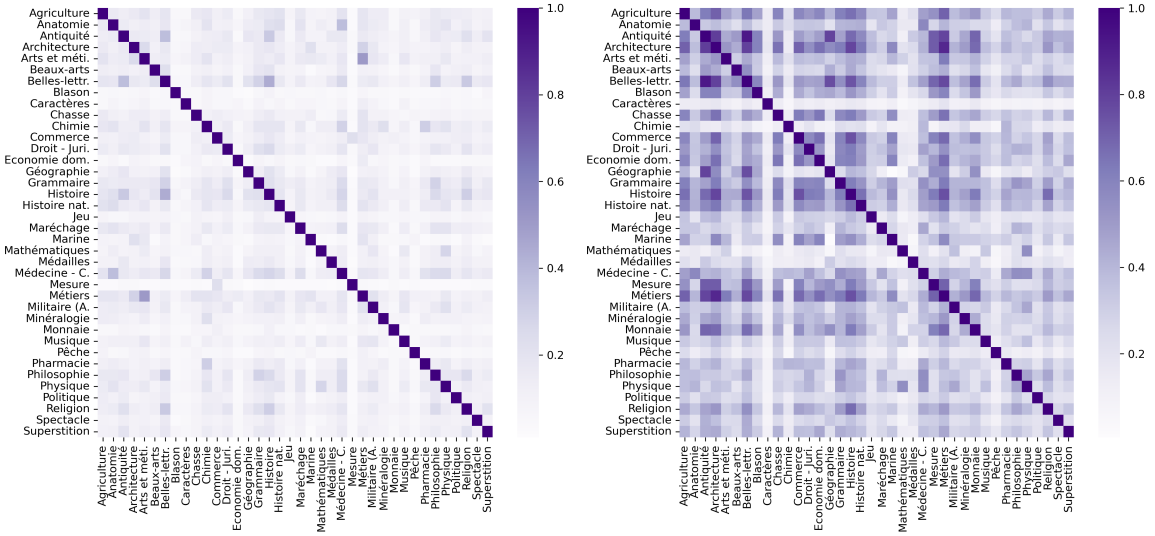
$$\langle c_i, c_j \rangle_{n,k,scal} = \frac{\langle V_{c_i,n,k}, V_{c_j,n,k} \rangle}{\|V_{c_i,n,k}\| \times \|V_{c_j,n,k}\|} \tag{4.2}$$

À partir de ces deux métriques, il est possible de générer des matrices de similarité pour chacune des configurations possibles, 18 au total : 3 valeurs possibles pour le paramètre n (la taille des n-grammes, pouvant être 1, 2 ou bien 3), 3 pour k (pour rappel 10, 50 ou 100) et 2 métriques à considérer chaque fois (*cardinale* ou *scalaire*). La figure 4.15 reproduit deux d'entres elles correspondant à $k = 100$ et $n = 2$ (100 2-grammes les plus fréquents par classe), pour les deux métriques possibles.

La première diagonale de ces matrices constitue une ligne de symétrie car la relation qu'elles représentent est symétrique : pour toute paire de classes c_i et c_j, $\langle c_i, c_j \rangle = \langle c_j, c_i \rangle$ (vrai pour chacune des deux métriques considérées, toutes deux définies à partir d'opérations symétriques – produits et sommes d'entiers, intersections d'ensembles). Par contraste, la matrice de confusion du modèle *SGD+TF-IDF* était asymétrique car le modèle ne confond pas les articles de *Géographie* avec ceux d'*Histoire* avec le même taux d'erreur qu'il confond ceux d'*Histoire* avec de la *Géographie* (un fait déjà visible sur le GPPV du modèle à la figure 4.14 p.111 et discuté à la section 4.2.1). Cette différence est une conséquence de la simplification opérée par rapport aux résultats des prédictions du modèle.

Les matrices sont également de plus en plus creuses à mesure que les paramètres n et k croissent, car la probabilité que deux classes partagent un de leurs n-grammes les plus fréquents décroît alors : d'une part plus un n-gramme est long, plus il tend à être unique et d'autre part plus la liste de n-grammes considérée est longue, plus des n-grammes différents ont des chances d'apparaître.

Enfin, la première métrique, la similarité *cardinale* présente l'avantage d'être moins «bruitée» que la similarité *scalaire*, c'est-à-dire que la couleur de fond des matrices de similarités générées est



(a) Pour la similarité *cardinale*

(b) Pour la similarité *scalaire*

Figure 4.15 – Similarités lexicale des 100 2-grammes les plus fréquents

plus claire, comme cela est visible sur la figure 4.15a (à gauche) par rapport à la figure 4.15b (à droite). Cela peut sembler contre-intuitif si l'on regarde le calcul de la similarité *cardinale* comme simplement un cas particulier de la similarité *scalaire* dans lequel tous les coefficients des vecteurs sont mis à 1, au coefficient multiplicatif de leur norme près. En effet, on pourrait s'attendre à ce que le produit scalaire des vecteurs de deux classes soit systématiquement très inférieur au décompte de leurs vecteurs de base en commun, en ce sens qu'avoir des coefficients différents ne peut que rendre les vecteurs moins colinéaires.

Mais cet effet est vaincu par la distribution de «masse» des coefficients dans les vecteurs : si deux vecteurs ont seulement très peu de composantes en commun, mais que leurs coefficients les plus élevés se trouvent précisément sur ces composantes, alors leurs projections sur le sous-espace des vecteurs qu'ils ont en commun produira des vecteurs de taille proche de leur taille initiale. Si de plus les distributions des composantes de ces vecteurs projetés ont des formes semblables (le même ordre quand on les classe par valeur de leurs coefficients, et avec le même ratio entre elles), alors leur produit scalaire peut rester assez proche du produit de leurs normes. Dans ce cas, la similarité *scalaire* donnera une valeur proche de 1, alors qu'ils ne partageront que quelques n-grammes, ce qui donnera une similarité *cardinale* très faible. En d'autres termes, les classes n'ont que peu en commun mais ce qu'elles ont en commun est précisément ce qui leur est le plus fondamental. Les composantes qu'elles ne partagent pas ne jouent alors qu'un rôle tout à fait marginal. Bien entendu, pour que cet effet s'applique, les deux classes doivent avoir au moins un n-gramme en commun, sans quoi leur produit scalaire sera simplement nul.

Cette configuration contre-intuitive se produit par exemple avec les 3-grammes des classes *Histoire* et *Religion*, visibles au tableau 4.5. Ces deux classes ne partagent que 25 de leur 100 3-grammes les plus fréquents (ce qui correspond donc à un score de similarité *cardinale* de 0.25) mais leur produit scalaire normalisé (la similarité *scalaire*) est de 0.61 car leurs projections sur le sous-espace des 25 3-grammes qu'elles ont en commun restent très proches d'elles (où leurs projections représentent respectivement 87.8% et 87.3% de leurs normes initiales).

Le tableau contient 10 n-grammes provenant d'informations sur l'emploi grammatical des articles («s.f.pl.» est l'abréviation de «substantif féminin pluriel», voir la section 3.1.2 p.47) et 2 qui sont des signatures ou des mentions (du même auteur, «c. d. j.» étant certainement l'abréviation du «Cheva-

n-gramme	Histoire	Religion
('s.', 'm.', 'nom')	199	78
('s.', 'm.', 'pl')	144	210
('m.', 'chevalier', 'jaucourt')	76	12
('avant', 'jesus', 'christ')	72	13
('s.', 'm.', 'terme')	33	15
('m.', 'pl', 'nom')	32	59
('s.', 'f.', 'pl')	31	12
('s.', 'f.', 'nom')	27	10
('m.', 'nom', 'donne')	27	16
('depuis', 'long', 'tems')	26	19
('s.', 'm.', 'mot')	24	11
('depuis', 'tems', '-là')	22	13
('a', '-t', '-il')	22	26
('a', 'long', 'tems')	19	9
('a', 'donné', 'nom')	18	16
('adj', 'pris', 'subst')	17	10
('long', 'tems', 'avant')	16	9
('c.', 'd.', 'j.')	13	10
('mot', 'vient', 'grec')	12	15
('s.', 'm.', 'plur')	12	27
('trévoux', 'chambers', 'gramme')	12	10
('auteur', 'esprit', 'loi')	11	11
('ajoute', '-t', '-il')	10	12
('mot', 'formé', 'grec')	9	16
('nom', 'a', 'donné')	9	19

Tableau 4.5 – Les 3-grammes les plus fréquents communs aux classes *Histoire* et *Religion* avec leurs comptes d'occurrences dans chacune

lier De Jaucourt»). Si les premiers sont sans doute peu pertinents pour cette étude (sauf à supposer que les proportions de termes féminins ou pluriels d'un domaine puissent avoir une signification), les signatures d'auteur peuvent avoir leur importance : d'une part parce que la capacité d'un même auteur à intervenir (ou être cité) dans deux domaines peut être un signe de leur proximité, d'autre part parce que dans ce cas son style personnel peut augmenter artificiellement leur ressemblance s'il a signé une part suffisamment grande des articles des deux domaines concernés. La notion de temps long dans le passé est traduite par 5 autres n-grammes : «avant jesus christ», «depuis long tems» et «depuis tems -là» (reliquat de «depuis ce tems-là» après élimination des mots vides, voir la section 4.1.2 p.93). Les autres n-grammes sont plus difficile à analyser : 2 montrent la prévalence de termes d'origine grecque, 4 semblent des tournures assez communes, et les 2 dernières concernent à nouveau des auteurs. Le fait de référer à Montesquieu (auteur de l'«Esprit des Lois») par une périphrase pourrait être propre au style d'un contributeur de ces deux domaines (Jaucourt ?), mais pourrait également être biaisé par l'observation des 3-grammes : «Montesquieu» n'a pas de raison d'être environné toujours des mêmes mots et son nom a donc très peu de chance d'obtenir assez d'occurrence pour remonter parmi les 3-grammes les plus fréquents, alors que l'expression «l'auteur de l'Esprit des Lois», devient après la préparation du corpus le 3-gramme «auteur esprit loi», unique, et donc décompté systématiquement à chacune de ses occurrences. Enfin le dernier 3-gramme restant, «trévoux chambers gramme» correspond vraisemblablement à une bibliographie sommaire et à une signature maltraitée par le lemmatiseur. En effet deux sources très communément employées par les auteurs de l'*EDdA* sont le *Dictionnaire universel de Trévoux*, dont plusieurs éditions ont été imprimées entre 1704 et 1771, et la *Cyclopaedia* de Chambers, qui a elle aussi connu plusieurs éditions entre 1728 et 1753 et dont la traduction en français était le but initial du projet de l'*EDdA* (voir section 3.1.1 p.44). Le mot «gramme», assez inattendu ici, a sans-doute été réintroduit par le lemmatiseur à la place d'un «g.», lettre employée par l'abbé Mallet, un contributeur important des domaines *Histoire* et *Religion* pour signer ses articles.

Il est difficile d'exploiter les matrices de similarités de la figure 4.15 au-delà des remarques structurelles déjà faites et de tirer des n-grammes plus que les quelques remarques qualitatives ci-dessus. De la même façon que pour l'étude de la matrice de confusion du modèle *SGD+TF-IDF* menée à la section 4.2.1, leur restriction aux PPV permet d'en tirer des graphes mettant plus clairement en évidence les ressemblances lexicales majeures entre les domaines. C'est ainsi que la matrice de la figure 4.15a permet après simplification de générer le graphe de la figure 4.16. La même procédure a été appliquée à chacune des 18 matrices de similarités correspondant au 18 configurations possibles (des paramètres n, k, et de la métrique considérée) pour générer 18 GPPV de similarités, qui ne sont pas tous reproduits ici mais qui ont été analysés.

Plusieurs critères distinguent assez nettement tous ces graphes du GPPV pour les confusions du modèle *SGD+TF-IDF* (voir figure 4.14 p.111). Alors que le GPPV du modèle était connexe (presque partagé en 2 composantes, rassemblées seulement par la classe *Mathématiques*), 15 se divisent en au moins 3 composantes (jusqu'à 6 sur celui de la figure 4.16 et 8 pour celui des 10 3-grammes ayant la plus forte similarité *scalaire*), et plusieurs graphes contiennent même des nœuds isolés, c'est-à-dire des domaines qui n'ont aucun n-gramme en commun avec aucun autre domaine. Les 3 autres graphes sont tous des GPPV générés pour la similarité *cardinale*. Tous les graphes pour cette mesure de similarité, même ceux ayant plus de 3 composantes non connexes possèdent 2 nœuds ou plus reliés à plusieurs autres nœuds, certains jusqu'à 9 comme la classe *Militaire* sur le GPPV des 50 3-grammes. Cet effet, très prononcé pour $k = 10$ s'explique par le faible nombre de valeurs pouvant être prises par les scores de similarité *cardinale*.

En effet pour chaque paire de nœuds considérée, elle prend une valeur de la forme $\frac{i}{k}$ avec $i \in \mathbb{N}, i \geq 0, i \leq k$ (k est le paramètre fixé pour le graphe considéré, pouvant valoir 10, 50 ou 100). Pour un nœud donné, en calculant sa similarité *cardinale* avec chacun des 37 autres nœuds, la métrique retournera donc une des seulement 11, 51 ou 101 valeurs possibles. Par application du «principe des

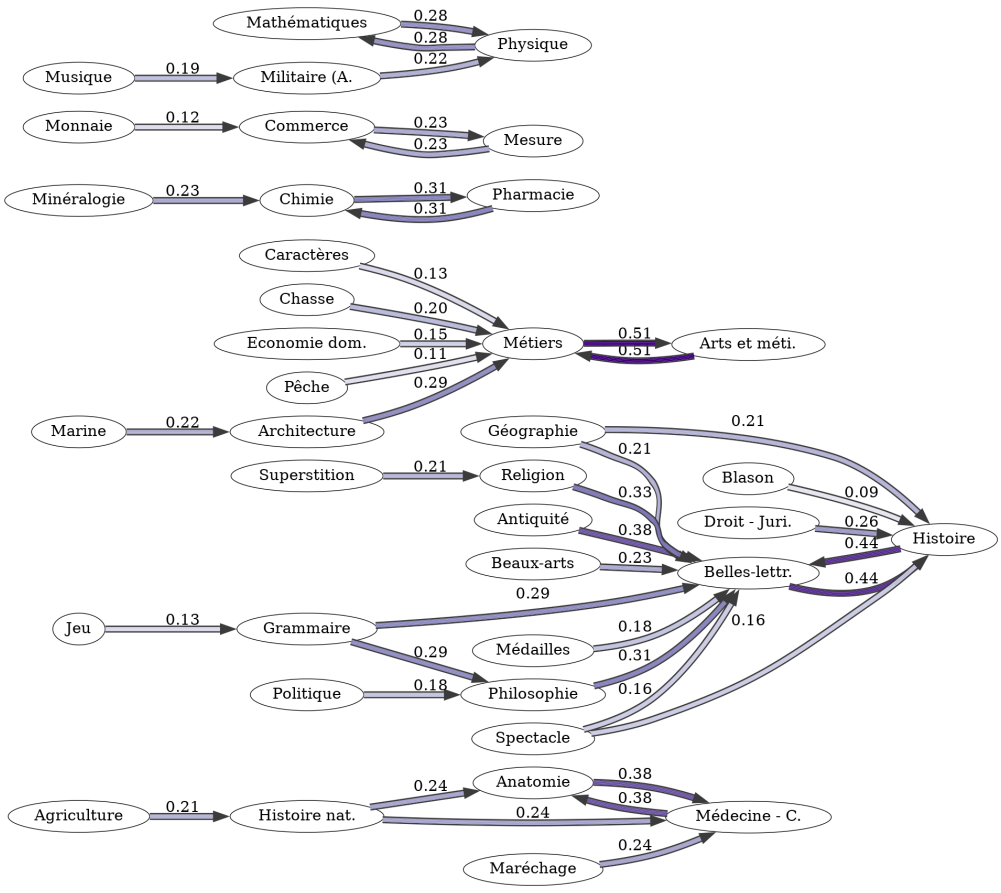


Figure 4.16 – Graphe des Plus Proches Voisins pour la similarité *cardinale* mesurée sur les 100 2-grammes les plus fréquents de chaque classe.

tiroirs» il y aura nécessairement plusieurs nœuds (de destination) avec la même similarité *cardinale* pour ce nœud considéré pour $k = 10$, des «collisions» dans les valeurs de sortie de la fonction qui a chacun des 37 autres nœuds attribue sa similarité avec le nœud de départ considéré. Cette certitude de collision devient seulement un risque, mais assez élevé, pour les deux autres valeurs de k. Si une collision se produit sur la plus haute valeur atteinte pour ce nœud, alors il y aura autant d'arêtes sortant de ce nœud que de classes impliquées dans la collision puisque ce sont ces arêtes avec la plus haute similarité qui sont utilisées pour tracer le graphe.

Malgré ces réserves, les GPPV de similarités contiennent des motifs similaires à ceux du GPPV des confusions du modèle de la figure 4.14 (p.111). Par exemple, on constate que la connexion entre *Mathématiques* et *Physique* est bien présente dans tous les graphes de similarités sauf 3 (ceux pour $k = 10$, $n = 1$ avec la similarité *scalaire* et $n = 3$ pour les deux mesures de similarité). Le triangle formé par *Mesure*, *Monnaie* et *Commerce* se trouve dans tous les graphes sauf ceux des 2-grammes pour la similarité *scalaire* (soit seulement 3 graphes également). On retrouve également des proximités fréquentes entre *Médecine*, *Anatomie*, *Pharmacie* et *Chimie* sur les graphes. En revanche le nœud *Métiers* n'est pas toujours aussi attracteur, cédant fréquemment sa place à des domaines comme les *Belles-Lettres*, la *Religion* ou l'*Histoire*. Quant à la *Géographie*, sa plus forte similarité est avec la classe *Antiquité* (jusqu'à 80% de similarité *scalaire* pour leur 50 2-grammes les plus fréquents), sauf dans 5 graphes (dont celui de la figure 4.16) où ses plus fortes ressemblances vont à l'*Histoire* et aux *Belles-Lettres*. Ces trois classes figuraient parmi les faux négatifs et les faux positifs les plus fréquents de la classe *Géographie* (voir la section 4.2.1), mais il est intéressant de noter que la classe *Histoire* qui apparaissait nettement en tête des erreurs n'obtient la plus forte similarité avec *Géographie* que sur relativement peu de GPPV de similarités.

Les erreurs les plus importantes faites par le modèle (correspondant donc au GPPV de sa matrice de confusion) présentent donc une certaine compatibilité avec les similarités lexicales qui existent entre les domaines. La ressemblance est davantage marquée pour les 2- et 3-grammes, et pour certains domaines. Les erreurs faites à propos de la *Géographie* en particulier ne suivent pas tout à fait le schéma de ses ressemblances lexicales.

4.2.3 Application de la centralité à la structure du modèle

Les deux sections précédentes tentent de dégager du sens d'objets complexes en les simplifiant localement (en considérant une colonne et une ligne particulière ou en limitant les arêtes partant de chaque nœud d'un graphe). La dernière étape restante pour analyser les liens entre domaines consiste donc logiquement à essayer de capturer leurs dynamiques d'ensemble. Pour y parvenir, cette section considère une mesure de centralité sur le graphe défini en considérant la matrice de confusion du modèle *SGD+TF-IDF*, C, comme une matrice d'adjacence pondérée. Il ne s'agit donc plus du GPPV de la section 4.2.1, engendré par la MPPV, mais de l'objet qui avait été initialement écarté au profit du GPPV car sa complexité ne permettait pas de le tracer pour l'exploiter visuellement. Ce graphe, correspondant à la matrice de confusion via l'équivalence décrite à cette même section 4.2.1, sera dénommé dans ce qui suit le «graphe de confusion» et noté $\mathcal{G}_C$. En cohérence avec les choix faits pour les GPPV dans les sections 4.2.1 puis 4.2.2, les arêtes de $\mathcal{G}_C$ représentent les faux négatifs du modèle (plutôt que les faux positifs).

Si un graphe représente des déplacements sur un réseau, la centralité de vecteur propre choisie pour cette étude correspond à une situation particulière d'équilibre des flux. Elle est définie comme une distribution de poids sur les nœuds caractérisée par le fait d'être inchangée après une étape de déplacement suivant les coefficients de ses arêtes. Dans le cadre de l'analyse de $\mathcal{G}_C$, l'interprétation dynamique du processus de prédiction des classes par le modèle donnée à la section 4.2.1 permet de traduire cette situation d'équilibre sur un ensemble donné d'articles. En considérant par exemple 1 000 articles de *Géographie*, d'après les scores mesurés sur le jeu de test le modèle en reconnaîtrait

correctement 985 comme étant de la classe *Géographie*, «presque» [7] 4 comme de l'*Histoire*, 2 comme de l'*Antiquité*, etc. Une distribution initiale d'articles exclusivement sur le nœud *Géographie* semblerait, après le «déplacement» causé par la prédiction du modèle, s'être étalée sur ses voisins en suivant les taux d'erreurs du modèle pour cette classe. Si au lieu de choisir 1 000 articles de la classe *Géographie* la distribution avait comporté 900 articles de *Géographie* et 100 d'*Histoire*, alors environ 6 articles d'*Histoire* aurait été pris pour de la *Géographie* par le modèle [8]. Leurs mouvements sur le graphe auraient dans une certaine mesure compensé ceux de la classe *Géographie* que le modèle aurait pris pour de l'*Histoire*. Une mesure de centralité des nœuds peut donc dans ce cas s'interpréter comme une proportion d'articles entre les différents domaines qui «résonne» d'une manière particulière avec le modèle en lui permettant de ne pas se tromper sur les quantités : pour un échantillon d'articles choisis dans chaque domaine proportionnellement à leur centralité, le modèle trouve en sortie exactement le bon nombre d'articles de chaque domaine (il se trompe sur certains articles de chaque classe mais dans l'ensemble ses erreurs se compensent).

Les poids attribués par la mesure de centralité permettent de se faire une idée de l'attractivité des nœuds les uns par rapport aux autres car la valeur en un nœud décrit une situation d'équilibre entre deux effets contraires. Elle peut être d'autant plus élevée (relativement aux autres, c'est-à-dire en supposant leurs poids fixés) que ce nœud reçoit un flux important des autres nœuds (en nombre – qu'il y ait beaucoup d'arêtes – ou en volume – des arêtes avec un coefficient élevé). Cette augmentation du poids est tempérée de manière linéaire par l'importance du flux qui quitte le nœud considéré (de même, en nombre ou en volume, ici encore seule la somme des coefficients des arêtes quittant un nœud donné a une importance) : supposant les coefficients de toutes les autres arêtes fixés, le poids d'un nœud sera d'autant plus faible que ses arêtes sortantes ont des coefficients importants. Une mesure de centralité élevée signifiera donc qu'un nœud est assez bien reconnu (flux de faux négatifs sortant bas) tout en attirant assez de faux négatifs des autres classes. Sur ce dernier point, c'est le comportement d'ensemble qui compte, le nœud n'a besoin d'être le plus fort attracteur d'aucun de ses voisins pour avoir une centralité élevée, ce qui permet de détecter des tendances qui étaient invisibles en réduisant les graphes à leurs PPV.

Pour formaliser le problème de la recherche d'une telle distribution stable qui capturerait les influences entre classes, il est utile de représenter les distributions d'articles par des vecteurs de $\mathbb{R}^{38}$ (cette dimension correspondant, pour rappel, au nombre de classes des *domaines regroupés*). Une fois encore, laisser les coefficients prendre comme valeurs des nombres réels permet de faire la transition d'une expérience initiale sur les *articles* où seuls des entiers peuvent avoir du sens à un contexte statistique où il est possible de représenter des distributions moyennes, des probabilités et plus seulement un vrai échantillon donné d'articles. Avec ce formalisme un peu plus algébrique [9], et en gardant comme dans la section précédente 4.2.1 la notation C pour la matrice de confusion de la figure 4.5 (p.99), la définition de la centralité comme une distribution de poids laissée stable par C se traduit par l'existence d'un vecteur

$$v \in \mathbb{R}^{38} \text{ tel que } v \cdot C = v \tag{4.3}$$

Cette expression coïncide avec la définition d'un vecteur propre (à gauche) associé à la valeur propre 1. Selon la définition de la matrice de confusion, chaque rangée représente une distribution de pro-

7. Après avoir appliqué le modèle une fois, chaque classe ne contiendra bien évidemment qu'un nombre entier d'articles, mais les nombres réels issus des scores du modèle et qui vont être considérés dans cette partie décrivent un comportement statistique (ils ont d'ailleurs été obtenus à l'origine en mesurant un nombre entier d'articles, les virgules traduisent seulement le fait que le choix arbitraire de «1000» articles n'est pas multiple du nombre d'articles de *Géographie* présents dans le jeu de test).

8. Selon les résultats du modèle, encore une fois, cela se produit pour 5.84 des articles de la classe *Histoire*.

9. Il est à noter qu'avec la convention de placer les vraies étiquettes de classe en lignes et les étiquettes prédites en colonnes, appliquer le modèle correspond à un simple produit matriciel avec le vecteur à gauche de la matrice.

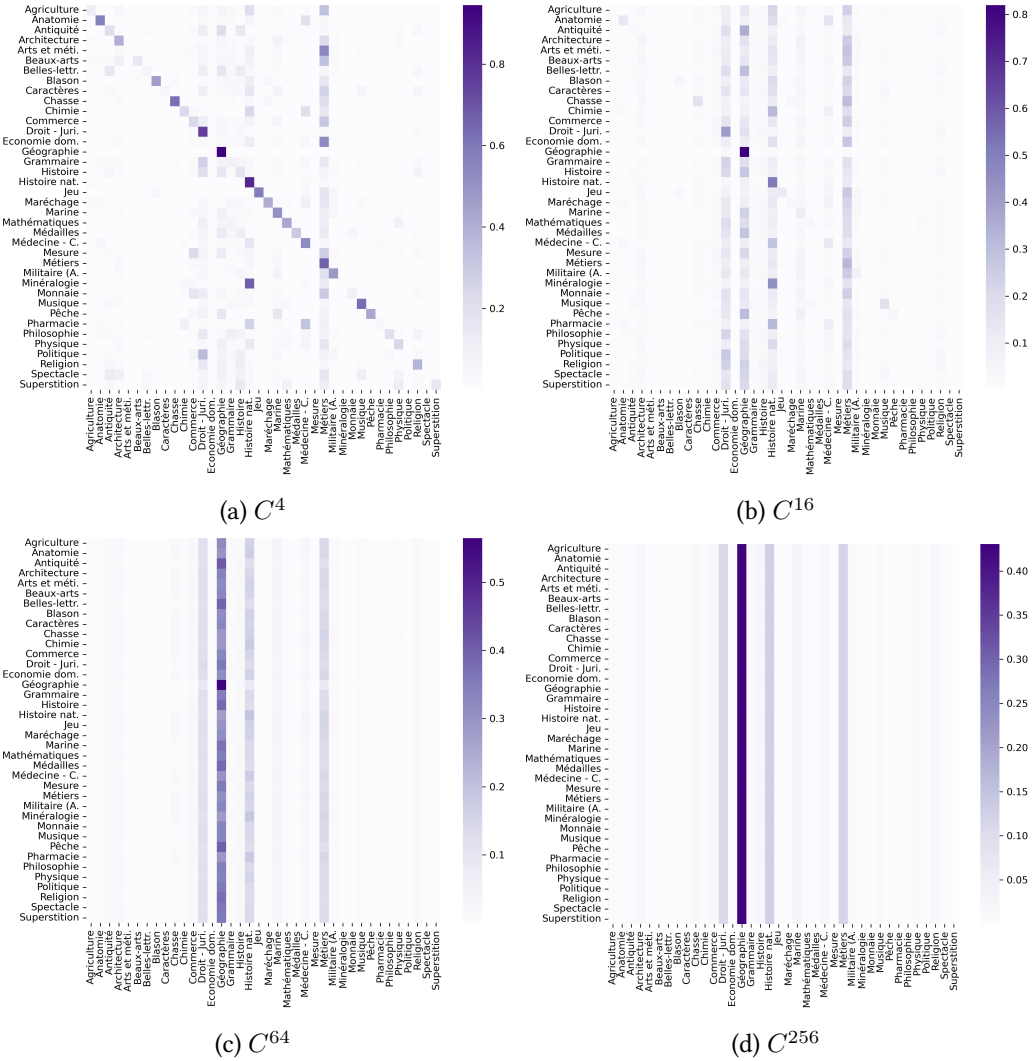
babilité (la probilité qu'un article donné issu de la classe correspondante soit prédit par le modèle dans chacune des classes possibles) : c'est-à-dire que C est stochastique à droite. En tant que telle, la somme des coefficients d'une rangée doit valoir 1. Étant données les règles de calcul des produits matriciels et en appelant v_1 le vecteur de $\mathbb{R}^{38}$ dont tous les coefficients sont égaux à 1, la propriété d'être stochastique à droite peut s'écrire :

$$C \cdot v_1 = v_1 \tag{4.4}$$

Cette expression montre que ce vecteur plein de 1 est un vecteur propre de la matrice C, mais seulement à droite. Or, les valeurs propres d'une matrice M sont les racines de son polynôme caractéristique $P_M[\lambda] = det(\lambda Id - M)$. Puisque par définition du déterminant, invariant par transposition, on a $det(M) = det(M^T)$, alors M et M^T ont le même polynôme caractéristique et de ce fait les mêmes valeurs propres. Ayant trouvé un vecteur propre à droite pour C associé à la valeur propre 1, on sait qu'il doit nécessairement exister un vecteur propre (à gauche cette fois) avec la même valeur propre, ce qui prouve que l'Équation 4.3 doit admettre au moins une solution. Ayant prouvé son existence, une manière pratique de calculer cette solution consiste à utiliser le théorème du cercle de Gershgorin qui, appliqué à une matrice stochastique, montre que toutes les valeurs propres doivent être inférieures ou égales à 1 en norme (somme des lignes). Par conséquent, itérer la matrice C de manière répétée va progressivement éliminer les «petits» coefficients (ceux strictement inférieurs à 1) en les faisant tendre vers 0 (parce que $\lim_{n\to\infty} \lambda^n = 0$ si $|\lambda| < 1$). Cela n'est en soi pas suffisant pour garantir la convergence des itérées de C car il pourrait y avoir plusieurs vecteurs propres associés à la valeur 1 si celle-ci avait une multiplicité > 1 [10]. Mais par contre, s'il y a convergence, alors ce sera nécessairement vers *le* vecteur propre (dont l'unicité serait alors prouvée) associé à cette valeur. Calculer les itérées de C donne par exemple pour les puissances 4, 16, 64 et 256 les matrices visibles dans la figure 4.17.

Après un peu plus de cent itérations, la matrice se stabilise et toutes ses lignes présentent une même distribution. Cette distribution est le vecteur propre recherché pour C et ses coefficients correspondent à la mesure de centralité de chaque nœud dans le graphe, explicitée sous forme d'histogramme à la figure 4.18. Cette même figure montre que la forme de la distribution diffère visiblement de celle du nombre d'articles par classe (voir la figure 3.34 p.83, dont la décroissance est bien plus continue) et présente trois groupes distincts : 1) la classe la plus centrale, 2) trois classes secondaires sur un pallier, et 3) toutes les autres en une longue traîne. De plus, alors que les classes ordonnées par centralité ont globalement des positions proches de leurs rangs par nombre d'articles, les classes les plus peuplées obtenant dans l'ensemble des scores de centralité plus élevés que celles les moins peuplées, l'ordre peut être localement assez différent. Par exemple, la classe *Droit - Jurisprudence*, 2ème classe par fréquence, n'est que 4ème par centralité, dernière du pallier secondaire. Juste après elle à la 5ème place se trouve la classe *Marine*, loin devant *Grammaire* (au 12ème rang avec une centralité presque 3 fois moindre) alors que cette dernière était plus fréquente que *Marine*. Il en est de même vers la fin de la distribution : une classe comme *Minéralogie*, avant-dernière en termes de fréquence n'est que la 8ème classe la moins centrale, avec une centralité plus de 5 fois supérieure à celle de l'avant-dernière. Bien que la mesure de centralité des classes semblent avoir un lien avec leurs tailles, toutes ces différences montrent que la taille seule ne suffit pas à expliquer en détail la distribution des centralités. Le modèle semble donc mieux identifier certaines classes grâce à leurs particularités intrinsèques.

10. l'application du théorème de Perron-Frœbenius permettrait de lever cette réserve en garantissant l'unicité de la valeur propre la plus grande en norme mais il s'applique sur des graphes connexes (matrices régulières). La densité élevée de $\mathcal{G}_C$ et la présence de boucles sur tous les nœuds (le modèle ne maltraite aucun domaine au point de ne reconnaître aucun de ses articles) engage à faire le pari de sa connexité, mais cela n'a pas été vérifié et le théorème ne peut pas être appliqué en toute rigueur. Le pragmatisme commande de passer outre : si les itérées de C convergent, l'unicité sera prouvée empiriquement, sans le théorème.



(a) C^4 (b) C^{16}

(c) C^{64} (d) C^{256}

Figure 4.17 – Des itérées de la matrice de confusion du modèle, C

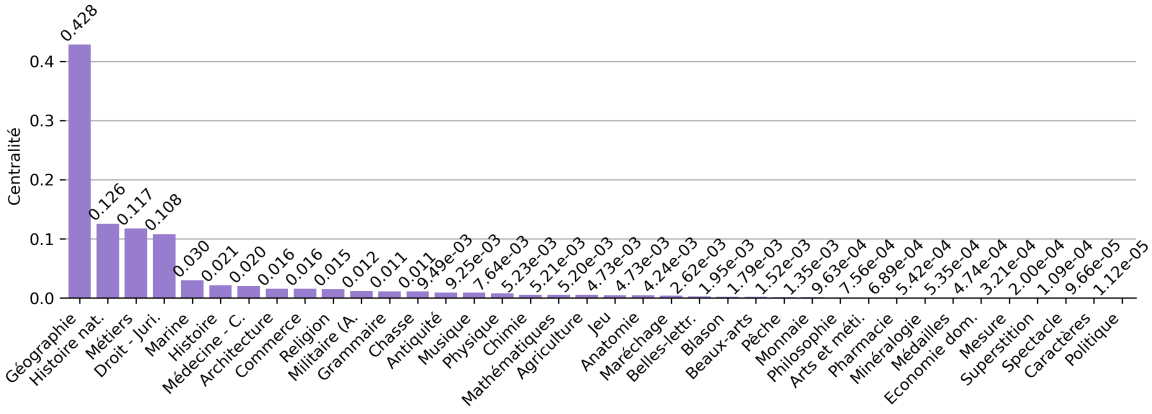


Figure 4.18 – La distribution par classe de mesures de centralité du graphe du modèle

En s'intéressant plus particulièrement à la classe *Géographie*, on constate qu'elle obtient le score le plus élevé (0.43). Ce résultat suggère que loin de jouer un rôle périphérique dans le graphe comme aurait pu le suggérer la figure 4.14 p.111, le domaine se place au centre du flux des erreurs de classification d'articles. En accord avec le nombre de faux positifs plus de trois fois supérieur à celui des faux négatifs mis en évidence à la section 4.2.1 ($3 \times 39 = 117 < 152$), le nœud *Géographie* du graphe reçoit un flux des autres nœuds du graphe supérieur au flux sortant d'après la remarque précédente sur l'interprétation de la centralité. Cette analyse montre donc une *Géographie* bien reconnue et détectée par le modèle dans de nombreux autres domaines, signe de l'importance au XVIII$^{\text{ème}}$ siècle de cette science au contact de nombreuses autres.

Au travers des différentes analyses précédentes, cette section amène plusieurs pistes de recherche à exploiter dans le chapitre 5. L'analyse qualitative initiale a montré des adjacences thématiques intéressantes entre la *Géographie* et des domaines comme l'*Histoire* et le *Commerce*. Les GPPV suggèrent quant à eux que les confusions du modèle entre la *Géographie* et les autres classes ne se fondent pas (ou au moins pas uniquement) sur des ressemblances lexicales (par opposition à des classes comme *Histoire* et *Métiers* qui attirent de nombreux faux négatifs d'autres classes et sont fréquemment la cible de nombreuses arêtes dans les GPPV de similarité lexicale). Au vu des résultats de la dernière section, cela signifierait que les contenus géographiques se distinguent assez bien de ceux des autres disciplines mais infusent largement dans l'*EDdA*, bien au-delà de ce qui a été initialement classé à *Géographie*.

4.3 Annotation des articles

Après avoir comparé différents modèles de classification automatique, l'étude des erreurs de prédiction a mis en évidence des ressemblances entre les domaines de connaissance. Ces parentés suggérent des pistes pour les analyses du chapitre 5 mais, avant de pouvoir les exploiter, il reste à appliquer une classification à l'ensemble des articles du corpus.

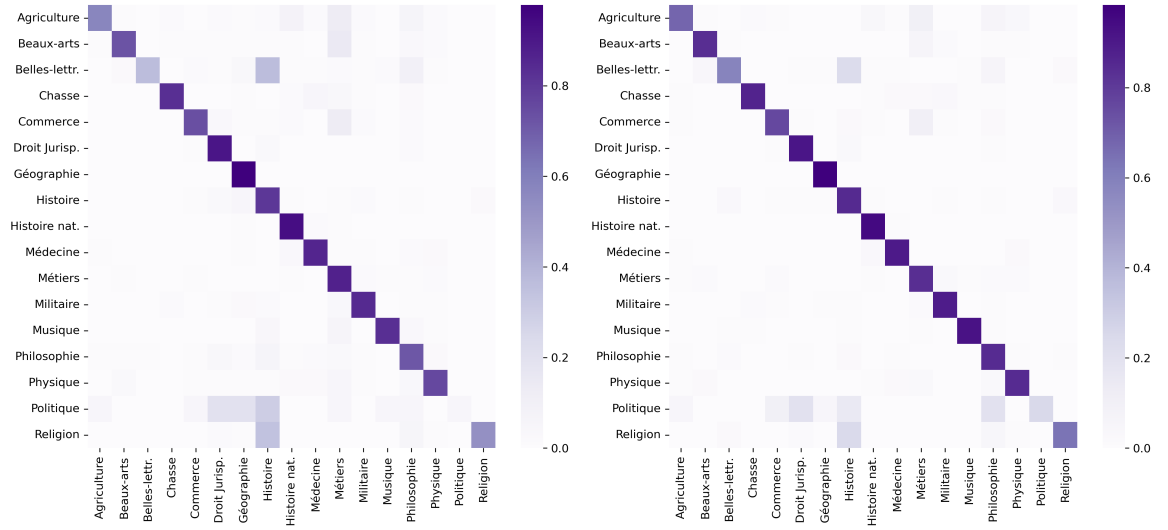
4.3.1 Choix d'un classifieur

Si le travail de comparaison des méthodes de classification décrit dans cette partie a été effectué sur le jeu des 38 «domaines regroupés», c'est celui des 17 «superdomaines» (les deux sont décrits à la section 3.3.2) qui a été plus utilisé pour les études contrastives. Il a donc été non seulement nécessaire de choisir une architecture définitive de modèle de classification, mais aussi de réappliquer les chaînes de traitement utilisées pour les comparaisons décrites à la section 4.1 p.92 pour la prédiction de superdomaines.

Superdomaines

Comme pour les domaines regroupés, une annotation en superdomaines des articles possédant un désignant explicite a pu être obtenue de manière immédiate en appliquant sur les corpus de 58 509 articles pourvus d'un désignant (voir le tableau 4.1 à la page 94) la fonction qui à un désignant normalisé (disponible dans les métadonnées de l'ENCCRE) associait son superdomaine. À partir de ce corpus annoté, de nouveaux modèles de classification utilisant ce jeu des superdomaines ont pu être entraînés en suivant exactement le même processus que précédemment. Au vu des résultats de l'étude comparative précédente, seuls des modèles *SGD+TF-IDF* (figure 4.19a) et *BERT* (figure 4.19b) ont été ainsi entraînés.

Les performances sont légèrement supérieures à celles obtenues sur les domaines regroupés, ce qui est cohérent avec la réduction du nombre de classes (moins il y a de classes, plus il devient difficile, statistiquement, de se tromper). La combinaison *SGD+TF-IDF* obtient des scores très semblables à ceux du modèle *BERT* avec des F-mesures respectives de 85% et 88% sur l'ensemble du jeu de test. Les



(a) Pour la combinaison *SGD+TF-IDF*

(b) Pour le modèle *BERT*

Figure 4.19 – Matrices de confusion des modèles entraînés sur les superdomaines

seuls éléments encore décelables hors de la diagonale sont les 3 colonnes correspondant à l'*Histoire*, les *Métiers* et la *Philosophie* et la ligne correspondant à la *Politique*, tant sur la figure 4.19a que sur la figure 4.19b. Les 3 colonnes montrent que le modèle a souvent tendance à confondre un assez grand nombre de classes avec une des 3 classes *Histoire*, *Métiers* ou *Philosophie*. La ligne montre que la *Politique* reste très mal reconnue par ces modèles avec le nouveau jeu d'étiquettes et que ses articles se retrouvent distribués entre un assez grand nombre de classes dans les prédictions du modèle au lieu d'être correctement assignés à la classe *Politique*. Le phénomène est naturellement moins prononcé pour le modèle *BERT* qui obtient dans l'ensemble de meilleurs résultats mais il reste observable sur les deux matrices.

Considérations énergétiques

Les résultats des évaluations ci-dessus étaient dans une certaine mesure prévisibles et cohérents avec l'état de l'art, mais le choix d'un classifieur repose généralement sur davantage que la seule comparaison de F-mesures. Si l'avantage des transformeurs sur les autres méthodes d'AP ne sont après tout que de quelques décimales, en revanche cette étude fait ressortir un avantage décisif en leur faveur : le peu de données d'entraînement qu'ils nécessitent. Les performances relativement décevantes des méthodes *CNN* et *BiLSTM* peuvent s'expliquer sans doute en grande partie par la rareté relative [11] des données d'entraînement et il est donc d'autant plus remarquable de voir les méthodes basées sur *BERT* s'en sortir aussi bien avec les mêmes quantités de données. Cela souligne l'intérêt du préentraînement de ces modèles, les quelques dizaines de milliers d'articles disponibles étant suffisants pour ajuster finement leurs paramètres.

La comparaison de méthodes traditionnelles d'AA avec ces méthodes d'AP plus récentes a une pertinence au-delà de l'intérêt historique si l'on prend en compte le coût des calculs. Il est bien sûr déjà remarquable de voir des combinaisons de méthodes comme *SGD+TF-IDF* ou *SVM+TF-IDF* obtenir de meilleurs résultats que les méthodes d'AP, pourtant bien plus récentes, à l'exception des transformeurs. Mais cela est d'autant plus impressionnant que leurs besoins en termes de stockage, de mémoire et de temps de calcul sont minimes comparés aux méthodes d'AP. La consommation énergétique occasionnée par l'entraînement et l'évaluation des modèles *SGD+TF-IDF* et *BERT* sur le jeu

11. par opposition aux mégadonnées qui ont nécessité l'invention de ces méthodes d'AP et qui consistent en des ensembles d'au moins plusieurs millions voire milliards d'objets

d'étiquettes des superdomaines a été mesurée à l'aide des outils développés par Lacoste et al. (2019). L'opération a été effectué pour les deux modèles sur la même machine du Centre Blaise Pascal de l'ENS de Lyon (Quemener & Corvellec, 2013) équipée d'une carte GPU de type NVIDIA RTX A2000 12GB et d'un CPU Intel Core i3-2120 d'une fréquence 3.30GHz.

L'ensemble de l'expérience a consommé 594 W·h comme le montre le tableau 4.6, ce qui représente une quantité d'énergie de l'ordre de celle utilisée pour faire fonctionner un réfrigérateur pendant une heure. Cela reste relativement modeste mais cette remarque a le mérite de matérialiser le coût d'un calcul qui peut trop facilement sembler se produire «hors du monde» dans un espace numérique sans conséquences, et ce d'autant plus quand le calcul a lieu sur un serveur distant et n'est pas suivi d'effets directement perceptibles comme le bruit du ventilateur et la chaleur dégagée par la machine depuis laquelle il est lancé. D'après l'Agence Européenne pour l'Environnement, l'électricité en France – où est situé le serveur qui a effectué le calcul – revenait à 68 g.eqCO_2 par kW·h en 2022 [12]. On peut donc estimer que 40.4 g.eqCO_2 ont été rejetés dans le cadre de cette expérience soit l'équivalent d'environ 6km de voyage en Eurostar pour une personne d'après les chiffres de la SNCF [13].

Tableau 4.6 – Coût en énergie des différentes phases de l'entraînement des modèles *SGD+TF-IDF* et *BERT*

Modèle	*SGD+TF-IDF*	*BERT*	Total
Prétraitement	41.3 W·h		41.3 W·h
Vectorisation	$1.59e^{-1}$ W·h		$1.59e^{-1}$ W·h
Entraînement	$4.48e^{-1}$ W·h	545 W·h	546 W·h
Évaluation	$9.60e^{-2}$ W·h	6.83 W·h	6.93 W·h
Total	42.0 W·h	552 W·h	594 W·h

Au-delà de la traduction en un coût physique concret qu'on peut en faire, cette valeur pose en soi plusieurs problèmes d'interprétation, notamment parce qu'elle dépend de l'efficacité de la production électrique à un endroit donné à un moment donné (dans un autre pays, la même expérience n'aura pas le même bilan carbone et rien ne garantit qu'il sera toujours possible d'effectuer cette expérience pour le même bilan carbone si les conditions de production venaient à changer en France) mais aussi parce que la production d'énergie dans certaines conditions a toujours une capacité finie, et que l'énergie utilisée à une certaine fin n'est pas disponible pour d'autres applications – si une grande quantité de personnes venaient à faire ce type d'expérience, l'énergie consommée pourrait venir à entrer en concurrence avec celle utilisée pour les transports ou l'industrie, au point de devoir potentiellement importer de l'électricité produite dans des conditions différentes en termes de bilan carbone (Lacoste et al., 2019, p. 4). Plus que le grammage équivalent CO_2 de l'ensemble, il est donc intéressant de considérer la répartition de l'énergie utilisée entre les différentes phases des différentes méthodes.

D'abord la distinction entre GPU et CPU, bien qu'elle ne soit pas reportée sur le tableau précédent différencie nettement les deux modèles. La consommation énergétique attribuée à la carte GPU pour l'entraînement du modèle *BERT* représente 67.8% de la consommation totale, fraction qui baisse à 27.5% pour le modèle *SGD+TF-IDF*. Disposer d'une GPU est en pratique indispensable pour utiliser un modèle *BERT* alors qu'il est tout à fait possible de s'en passer pour *SGD+TF-IDF*. Mais le point le plus saillant du tableau 4.6 reste la différence d'ordre de grandeur entre la consommation de l'assemblage *SGD+TF-IDF* et celle de *BERT*, d'un facteur de l'ordre de 10. Sur la phase d'évaluation, le facteur est de l'ordre de 70, et sur la phase d'entraînement il atteint même plus de 1 200. En fait, même le coût

12. https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emission-intensity-of-1
13. https://ressources.data.sncf.com/explore/dataset/info-ges/information/

de la préparation du vectoriseur *TF-IDF* additionnée à l'entraînement et l'évaluation du modèle *SGD* a coûté presque 10 fois moins d'énergie que la seule évaluation de *BERT* (0.7 W·h contre 6.83 W·h). Enfin, alors que la phase de prétraitement pour filtrer les mots vides et lemmatiser le reste du texte est souvent perçue comme un préalable décorrélé de l'apprentissage automatique en lui-même, le fait que *BERT* puisse s'en passer est en réalité un avantage important puisque c'est en fait à cette phase que la préparation du modèle *SGD+TF-IDF* consomme la très large majorité (> 98.3%) de l'énergie qu'il requiert. Devant ce constat, un modèle *SGD+TF-IDF* a été réentraîné sur le texte non prétraité, obtenant des résultats tout à fait similaires à ceux obtenus avec prétraitement [14], ce qui s'explique par la capacité de la méthode *TF-IDF* à discriminer les mots présents dans un grand nombre de documents (pour les mots vides, probablement tous). Le modèle obtenu occupe seulement 4% de place en plus et son entraînement a consommé 15% d'énergie en plus ($5.15e^{-1}$ W·h), ce qui amène à s'interroger sur la réelle pertinence de cette phase de prétraitement pour les modèles *SGD+TF-IDF* quand le nombre de prédictions à faire est limité.

Dans le cadre des présents travaux de thèse, il ne s'agit en effet pas de mettre en place un service de classification mais simplement d'appliquer un jeu d'étiquettes une seule fois sur un ensemble d'articles en vue de leur exploitation. Bien entendu, les performances très légèrement – mais strictement – supérieures de *BERT* incitent à le préférer. Mais surtout, le nombre d'occurrences où le modèle est utilisé est donc assez limité et dû aux seules expérimentations de cette étude, et une fois la classification obtenue il n'a plus besoin de servir. Pour cette raison, c'est effectivement le classifieur *BERT* qui a été retenu pour étiqueter l'ensemble du corpus malgré les points soulevés précédemment qui montrent la pertinence de *SGD+TF-IDF*.

Le modèle *BERT* a donc été utilisé pour prolonger la classification en superdomaines sur les articles de l'*EDdA*. La répartition des articles de l'*EDdA* à l'issue de cette classification est visible à la figure 4.20. Suite aux regroupements introduits par les superdomaines toutes les classes à part la *Politique* comportent au moins 1 000 articles.

4.3.2 Extension à *LGE*

Pour *LGE* la situation est un peu plus compliquée. Il a été montré à la sous-section 3.1.2 (p.49) que certains articles étaient pourvus de désignants – bien qu'ils semblent plus rares que dans l'*EDdA* – mais l'encodage des articles de cette encyclopédie demeure trop superficiel et n'y donne pas accès (voir la section 3.1.1 page 46). Pour obtenir une classification de ses articles, il faut donc utiliser les prédictions du modèle *BERT* discuté précédemment dans cette section. Cela pose naturellement un problème épistémologique majeur : l'ensemble de classes des superdomaines a été conçu pour représenter de manière simplifiée les domaines de connaissance du XVIII^{ème} siècle, sans prise en compte particulière de l'état des sciences au XIX^{ème} ni même des domaines choisis par les éditeurs de *LGE*. L'application de ces domaines, de surcroît par un système automatisé, ne peut donc pas se faire sans une phase d'évaluation tâchant de valider la pertinence des superdomaines associés aux articles.

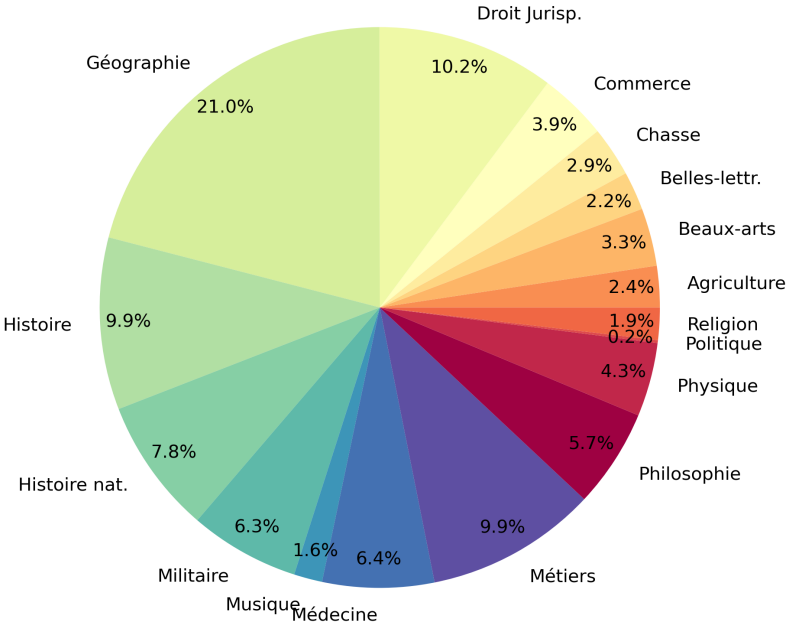
Le fait le plus saillant parmi les prédictions du modèle *BERT* sur les articles de *LGE* est le taux surprenamment élevé d'éléments dans la classe *Commerce*, de l'ordre de 8%. Ce taux est bien supérieur à celui de la même classe dans l'*EDdA* (3.9%) ou même à sa fréquence perçue en parcourant les pages de *LGE*. Un échantillonnage des articles ainsi classés fait ressortir un grand nombre d'articles très courts sur le type de BOULAINCOURT (La Grande Encyclopédie, T7, p.655) présenté à la figure 4.21.

Cet article relève évidemment de la *Géographie* et non pas du *Commerce*. À l'image de la large majorité des articles de l'échantillon, il comprend l'abréviation «Com.», qui signifie évidemment «com-

14. la F-mesure reste à 85%, seuls les scores individuels bougent, certaines classes comme la *Politique* étant moins mal reconnues (17% contre 8%) alors que d'autres perdent un peu comme le *Commerce* (77% contre 79%) – la *Géographie* est essentiellement inchangée (97% contre 96%)

Domaine	Taille
Politique	124
Musique	1213
Religion	1380
Belles-lettres	1652
Agriculture	1769
Chasse	2138
Beaux-arts	2474
Commerce	2930
Physique	3160
Philosophie	4223
Militaire	4709
Médecine	4781
Histoire naturelle	5797
Métiers	7332
Histoire	7361
Droit Jurisprudence	7600
Géographie	15 547

(a) Décompte absolu



(b) Proportions relatives

Figure 4.20 – Nombre d'articles par domaine dans l'*EDdA*

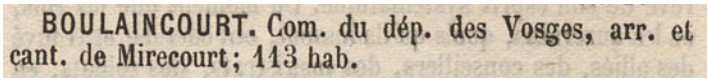

BOULAINCOURT. Com. du dép. des Vosges, arr. et cant. de Mirecourt; 113 hab.

Figure 4.21 – Article BOULAINCOURT, *LGE*, T7, p.655

mune» dans le contexte. Le terme est anachronique dans ce sens au XVIII^{ème} siècle : comme le montre l'article COMMUNE ou COMMUNES (L'Encyclopédie, T3, p.725), il désigne alors une association d'habitants d'une localité, et pas la localité elle-même. Cette entrée possède d'ailleurs le désignant «Jurisp.» et relève clairement du droit plus que de la géographie. Il est difficile de déterminer avec certitude les raisons qui ont orienté le choix du modèle mais les occurrences répétées de «Com.» pouvant évoquer un désignant mal filtré dans ces articles incorrectement classés en *Commerce* sont pour le moins troublantes. Il n'est pas facile non plus de tester cette hypothèse en contrôlant la lemmatisation de cette abréviation puisque *BERT* attend en entrée le texte de l'article. Il ne procède pas à un découpage en mots en interne (il n'y a donc pas de lemmatisation) mais se contente de grouper des graphèmes qui apparaissent fréquemment ensemble (algorithme WordPiece).

De manière pragmatique, la solution la plus immédiate pour corriger le problème au vu de la régularité de ces articles relativement brefs et possédant le motif «Com. du dép.» («commune du département», autre anachronisme) sur leur première ligne est de les identifier à l'aide d'une expression régulière et d'un filtrage sur leur taille. En effet, le motif peut survenir dans des articles plus généraux sur un département, comprenant notamment des sections historiques ou économiques et les erreurs de segmentation (voir la section 3.1.3 page 61) peuvent préfixer de tels articles géographiques brefs à d'autres articles sur des sujets divers. Pour éviter d'introduire trop de bruit dans de futurs usages qui pourraient être faits de ces articles, un seuil de 50 tokens est fixé empiriquement après échantillonnage d'articles ayant une occurrence de l'expression précédente sur leur première ligne.

Plus précisément, le motif est sujet à une certaine variation sans doute en partie du fait de l'envergure de projet de *LGE* – par application de ce constat empirique formulé à la section 3.3.3 qui énonce qu'à cette échelle tout motif subit des mutations (voir p.87) – mais surtout à cause des erreurs d'OCR qui fournissent un rendu imprécis des caractères présents sur le papier. Certains 'C' majuscules, peut-être à cause d'une tache sur le papier ont été lus comme 'G', certains '.' à la fin des abréviations ont pu être pris pour des ',' voire être omis entièrement. Pour ces raisons, l'expression régulière écrite pour rechercher ce motif a été largement étendue pour s'adapter à tous les cas rencontrés empiriquement dans les articles comme le montre l'extrait de code source 4.1.

Code source 4.1 – Expression régulière utilisée pour repérer les entrées de communes

```
[cCG]\(o\(m\|[ri]n\)\{1,2\}\|ant\)[-.,]\?  \?d[ue]  \?d[ée]p[-,.]\?
```

En lisant les blocs à partir des plus profonds pour remonter à l'expression dans son ensemble, voici comment elle peut s'analyser. D'abord, la lettre 'm' est parfois perçue par l'OCR comme la séquence ”rn” voire ”in”, ce qui est à l'origine de la séquence `\(m\|[ri]n\)` qui sert à capturer «un m» : soit R_m cette séquence. L'abréviation n'est pas toujours la même et il y a plusieurs occurrences où les deux 'm' du mot «commune» ont été conservés : c'est là le sens de R_m`\{1, 2\}` qui tolère une séquence de 1 ou 2 fois l'expression précédente pour absorber les différentes réalisations possibles de «m» ou «mm». Ce bloc, préfixé du o reconnaît évidemment la séquence «omm» (et les différentes combinaisons qui peuvent le remplacer, donc) : ce sera $R_{omm?}$ dans ce qui suit. L'ensemble `\(`$R_{omm?}$`\|ant\)` reconnaît donc la séquence précédente ou bien «ant» : cette disjonction a été ajoutée pour prendre en compte quelques villes qui n'étaient pas mentionnées en tant que «commune» mais en tant que chef-lieu de canton «Ch.-l. de cant. du dép.» comme l'article HEYRIEUX (La Grande Encyclopédie, T20, p.60) présenté à la figure 4.22 [15]. Le préfixe `[cCG]` sert à capturer un ”C”, qui est parfois lu ”c” ou ”G” par l'OCR, pour former $R_{Comm?|Cant}$. Le motif `[-.,]\?`, qui survient juste après $R_{Comm?|Cant}$ ainsi qu'à la toute fin de l'expression, permet d'absorber différentes lectures du '.' (point), parfois ”,” voire ”-”, ainsi que son absence potentielle (`\?`). Ce `\?` survient aussi après les deux espaces qui manquent entre les mots de l'expression dans certains textes. Les deux derniers ensembles de caractères prennent en compte ”du” et ”de” (on trouve «du dép.» ou «de dép.») et bien sûr l'absence possible de l'accent de «dép.».

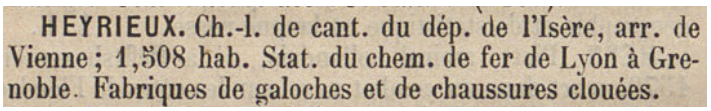

HEYRIEUX. Ch.-l. de cant. du dép. de l'Isère, arr. de Vienne ; 1,508 hab. Stat. du chem. de fer de Lyon à Grenoble. Fabriques de galoches et de chaussures clouées.

Figure 4.22 – Article HEYRIEUX, *LGE*, T20, p.60

À l'issue de ce traitement, 20 116 articles sont identifiés (soit 14.9% du total des articles de *LGE*) et la classe *Géographie* leur est assignée manuellement. Ces articles constituent un sous-corpus des Communes qui ne fait pas l'objet d'une exploitation particulière dans le cadre de cette thèse mais qui pourrait intéresser de futures études. Il serait également intéressant d'utiliser ces articles pour améliorer le modèle de classification mais cela reste à faire.

La classification utilisée pour les articles de *LGE* est donc composite : elle assigne la classe *Géographie* aux articles du sous-corpus des Communes défini ci-dessus, et la classe prédite par le modèle *BERT* entraîné pour les superdomaines sur l'*EDdA* pour les autres articles. Une fois cette définition posée, il devient possible d'évaluer la qualité de cette annotation afin de juger de son utilité. Le diffé-

15. Il est difficile de savoir pourquoi ces entrées peuvent être également affectées par l'erreur de classification en *Commerce* puisqu'elles ne comportent pas le motif «Com.». Cette erreur paraît moins systématique que pour les entrées où ce motif apparaît et, s'agissant de villes plus importantes (puisque chefs-lieux), celles qui subissent cette erreur de classification mentionnent souvent des produits issus de la ville – comme les chaussures de la figure 4.22 – ce qui pourrait fournir une piste d'explication.

rentiel conceptuel entre ces classes prévues pour le XVIII^{ème}s. et ces articles rédigés au XIX^{ème}s. empêche naturellement les articles de *LGE* de correspondre parfaitement aux classes proposées. Ainsi, la présence des biographies discutées précédemment (à la section 3.1.2 p.49) perturbe évidemment le système de classes des superdomaines modelé sur l'*EDdA* qui les exclut volontairement ; de même des domaines anachroniques comme l'Industrie n'existent pas parmi les superdomaines. Une telle évaluation ne peut donc pas exiger une rigueur comparable à celle menée sur les modèles de classification comparés à la section 4.1.3 (p.94 et seq.) où il s'agissait essentiellement de «compter les points», mais doit au contraire se contenter de déterminer si la classe proposée fait assez sens pour permettre d'opérer des comparaisons intéressantes dans les analyses contrastives du chapitre suivant.

Dans cette optique, un échantillon de 40 articles pour chacun des 17 superdomaines (soit $n = 680$ articles au total) a été sélectionné aléatoirement pour une validation manuelle des superdomaines qui leur étaient associés. Plusieurs problèmes gênent aussi l'évaluation, dont les problèmes de segmentation qui affectent encore un nombre perceptible d'articles de *LGE* ce qui cause la présence d'agrégats de plusieurs articles parmi les textes de l'échantillon pour lesquels la définition d'une classe fait peu sens. La présence d'articles qui sont de purs renvois sans contenu propre rend également difficile la détermination d'une classe sans ambiguïté. Pour toutes ces raisons, l'évaluation a été plutôt permissive, c'est-à-dire que la «charge de la preuve» a plutôt pesé sur le rejet de la classification proposée par le modèle que sur son acceptation : le refus d'une annotation suggérée par le modèle devait être motivé par au moins un superdomaine qui conviendrait mieux de manière évidente. Pour un article qui aurait accepté plusieurs classes, la proposition du modèle a été acceptée si l'une au moins correspondait (même si ça n'était pas forcément la première mentionnée dans le texte). C'est ainsi que la plupart des biographies se retrouvent en *Histoire* ou dans la discipline où la personne s'est illustrée. Pour les renvois, la classe proposée a été acceptée sauf quand la cible du renvoi relevait manifestement d'une autre classe. Pour les agrégats de plusieurs articles, la classification a été rejetée quand vraiment aucune des entrées contenues n'avait de rapport avec la classe proposée et acceptée sinon.

Avec cette convention, la méthode composite (*Géographie* pour les Communes, la prédiction de *BERT* pour les autres articles) a prédit une classe acceptable pour 590 articles de l'échantillon, soit 86.8% de la totalité. S'agissant d'une évaluation binaire, qui se contente de dire si la classe est acceptable sans fournir de réponse attendue dans le cas contraire, il n'est pas possible de calculer de score de rappel ni de générer de matrice de confusion pour observer les biais du modèle. Le score mesuré correspond à une précision puisqu'il compte le nombre de vrais positifs divisé par le nombre total d'élements. Le tableau 4.7 montre que le modèle obtient une assez bonne précision sur la plupart des domaines (supérieure à 90% sur 7 domaines). Les domaines qui mettent le plus le modèle en difficulté sont la *Chasse*, le *Commerce* et dans une moindre mesure le domaine *Militaire*. La *Politique*, qui est pourtant si mal reconnue par le modèle dans l'*EDdA* obtient le score tout à fait surprenant de 95%. D'une part, il est à noter que lors de l'évaluation sur l'*EDdA*, la précision – certes bien plus faible – était tout de même de 56%, et c'est le très mauvais rappel sur cette classe qui donnait une F-mesure aussi faible. D'autre part, les articles de *LGE* classés en *Politique* sont assez prototypiques, très longs pour la plupart, ou consistent en des biographies d'hommes politiques (et sans doute un grand nombre d'articles validés auraient aussi pu être acceptés en *Histoire*, donc il faut garder à l'esprit que les prédictions du modèle sont par construction acceptables pour un humain, mais pas nécessairement comparables à celles faites sur l'*EDdA*).

Le nombre relativement élevé d'articles sélectionnés pour cette évaluation permet d'accorder une certaine confiance aux résultats obtenus. S'agissant d'une évaluation binaire, le processus peut se modéliser par le tirage sans remise (chaque article ne peut être sélectionné qu'une fois dans l'échantillon) d'une variable aléatoire booléenne pour chaque article, pouvant prendre la valeur `vrai` si la classification proposée par le modèle pour cet article est jugée pertinente et `faux` sinon. Il s'agit de variables de Bernoulli, non indépendantes en toute rigueur : la probabilité que la première variable

Domaine	Précision
Chasse	67.5
Commerce	70.0
Militaire	77.5
Histoire	80.0
Physique	82.5
Agriculture	85.0
Droit Jurisprudence	85.0
Métiers	85.0
Belles-lettres	87.5
Philosophie	87.5
Médecine	92.5
Religion	92.5
Politique	95.0
Géographie	95.0
Beaux-arts	97.5
Musique	97.5
Histoire naturelle	97.5

Tableau 4.7 – Scores de précision par domaine obtenus par le modèle *BERT* pour la classification en superdomaines sur l'échantillon de *LGE*

vale `vrai` est égale à la précision du modèle dans son ensemble, mais en fonction du résultat de ce premier tirage, la probabilité de succès de la deuxième variable est affectée. Toutefois, les dimensions de l'expérience font que l'échantillon reste assez petit par rapport à l'ensemble de la population ($100 \times 680 = 68\,000 < 130\,000$) donc les variables peuvent être considérées indépendantes (le résultat de chaque tirage affecte peu la proportion d'articles dont la classification est satisfaisante parmi les articles restants). Le Théorème Central Limite s'applique donc et permet de modéliser le comportement de la somme de ces variables par une loi normale.

La moyenne empirique m constitue un estimateur sans biais de la proportion réelle p. C'est donc la valeur la plus probable compte tenu du résultat de l'expérience, c'est-à-dire que toute autre valeur, prise individuellement, est moins probable que m mais pas impossible. La vraisemblance de l'estimation dépend de la taille de l'intervalle que l'on considère autour de m. Si l'on prend un intervalle trop petit (par exemple $10^{-6}\%$) il est plus probable que p soit hors de cet intervalle que dedans. En particulier, il est vain de croire que p puisse être exactement égale à m (intervalle de taille nulle) : bien que ce soit la valeur la plus vraisemblable, la probabilité de ce nombre précis est nulle. L'ensemble des valeurs réelles possibles pour p suit une distribution en cloche autour de la valeur m, la loi normale, dont la probabilité décroît au fur et à mesure que l'on s'en éloigne. Il faut donc accepter de perdre en précision pour obtenir un intervalle assez probable pour connaître p avec une confiance suffisante. La taille du demi-intervalle dépend de la moyenne empirique m, de la taille de l'échantillon n et d'un facteur multiplicatif lié seulement pour une loi donnée à la précision attendue, son quantile. Ainsi pour que p ait plus de 95% de chance de faire partie de l'intervalle proposé (encadrement à 95% de confiance), le quantile $z_{97.5\%}$, d'ordre $1 - \frac{5}{2} = 97.5\%$ permet d'écrire en reprenant n pour désigner le nombre total d'articles échantillonnés, l'encadrement de la formule 4.5.

$$m - z_{97.5\%} \times \sqrt{\frac{m \times (1-m)}{n}} \leq p \leq m + z_{97.5\%} \times \sqrt{\frac{m \times (1-m)}{n}} \tag{4.5}$$

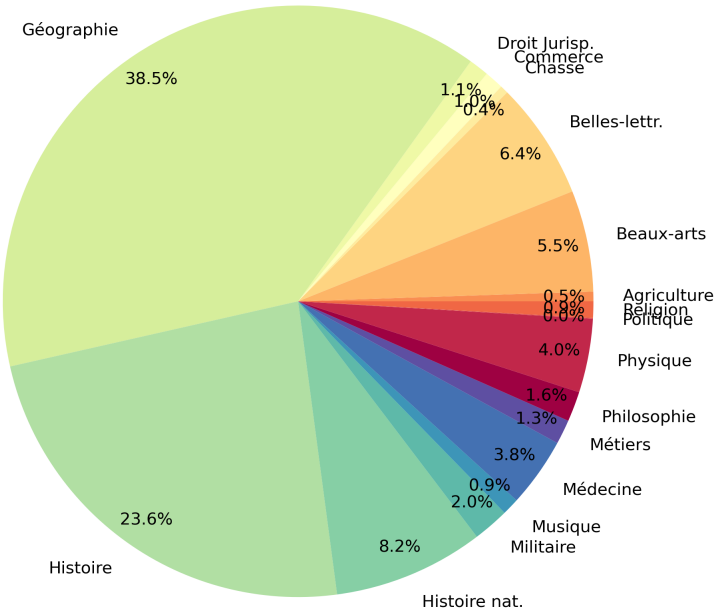
Une application numérique à partir des valeurs $m = \frac{590}{680}$, $n = 680$ et $z_{97.5\%} = 1.96$ (donnée par les tables numériques de la loi normale) permet d'obtenir l'encadrement 4.6. Cet encadrement signifie que l'étiquetage par le modèle des domaines de connaissance dans *LGE* a 95% de chances d'avoir une qualité au moins égale à 84.2% et inférieure à 89.3%.

$$84.2\% \leq p \leq 89.3\% \tag{4.6}$$

La formule 4.5 montre que la largeur de l'intervalle dépend de la taille de l'échantillon n. Par conséquent, la précision de l'encadrement est moins bonne pour chaque classe individuellement (pour lesquelles $n = 40$) que pour l'ensemble des articles. La présence du produit $m \times (1 - m)$ au numérateur du demi-intervalle dans 4.5 montre également que la précision est meilleure pour les classes les mieux reconnues (car ce produit toujours positif s'annule en 0 et en 1 et atteint sa valeur maximale – 0.25 – en $m = 0.5$). Ainsi, avec le même degré de confiance de 95%, il est par exemple possible d'affirmer que la précision du modèle sur la classe *Histoire naturelle* est d'au moins 92.7%. Par contraste, la précision la moins bonne mesurée sur la classe *Chasse* (67.5%) ne peut pas être encadrée aussi finement et pour conserver le même indice de confiance il n'est possible de fournir que la valeur 52.3% comme borne inférieure (avec optimisme, le même argument renversé implique que seules des valeurs au-delà de 82.0% ont moins de 5% de chance d'être la vraie précision du modèle sur cette classe). Pour la classe *Géographie* qui intéresse principalement les présents travaux, il est possible d'affirmer à 95% de confiance que la précision est au moins supérieure à 88.2%. Les bons résultats du modèle sur la très large majorité des classes – et la *Géographie* en particulier – permettent de valider l'emploi de cette annotation pour les analyses conduites tout au long du chapitre 5. La répartition des nombres d'articles de *LGE* par superdomaine est donnée par la figure 4.23.

Domaine	Taille
Politique	64
Chasse	605
Agriculture	706
Religion	1212
Musique	1256
Commerce	1404
Droit Jurisprudence	1471
Métiers	1802
Philosophie	2216
Militaire	2660
Médecine	5116
Physique	5429
Beaux-arts	7466
Belles-lettres	8660
Histoire naturelle	11 105
Histoire	31 779
Géographie	51 869

(a) Décompte absolu



(b) Proportions relatives

Figure 4.23 – Nombre d'articles par domaine dans *LGE*

Cette section conclut les travaux en classification automatique effectués dans le cadre de cette thèse.

La comparaison de classifieurs opérée à la section 4.1 a montré la pertinence de l'emploi d'un modèle de type *BERT* pour appliquer une classification en domaine aux articles. Plus inattendu, cette comparaison a aussi révélé l'intérêt spécifique de méthodes classiques d'AA telles que *SVM* et *SGD* associées à une vectorisation *TF-IDF* sur des volumes de données de l'ordre de grandeur de ceux présents dans le corpus d'étude. Ces modèles nécessitent en effet relativement «peu» de données pour produire des résultats exploitables par rapport aux méthodes d'AP autres que *BERT*. La sous-section 4.3.1 a en outre mis en évidence le fait que *SGD*, certes légèrement moins performant que *BERT*, s'avère en plus particulièrement efficace du point de vue de la consommation énergétique. Les études conduites à la section 4.2 sur les erreurs faites par ce modèle sur les articles de l'*EDdA* suggèrent des pistes intéressantes à explorer dans le chapitre 5, notamment sur la présence de contenus non géographiques dans les articles de la classe *Géographie*. À l'issue de ces travaux de classification, la totalité des articles du corpus – aussi bien ceux issus de l'*EDdA* que de *LGE* – a été classée selon ces superdomaines, ce qui rend possible les études contrastives développées dans le dernier chapitre.

Chapitre 5

Études contrastives

5.1 Tracer les contours de la géographie 133
5.1.1 La place de la géographie 134
5.1.2 Des articles qui changent de domaine 146
5.2 La biographie cachée 157
5.2.1 Spécificité des lemmes «naître» et «mourir» en *Géographie* 158
5.2.2 L'influence des domaines 164
5.2.3 Différents profils de passages biographiques 166

Ce dernier chapitre qui s'ouvre propose une méthodologie fondée sur une analyse du corpus à différents niveaux de granularité pour montrer la pertinence des données préparées dans le cadre de cette thèse. Le chapitre 3 fournit un corpus structuré, annoté en syntaxe et développe la notion de domaine de connaissance, centrale pour la classification qui est appliquée aux articles au chapitre 4. Ces efforts amènent deux directions de travail pour le présent chapitre.

En gardant bien à l'esprit les réserves émises à l'Introduction sur la différence entre d'une part la Géographie en tant que discipline et d'autre part les discours qui peuvent relever de la géographie (voir 1.2 p.11), l'examen de mesures effectuées à l'échelle des articles permet de mettre en lumière des différences en diachronie entre l'*EDdA* et *LGE* ainsi qu'entre domaines de connaissance. La première étape que représente la section 5.1 consiste donc à estimer les variations de la place accordée à la discipline géographique dans les œuvres du corpus au travers de mesures des articles de *Géographie* aux XVIII^{ème} et XIX^{ème} siècles.

Dans un second temps, une étude plus fine des textes basée sur des phénomènes au niveau des phrases, notamment syntaxiques, révèle des mélanges thématiques au sein des articles. La section 5.2 permet ainsi d'identifier des biographies au sein des articles de *Géographie* et d'étudier la relation particulière qu'entretiennent ces contenus avec la discipline.

5.1 Tracer les contours de la géographie

Cette première section articule deux approches complémentaires. Une première étude purement quantitative établit le périmètre accordé à la *Géographie* et aux autres domaines de connaissance dans l'*EDdA* et dans *LGE*. Dans un second temps, une étude qualitative vient compléter la précédente en regardant l'évolution du traitement qui est fait d'un ensemble fixe de thématiques par rapport à

la notion de domaine de connaissance.

5.1.1 La place de la géographie

La revue quantitative qui débute s'intéresse à plusieurs critères, du nombre d'articles consacrés à chaque domaine aux densités d'entités nommées qu'on peut y rencontrer.

Tailles

À l'issue des sections 4.3.1 et 4.3.2, la répartition du nombre d'articles par domaine était connue. Mais un parcours rapide des pages de l'*EDdA* comme de *LGE* montre une grande disparité de taille d'une entrée à une autre. Le phénomène a déjà été décrit par Laramée (2017, p. 168) pour la géographie dans l'*EDdA* où il peut au moins s'expliquer par des différences de conceptions très marquées entre les auteurs quant aux buts de la Géographie, particulièrement entre Diderot et le chevalier de Jaucourt. Mais il s'observe également pour des articles d'autres domaines, ainsi que dans toute *LGE*. Il est dès lors naturel de s'intéresser aux éventuels liens entre taille des articles et domaines de connaissance.

À plusieurs échelles, plusieurs métriques peuvent être considérées. La mesure qui se présente le plus immédiatement à l'esprit pour mesurer la taille d'un ensemble de textes est sans doute le nombre d'unités lexicales qu'il contient. Sous une apparente simplicité c'est un problème intrinsèquement «délicat» (voir la section 2.4.3 p.34). En diachronie, la question se complique encore car la segmentation «naturelle» pour les locuteurs change suivant les époques des parties du corpus. Si le français classique de l'*EDdA* diffère heureusement assez peu de celui déjà quasi contemporain de *LGE*, on y trouve encore tout de même des formes similaires à celles rapportées par Diwersy et al. (2017, p. 29). Ainsi, l'adverbe «très» demeure préfixé à de nombreux adjectifs, bien que séparé d'un tiret [1] dans des formes comme «très-subtil» à l'article MALEBRANCHISME (L'Encyclopédie, T9, p.942) ou «très-philosophique» dans l'article DICTIONNAIRE (L'Encyclopédie, T4, p.958) cité à la section 2.2.1 (p.24). À l'inverse certaines formes ne sont pas encore agglutinées telles que «par ce que» dans l'entrée ATTAQUES *d'une place* (L'Encyclopédie, T1, p.829).

Les «mots» sont donc une abstraction utile mais inatteignable en pratique : la linguistique de corpus requiert de faire des choix de segmentation toujours teintés d'une part d'arbitraire pour produire une séquence de tokens que les outils numériques sont à même de manipuler. Ceux opérés dans le cadre de ces travaux suivent simplement les décisions de l'annotateur Stanza qui intègre la phase de tokenisation (voir la section 3.3.1 p.81). Entraîné pour du français contemporain, il laisse l'amalgame «très-subtil» en un seul token (qu'il annote correctement comme un adjectif alors que la forme «très-opiniâtre» qui apparaît peu après dans le texte est étiquetée en tant que nom commun), et ne reconnaît pas «parce que», considérant la séquence «par ce que» comme trois tokens distincts. Le seul choix restant à l'issue des traitements concerne les amalgames, représentés à l'aide de sous-tokens disponibles dans le format CoNLL-U. À la suite de Vigier (2017, p. 101), un seul token a été produit pour les représenter dans des outils permettant des recherches en surface comme TXM. Les POS et lemmes affectés à ces tokens sont obtenus en concaténant les POS et les lemmes de leurs sous-tokens, séparés par un caractère '+'. C'est à ces conventions que renvoient les nombres de «mots» considérés dans cette partie.

Mais la taille des textes peut également être considérée à l'échelle plus fine des signes, en comptant précisément la place qu'ils occupent et là aussi, cette intention peut être réalisée au moins de deux manières. L'emploi d'outils informatiques amène à utiliser la place occupée sur le disque, d'accès immédiat puisque disponible parmi les métadonnées exposées par le système de fichier au même titre que son nom ou sa date d'accès, sans qu'il soit même nécessaire d'en lire le contenu. Du point

1. en moyen français on trouve des formes simplement agglutinées comme «trèsardent» rapportée par Diwersy et al. (2017, p. 29)

de vue de la complexité, cette information est donc accessible en temps constant par rapport à la taille du fichier ($O(1)$). Le nombre d'octets occupés par un fichier contenant le texte d'un article dépend de certaines conventions discutées à la section 3.2.3 notamment pour l'encodage des caractères ou la représentation des fins de ligne. Si les fichiers créés pour *LGE* les réalisent par des `"\n"` et sont distribués ainsi, il n'est pas exclu que certaines plateformes ou éditeurs de texte les remplacent par des `"\r\n"` en les enregistrant, ce qui changerait leur contenu et donc leur taille en octets. De plus, si au lieu d'utiliser cette version ils sont recréés à partir des fichiers ALTO avec `soprano`, les choix de la plateforme sur laquelle le traitement a lieu pourront de la même façon avoir une influence sur la taille des fichiers produits. Enfin pour l'*EDdA*, puisqu'il n'est de toute façon pas possible de redistribuer les fichiers, il n'est pas possible de garantir les choix d'encodage de l'éventuelle version qui pourrait être récupérée par quelqu'un tentant de reproduire ces travaux. Le nombre d'octets, s'il est très simple d'accès, ne constitue donc pas un invariant absolu des textes.

D'un point de vue plus typographique, la place occupée sur les pages par un article correspond davantage au nombre de caractères que contient son texte. Ce dernier est en revanche plus long à calculer puisqu'il faut parcourir tout le contenu du fichier une fois, la complexité est donc ici linéaire ($O(n)$ où n représente la longueur du fichier). Ce calcul peut toutefois être fait une seule fois pour chaque article lors de son indexation et ce choix ne pénalise donc pas les nombreuses comparaisons faites ultérieurement. C'est donc ce nombre de caractères qui est utilisé à l'échelle des signes, mais il est affecté par les mêmes paramètres que le décompte des octets pour les mêmes raisons car il est impossible de se défaire totalement des détails d'implémentation en linguistique outillée. En outre ce décompte inclut la présence des caractères de fin de ligne. En effet, il apparaîtrait difficilement justifiable de compter les espaces, qui prennent de la place entre deux mots sur la même ligne mais pas ces autres séparateurs alors que deux mots séparés par la fin d'une ligne prennent visuellement plus de place sur une page qu'un seul mot dont la longueur serait la somme des longueurs de ces deux mots. Il est également à noter que tous les fichiers se terminent par une fin de ligne, ce qui ajoute un caractère supplémentaire au décompte.

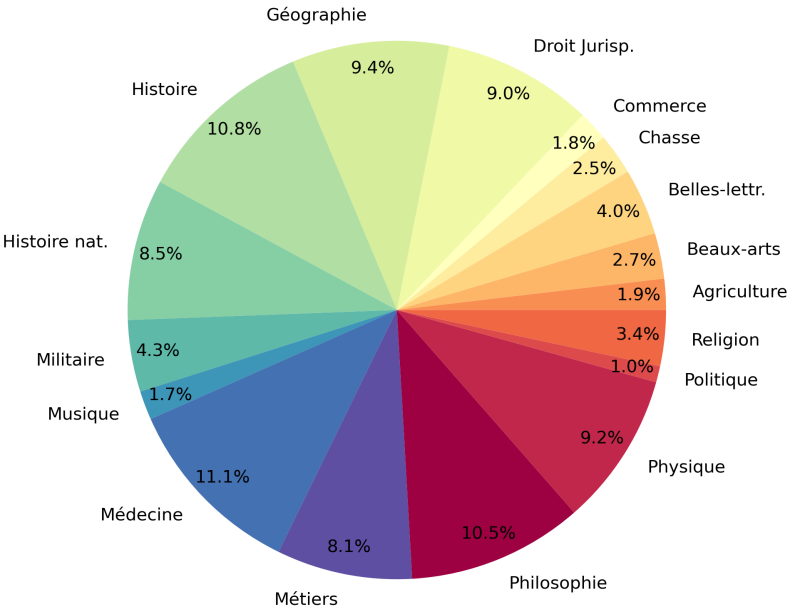
Nombre de mots

La chose la plus évidente lorsque l'on compare la figure 5.1 représentant le nombre de mots par domaine dans l'*EDdA* avec celui représentant le nombre d'articles pour la même partition (figure 4.20 p.127) est la faible importance relative des articles de géographie en termes de taille. En nombre d'articles, la *Géographie* est très nettement le domaine le plus représenté avec plus de 20% des articles de l'*EDdA* qui lui sont consacrés, soit presque deux fois plus d'articles que le deuxième domaine le plus représenté, *Droit et Jurisprudence*. Malgré cette très forte présence, ses articles ne contiennent que 9.3% du nombre de mots, ce qui n'en fait que le 4$^{\text{ème}}$ domaine le plus volumineux, avec de surcroît un nombre de mots assez voisin de celui des 5$^{\text{ème}}$ et 6$^{\text{ème}}$ domaines selon le même critère (respectivement *Physique* et *Droit et Jurisprudence*). Il est même remarquable que la *Géographie* soit couverte en moins de mots que la *Philosophie*, à laquelle ne sont consacrés que 5.47% des articles, soit près de 4 fois moins qu'à la *Géographie*. Les articles de géographie apparaissent donc comme particulièrement brefs par rapport à ceux des autres domaines de connaissance.

Cette caractéristique est encore plus visible sur la figure 5.2 qui représente la distribution des nombres de mots par article dans chacun des domaines pour l'*EDdA*. Le centre de masse de la *Géographie* y est particulièrement décalé vers les petites valeurs de nombre de mots et, alors que l'échelle logarithmique rend la plupart des distributions assez plates, les faisant ressembler à des crayons avec un long corps et une mine plus ou moins affûtée, la distribution de la *Géographie* présente une forme assez particulière, directement décroissante (une «mine» gigantesque sans «corps»). La pente de décroissance est plus ou moins visible selon les domaines mais la plupart présentent au moins cet effet de rupture de pente (la «mine» dans la métaphore précédente), par exemple les domaines *Histoire* et *Militaire*. Seul le domaine *Métiers* semble comme la *Géographie*

Domaine	Taille
Politique	232 767
Musique	412 388
Commerce	432 421
Agriculture	443 352
Chasse	591 949
Beaux-arts	651 842
Religion	797 139
Belles-lettres	948 090
Militaire	1 024 581
Métiers	1 936 895
Histoire naturelle	2 018 865
Droit Jurisprudence	2 137 346
Physique	2 194 062
Géographie	2 241 549
Philosophie	2 506 523
Histoire	2 569 751
Médecine	2 647 193

(a) Décompte absolu



(b) Proportions relatives

Figure 5.1 – Nombre de mots par domaine dans l'*EDdA*

décroître constamment, mais sa pente est tout de même moins forte que celle de la *Géographie* et il possède bien plus d'articles avec un nombre élevé de mots (la différence devient flagrante au-dessus de quelques centaines de mots). La figure 5.2 montre la raréfaction des articles de *Géographie* en raison logarithmique de leur taille (la mine présente des bords bien droits caractéristiques d'une relation linéaire). La différence de densité entre les articles de 100 mots et ceux de 10 mots est par exemple sensiblement la même qu'entre les articles de 1 000 mots et ceux de 100 mots.

Il est naturel après avoir fait cette remarque de se demander si cette signature particulière persiste 130 ans plus tard. La taille de la *Géographie* dans *LGE* aussi est bien moins importante en nombre de mots qu'en nombre d'articles. Elle représente en effet une part encore plus importante du nombre d'articles (37.8% soit plus d'un tiers alors qu'elle ne représentait «que» 20.74% des articles de l'*EDdA*), mais cette fois encore cela ne suffit pas à en faire le domaine le plus volumineux : malgré cette surreprésentation accrue, la *Géographie* n'arrive qu'à la deuxième place des disciplines occupant le plus de mots en comprenant «seulement» 22.87% du texte. Là encore, les articles de *Géographie* apparaissent donc relativement brefs par rapport à ceux des autres disciplines.

Il faut toutefois remarquer que les proportions d'articles classés en *Histoire* et en *Géographie* semblent en contradiction avec les intentions des auteurs de *LGE* dans l'Avant-Propos (La Grande Encyclopédie, T1, p.XI) mentionnées à la section 3.1.2 (p.55) puisqu'il y est question de 15% du volume consacré à *Histoire et Géographie*, alors que la seule *Géographie* prédite par le modèle en totalise déjà une fois et demie plus. S'il faut bien sûr garder en tête que l'ensemble de 17 domaines utilisés dans cette étude ne correspond pas au découpage en 15 «sciences» de l'Avant-Propos, il reste vrai que le modèle a néanmoins prédit qu'un article sur deux appartenait à la *Géographie* ou à l'*Histoire*, et devrait donc être compté dans *Histoire et Géographie*, ce qui ferait un volume bien trop élevé. Parmi les raisons qui pourraient expliquer cet écart se trouve l'inadéquation potentielle d'un modèle entraîné sur l'*EDdA* pour traiter des textes de *LGE*. Un modèle d'apprentissage automatique intègre un certain nombre de biais statistiques du corpus sur lequel il s'entraîne et a tendance à les reproduire sur les données sur lesquelles on l'applique. Toutefois, le déséquilibre prédit en faveur de la *Géographie* est encore

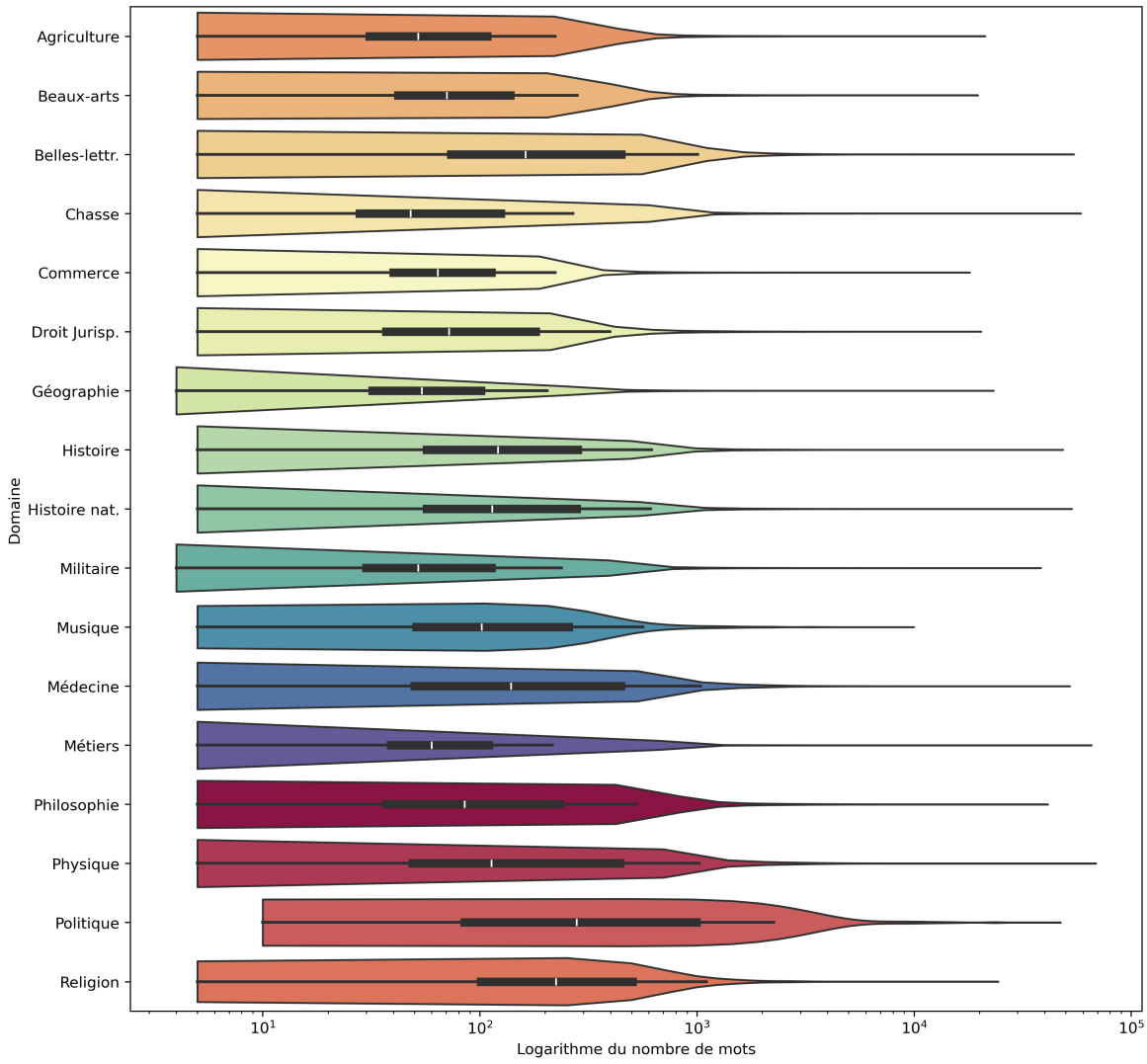
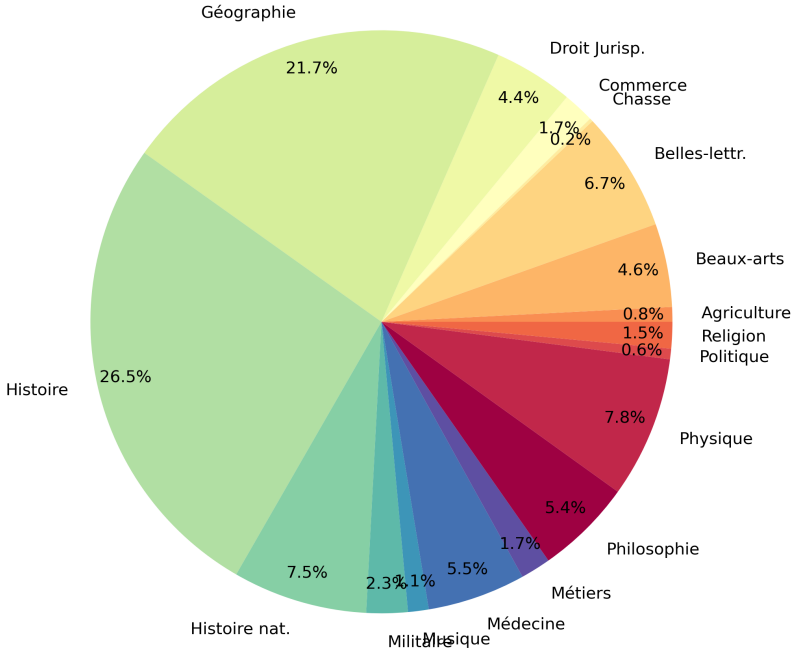


Figure 5.2 – Distribution des nombres de mots par article au sein des différents domaines de l'*EDdA*

plus important que celui constaté dans l'*EDdA*, et de plus l'*Histoire* était moins présente que la *Géographie* alors qu'elle l'est encore davantage dans les prédictions du modèle sur *LGE*. Si les erreurs du modèle pourraient sans aucun doute gonfler un peu les chiffres, elles paraissent difficilement pouvoir changer 15% en 50%. Une autre explication vraisemblable pourrait résider dans la difficulté à décider à l'avance avec assez de précision du contenu d'une œuvre aussi vaste qu'une encyclopédie, impliquant autant de contributeurs sur une période aussi longue. D'ailleurs, le même Avant-Propos annonce 25 tomes de 1200 pages chacun, quand on sait qu'il y en a eu finalement 31 au total, dont le dernier dépasse allègrement les 1 300 pages. Mais là encore, même si l'*Histoire et Géographie* était effectivement prévue pour être la science occupant le plus de place dans *LGE*, on a peine à croire qu'elle ait pu plus que tripler de volume sous les plumes de ses contributeurs. Il n'est pas possible de conclure avec les données disponibles quant aux raisons de cette différence, mais c'est un élément à garder en tête qui sera utile tout au long des réflexions de ce chapitre.



Domaine	Taille
Chasse	101 165
Politique	321 907
Agriculture	447 881
Musique	629 295
Religion	809 072
Métiers	919 074
Commerce	955 416
Militaire	1 259 905
Droit Jurisprudence	2 417 252
Beaux-arts	2 539 133
Philosophie	2 947 525
Médecine	2 994 261
Belles-lettres	3 653 929
Histoire naturelle	4 109 638
Physique	4 310 962
Géographie	11 947 463
Histoire	14 572 388

(a) Proportions relatives

(b) Décompte absolu

Figure 5.3 – Nombre de mots par domaine dans *LGE*

Quoi qu'il en soit, en s'intéressant aux distributions d'articles en fonction de leur nombre de mots, la figure 5.4 montre cette fois un profil assez différent et beaucoup moins marqué où davantage de disciplines voient leurs articles diminuer en nombre avec l'augmentation de leur taille mais avec des pentes bien moins prononcées. Ainsi, si elles semblent s'accentuer légèrement à partir de 1 000 mots, les pentes de décroissance en fréquence de l'*Histoire* et de l'*Histoire naturelle* sont assez droites et ressemblent globalement à celle de la *Géographie*, dont la pente, si elle semble là aussi régulière, est bien moins marquée, signe d'un étalement bien plus progressif. De plus, les nombres de mots minimum par domaine varient bien plus (ce qui se traduit sur la figure 5.4 par le fait que les «crayons» sont bien moins alignés à gauche qu'ils ne l'étaient pour l'*EDdA* à la figure 5.2). Il y a donc plus d'articles de géographie longs dans *LGE* qu'il n'y en avait dans l'*EDdA* et surtout, la répartition en nombre de mots des articles ressemble davantage à celle des autres disciplines. Cela suggère que le style des articles de *Géographie* perd de sa spécificité.

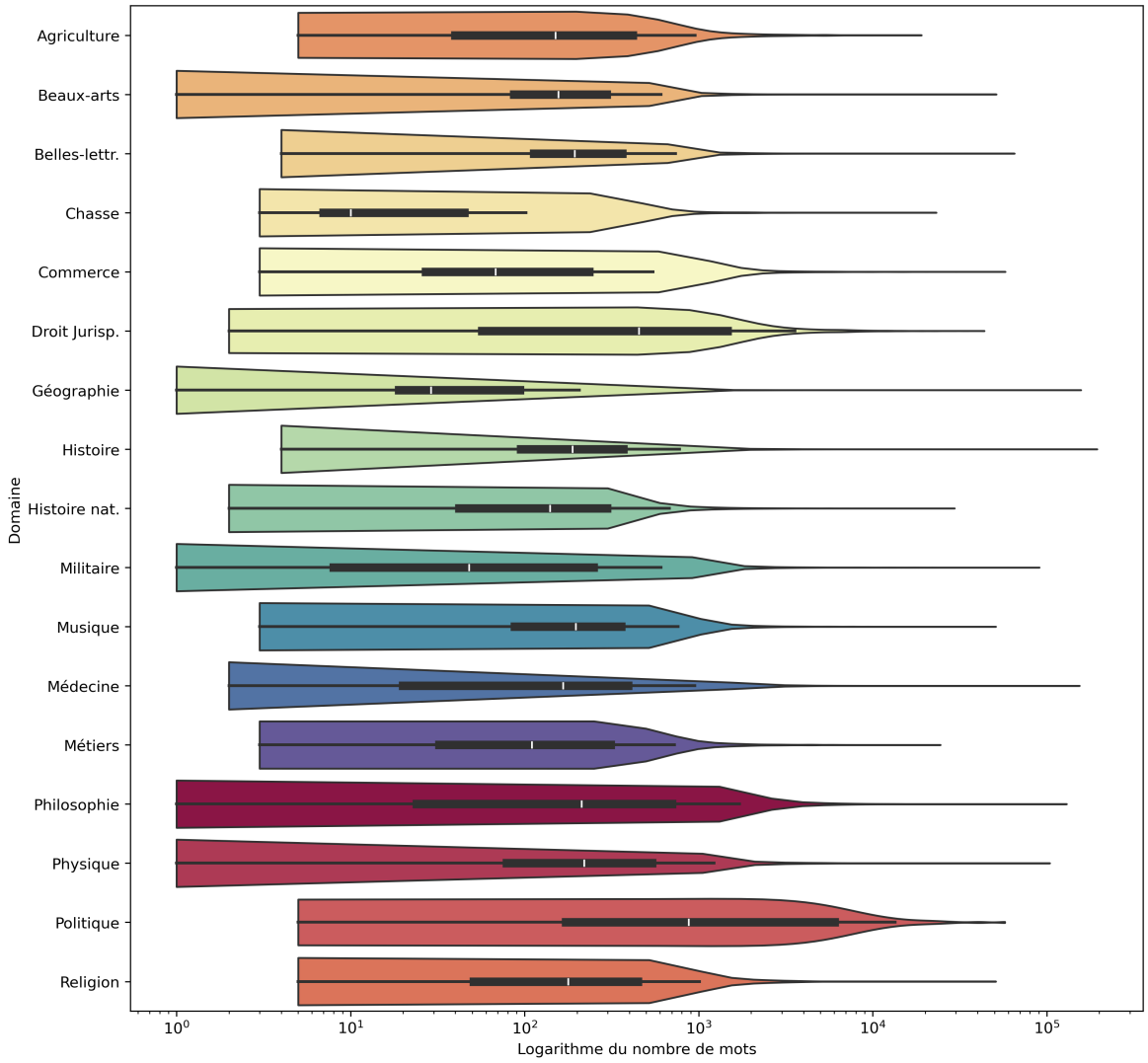


Figure 5.4 – Distribution des nombres de mots par article au sein des différents domaines de *LGE*

Longueur moyenne des mots

L'intérêt de mesurer aussi les tailles d'articles à l'échelle des signes comme décrit plus haut est de pouvoir calculer la longueur moyenne des mots de chaque domaine pour vérifier s'il existe un lien de dépendance entre celle-ci et les domaines. Il paraît plutôt vraisemblable que le rapport entre nombre de lettres et nombre de mots, c'est-à-dire en fait la longueur moyenne d'un mot, soit une constante propre à la langue et que la mesure ne puisse varier que très peu sur des échantillons assez grands. Le tableau 5.1 montre qu'au contraire ce rapport n'est pas constant et qu'il atteint son minimum pour la *Géographie* dans l'*EDdA*.

Domaine	dans l'*EDdA*	dans *LGE*
Agriculture	4.82	5.13
Beaux-arts	4.9	5.03
Belles-lettres	4.97	4.95
Chasse	4.89	4.99
Commerce	4.87	5.19
Droit Jurisprudence	5.07	5.16
Géographie	4.7	4.97
Histoire	5.0	5.03
Histoire naturelle	4.88	5.16
Militaire	4.84	5.1
Musique	4.75	4.95
Médecine	5.1	5.25
Métiers	4.77	5.11
Philosophie	4.96	5.14
Physique	5.0	5.16
Politique	5.13	5.3
Religion	4.98	5.13
Moyenne	4.92	5.1
Écart-type	0.12	0.1

Tableau 5.1 – Longueur moyenne des mots par domaine pour chacune des deux encyclopédies étudiées

Les articles de *Géographie* y sont donc particulièrement courts, non seulement en nombre de mots mais même, typographiquement, en place occupée sur les pages de l'*EDdA*. La présence dans ces articles d'abréviations notamment pour les points cardinaux («N.», «S.») et les coordonnées («Long.», «lat.») contribue sans doute à abaisser un peu la moyenne dans ce domaine mais ne suffit probablement pas à expliquer la brièveté particulière des articles de *Géographie* car des abréviations d'une lettre se rencontrent également dans les autres domaines. Outre les initiales de noms de personnes qui concernent toutes les lettres et peuvent apparaître dans de nombreux contextes, la séquence «S.» peut aussi représenter le mot «Saint» et se rencontre alors très fréquemment dans toute l'*EDdA*, notamment en *Histoire* et en *Religion* où elle a respectivement 1 861 et 1 409 occurrences (il y en a 1 877 en *Géographie*).

Dans *LGE*, la longueur moyenne des mots dans les articles augmente globalement pour atteindre 5.1 soit presque deux écarts-types de plus que la moyenne dans l'*EDdA*, ce qui montre un mouvement général de complexification du discours avec l'emploi plus fréquent de mots techniques plus longs quels que soient les domaines. Cette nouvelle valeur est même supérieure à la longueur moyenne des mots dans tous les domaines de l'*EDdA* sauf la *Politique*. La *Géographie* n'échappe pas à cet allongement du vocabulaire. Si elle reste plus d'un écart-type sous la moyenne de longueur des

mots dans *LGE* (4.97 contre 5.1) elle perd tout de même sa place de domaine aux mots les plus brefs, détrônée de peu par la *Musique* et les *Belles-lettres*. Mais surtout, la longueur moyenne de son vocabulaire, 4.97, est supérieure à celle dans l'ensemble de l'*EDdA*, tout domaine confondu.

Cela est d'autant plus remarquable que les abréviations se multiplient pourtant dans *LGE*. Parmi les 20 lemmes les plus spécifiques de la *Géographie* dans cette œuvre figurent cinq abréviations : «hab.» (habitant), «dép.» (département), «arr.» (arrondissement), «cant.» (canton) et «kil.» (kilomètre), toutes spécifiques «à saturation» (voir la section 2.5.2 p.39) et de surcroît exclusives à la *Géographie* (elles sont sous-spécifiques de tous les autres domaines). Par comparaison, les 40 lemmes les plus spécifiques de la *Géographie* dans l'*EDdA* ne contenaient que quatre abréviations de mots pleins (pour «Sud», «ancien(ne)?», «Longitude» et «latitude», les autres abréviations sont ou bien les désignants du domaine, ou des signatures d'auteurs). Pour que la longueur moyenne des mots dans les articles de *Géographie* parviennent à croître autant entre l'*EDdA* et *LGE*, il est donc raisonnable de faire l'hypothèse de l'apparition dans ce domaine d'un grand nombre de termes techniques beaucoup plus longs à même de contrebalancer cet effet, ce qui corrobore également l'hypothèse de sa disciplinarisation.

Annotations géo-sémantiques

Intuitivement, la Géographie en tant que science descriptive de l'espace terrestre renvoie à la notion de lieu et en particulier de lieu nommé. Le concept d'Entité Nommée (EN) paraît donc un angle d'approche tout à fait approprié pour étudier les articles encyclopédiques du corpus et on peut faire l'hypothèse que ces entités seront particulièrement présentes dans les articles de *Géographie*.

Ce concept apparu dans les années 1990 généralise la notion de noms propres à tout groupe de mots renvoyant à des objets nommés, uniques (Nadeau & Sekine, 2007, p. 3) (ou supposés tels dans le contexte). Central dans les tâches d'extraction d'information, il se définit en creux comme la matière précieuse à extraire qui justifie de s'intéresser à ces tâches. Il s'agit de les trouver dans des textes tels que des articles de presse ou des rapports d'activité ; de les classer suivant une typologie préétablie en relation avec l'usage que l'on souhaite faire du document ; de les rattacher sans ambiguïté à des entrées présentes dans des bases de connaissance et enfin de découvrir les relations qu'elles entretiennent (Ehrmann et al., 2016, p. 3350). Si elles semblent inclure dès le début les expressions numériques comme les dates ou les quantités pourvues d'une unité (Sekine et al., 2002), la définition et l'organisation de classes assez complètes fait l'objet de travaux conséquents qui aboutissent à la publication de jeux d'étiquettes complexes (Sekine & Nobata, 2004) jusqu'à des schémas d'annotation (Rosset et al., 2011). L'extraction des relations spatiales entre entités ou *Spatial Role Labeling*, une tâche introduite par Kordjamshidi et al. (2011) pour SemEval2012 est une opération difficile à cause de l'ambiguïté des prépositions et du vocabulaire dans de nombreux langages – illustrée par exemple pour l'anglais par Kordjamshidi et al. (2010, p. 3).

Mais les discours géographiques, au-delà de leur densité en EN, présentent aussi un intérêt dans la manière dont ils les mettent en relation ainsi que dans le vocabulaire qu'ils utilisent pour les catégoriser. On peut ainsi penser, dans le contexte précis de ce corpus d'étude à la différence significative entre des termes comme «paroisse» et «commune». C'est pourquoi une annotation géo-sémantique riche a été retenue plutôt qu'une annotation en entités nommées classiques pour examiner l'hypothèse formulée au début de cette sous-section. Le corpus a été annoté avec un modèle SpaCy spancat personnalisé [2] (Moncla et al., 2024a). Ce modèle a pour tâche de catégoriser les entités ou spans (ensemble de tokens) avec imbrications possibles selon une version simplifiée du schéma d'annotation décrit dans Moncla & Gaio (2015). Le schéma retenu pour l'annotation du jeu de données GeoEDdA (Moncla et al., 2024b) qui a servi pour l'entraînement et l'évaluation de ce modèle d'annotation comporte 12 étiquettes :

2. https://huggingface.co/GEODE/fr_spacy_custom_spancat_edda

— 3 types de noms propres : lieux (`NP-Spatial`), personnes (`NP-Person`) et autres (`NP-Misc`)
— 2 types de noms communs : lieux (`NC-Spatial`) et personnes (`NC-Person`)
— 1 pour les relations spatiales (`Relation`)
— 1 autre pour les coordonnées géographiques (`LatLong`)
— 1 marqueur pour les vedettes d'articles (`Head`)
— 1 pour leurs désignants (`Domain-mark`)

À ces neuf types atomiques s'ajoutent trois types d'Entités Nommées Étendues (ENE), reprenant le concept défini par Gaio & Moncla (2017):

— 1 de lieux (`ENE-Spatial`)
— 1 de personnes (`ENE-Person`)
— 1 non classées (`ENE-Misc`)

Ces entités permettent de capturer et de structurer le contexte local des entités nommées (NP-*). Elles sont construites à partir des NC-* et des NP-* (ou d'une autre ENE-*) et peuvent également inclure les Relations. De manière systématique le type de l'ENE (Spatial, Person, Misc) hérite du type du NC-* sur lequel elle est bâtie. Ainsi la figure 5.5 montre comment un `NC-Spatial` peut se combiner avec un `NP-Spatial` pour former une `ENE-Spatial`. En effet, une «ville d'Italie» est bien une ville, donc l'ENE dans son ensemble désigne une ville et est donc de type spatial. Dans ce cas le nom propre impliqué est aussi spatial donc il serait possible de croire que c'est le `NP` qui gouverne le type de l'`ENE` ou qu'un `NP` et un `NC` doivent s'accorder pour former une `ENE` mais un exemple comme «place de la Libération» permet de dissiper ces mauvaises intuitions : la «Libération» est un nom propre renvoyant à un événement (elle serait ici annotée `NP-Misc`), «place» est un `NC-Spatial` et l'ensemble est bien une `ENE-Spatial` (une fois encore de manière évidente parce que la «place de la Libération» est en particulier une place, donc bien un lieu). Ayant choisi de n'annoter que les noms communs de lieux et de personnes, les `ENE-Misc` sont dans ce schéma d'annotation dépourvues de `NC` (on pourrait ainsi annoter «le débarquement de Normandie» `ENE-Misc` car il s'agit d'un événement, «débarquement» ne serait pas annoté mais «Normandie» constitue évidemment un `NP-Spatial`). Enfin, les `ENE` peuvent à leur tour se combiner entre elles comme cela est également visible à la figure 5.5 (où elles atteignent quatre niveaux de profondeur).

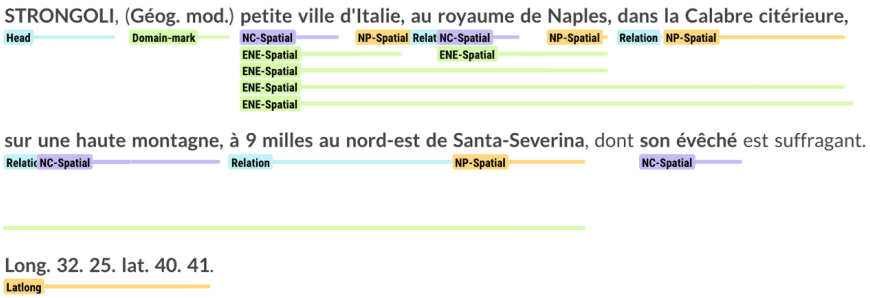


Figure 5.5 – Annotations de l'article Strongoli (L'Encyclopédie, T15, p.547)

La figure 5.6 représente les densités des différents types d'entités par domaine dans l'*EDdA*, c'est-à-dire pour chaque domaine de connaissance le nombre d'entités d'un des types décrits ci-dessus divisé par la taille de ce domaine. Les colonnes ne peuvent pas se sommer directement puisque les cellules de la matrice représentent des densités dans des domaines qui n'ont pas les mêmes nombres de mots. L'histogramme en haut de la figure représente donc leurs sommes pondérées, c'est-à-dire en fait les décomptes d'entités d'un type donné dans toute l'*EDdA*, divisés par le nombre de mots de cette encyclopédie.

Les sommes horizontales s'effectuent au contraire sans problème, mais il faut garder à l'esprit qu'elles représentent un nombre d'entités et pas un nombre de mots. D'abord, il n'y a pas de correspondance

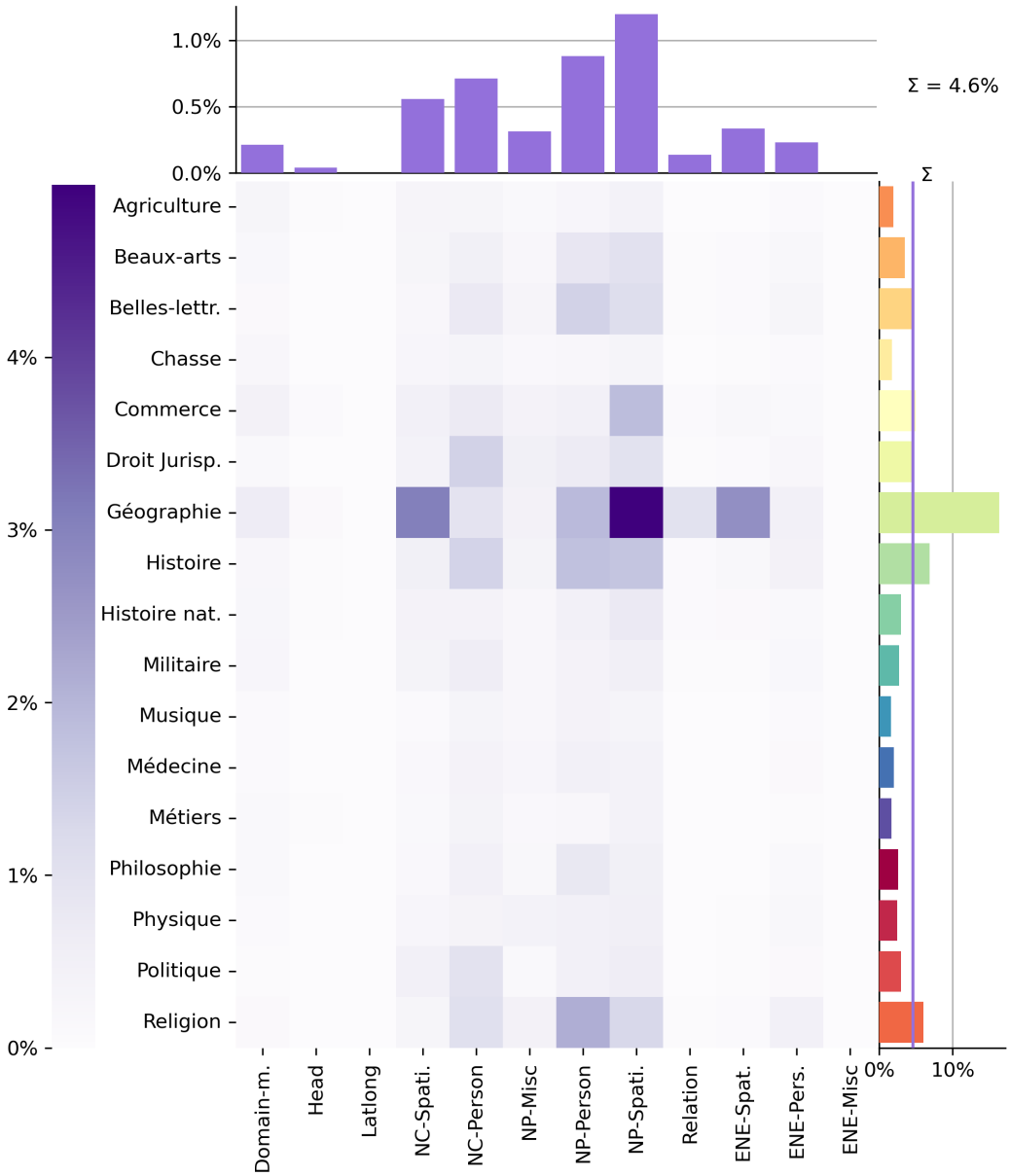


Figure 5.6 – Densité d'entités dans l'*EDdA* par domaine et par type

un-à-un entre les mots et les entités, même atomiques, ce qui est visible par exemple à la figure 5.5 dans «petite ville» : le syntagme nominal entier a été étiqueté `NC-Spatial` alors qu'il comporte deux mots. Ensuite, la nature imbriquée des ENE entraîne que les mêmes mots individuels peuvent participer à plusieurs entités. Ainsi, la séquence «royaume de Naples» de la même figure 5.5 contribue à trois entités : une pour «royaume», `NC-Spatial`, une pour «Naples», `NP-Spatial`, et une pour l'ensemble de l'expression, `ENE-Spatial`. Le fait que l'expression comporte également trois mots n'a pas de lien numérique avec ce fait, et sa participation à trois autres `ENE` dans lesquelles elle est imbriquée en profondeur fait que la densité d'entités rapportée à ce passage serait même supérieure à une entité par mot. Le compte obtenu pour chaque article rapporté à son nombre de mot, division nécessaire pour rendre compte fidèlement de la distribution des entités sans biais dûs à sa taille, ne représente donc pas une proportion de mots qui seraient des entités mais simplement une densité abstraite pour chacun et qui n'a pas réellement d'interprétation intuitive directement visible dans les articles.

Le trait le plus visible de la figure 5.6 est évidemment la prépondérance de la *Géographie* par opposition aux autres domaines dans la répartition des entités. Les plus présentes sont celles de lieu : `ENE-Spatial`, `NC-Spatial` et surtout `NP-Spatial` qui atteint environ 5%, ce qui est supérieur à la densité moyenne de tous les types d'entités confondus dans l'ensemble de l'*EDdA* (Σ, valant 4.6%). La ligne *Géographie* dans son ensemble totalise environ 16% d'entités et se dégage très nettement des autres domaines, puisque seules les classes *Histoire* et *Religion* arrivent à dépasser Σ de peu, avec une densité plus de deux fois moindre. Les ENE sont dans leur ensemble assez rares mais la catégorie `ENE-Spatial` atteint tout de même une densité assez élevée dans la classe *Géographie*, de l'ordre de celle de `NC-Spatial` ce qui est cohérent avec la remarque faite ci-dessus sur le fait qu'une `ENE-Spatial` était censée être bâtie autour d'un `NC-Spatial`. L'inverse n'est pas vrai, ce qui se voit au fait que la colonne `NC-Spatial` est légèrement plus foncée sur la ligne *Géographie* que celle de `ENE-Spatial`. Cette différence s'accentue pour tous les autres domaines : là où le rapport `NC-Spatial` / `ENE-Spatial` (calculé à partir des données numériques de la matrice, difficile à évaluer à partir des seules couleurs de la figure) atteint son minimum, 1.1, en *Géographie*, il monte ensuite à 2.5 en *Commerce*, deuxième domaine pour lequel ce rapport est le plus bas et s'élève même à plus de 10 en *Agriculture*. Bien qu'encore atténuée, une répartition similaire concerne les `Relations`, présentes quasi-exclusivement en *Géographie*. Si des éléments spatiaux sont donc présents à l'intérieur de syntagmes nominaux dans les autres domaines, la *Géographie* se caractérise dans l'*EDdA* par leur mise en relation au sein des composés structurés que constituent les ENE.

Les autres types d'entités particulières, celles de personnes, présentent une distribution beaucoup plus homogène entre les domaines. Les `NC-Person` sont ainsi bien plus uniformément réparties que les `NC-Spatial`, et les `NP-Person` atteignent une densité proche de 2%, son maximum, dans trois domaines. Il est à noter que ce dernier type d'EN est bien représenté en *Géographie*, où il très légèrement plus fréquent qu'en *Histoire* alors que les EN de type *NP-Spatial* étaient bien moins fréquents en *Histoire* qu'en *Géographie*. Les entités de personnes n'apparaissent donc pas comme caractéristiques d'un domaine en particulier mais s'observent avec des combinaisons de densités variées entre les `NC-Person`, `NP-Person` et `ENE-Person` en fonction des domaines.

La figure 5.7 qui montre les mêmes statistiques mais pour *LGE* offre un aspect tout à fait différent. La *Géographie* y conserve la place de domaine le plus dense en entités mais son score est bien plus proche de ceux de l'*Histoire*, des *Beaux-arts* et des *Belles-lettres* les trois autres domaines à dépasser la densité moyenne à l'échelle de toute cette seconde encyclopédie. Cette proximité se retrouve dans les lignes de la matrice qui sont bien plus uniformes : celle de la *Géographie* ne ressort pas nettement des autres et présente des similitudes fortes avec celles des trois autres domaines les plus denses. La seule particularité réellement identifiable reste au niveau des types utilisés pour les lieux : `ENE-Spatial` et `NC-Spatial`, bien que très atténués par rapport à l'*EDdA* se dégagent de la ligne *Géographie* et la distinguent des autres semblables. La *Géographie* conserve la plus haute densité de `NP-Spatial`

mais est considérablement rattrapée par la plupart des domaines. Très schématiquement, là où la figure 5.6 pour l'*EDdA* était marquée par la ligne *Géographie*, la figure 5.7 de *LGE* se présente plus verticalement et se structure le long de la colonne `NP-Spatial`. Le résultat sur l'histogramme du haut de la figure est net : cette classe creuse l'écart et s'élève nettement au-dessus des autres à près de 3% de densité dans la totalité de *LGE*, soit environ deux fois et demie plus que dans l'*EDdA*. La densité totale d'entités dans cette encyclopédie s'est beaucoup accrue, atteignant 7.5% soit plus d'une fois et demie celle dans les pages de l'*EDdA*.

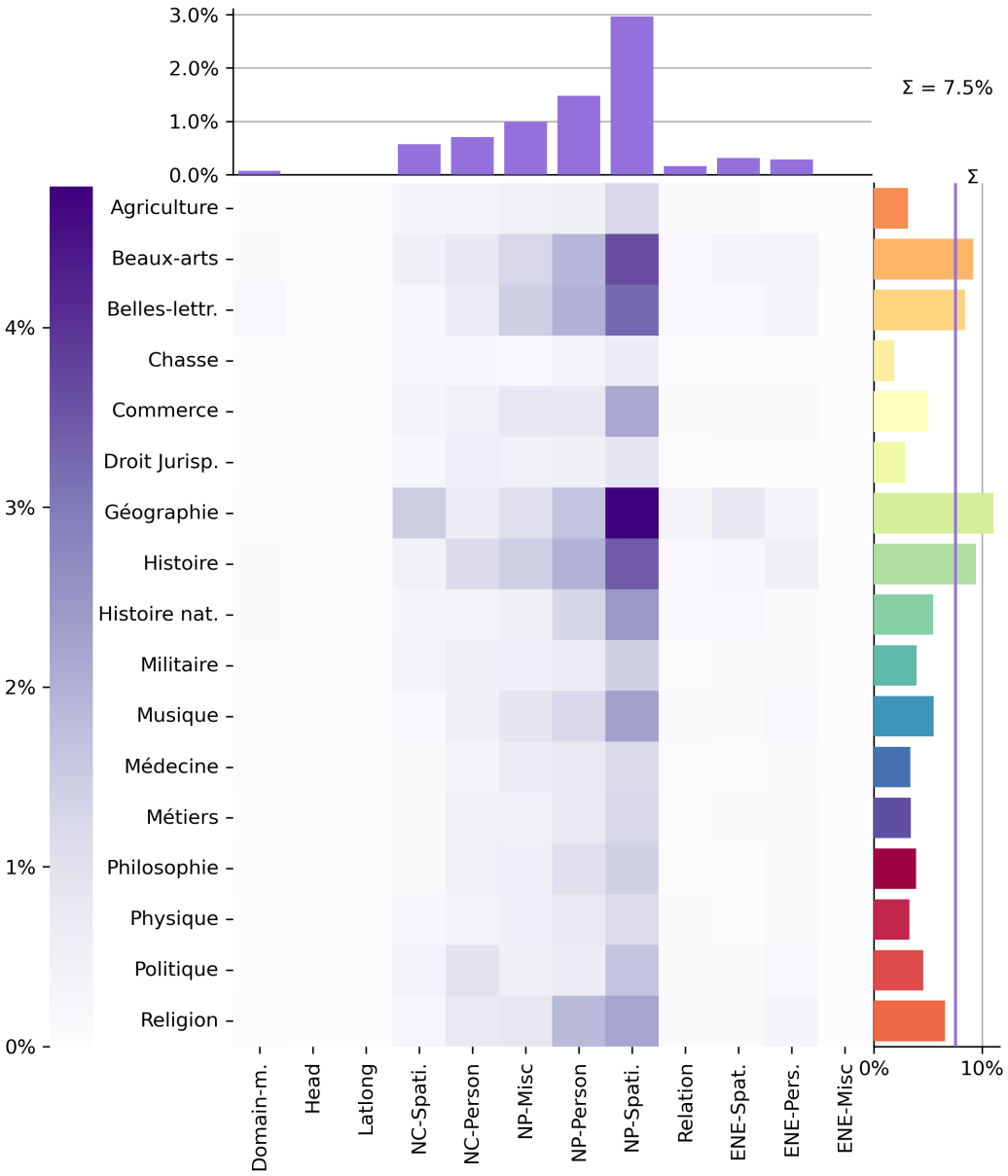


Figure 5.7 – Densité des ENE dans *LGE* par domaine et par type d'ENE

Il y a donc dans ces encyclopédies deux mouvements simultanés et contraires au niveau des entités entre le XVIII^{ème} et le XIX^{ème} siècle. D'une part les entités atomiques se font plus fréquentes, surtout les noms propres, et se répandent dans de nombreux autres domaines que la *Géographie*, y compris les `NP-Spatial` détectés autant dans des domaines où ils peuvent être attendus (le *Commerce* mais aussi l'*Histoire naturelle* quand on tient compte de ce que la discipline doit aux grandes explorations du siècle précédent) que dans d'autres où ils le seraient moins comme les *Beaux-arts*, la *Musique* ou

même la *Religion*. Concernant les entités de lieu, le mouvement se limite aux seuls noms propres. Les `NC-Spatial` tout comme les `ENE-Spatial`, bien moins fréquents, restent davantage caractéristiques de la *Géographie*, qui se distingue par contraste des autres domaines en conservant l'usage de ces noms communs qui deviennent un vocabulaire disciplinaire. Malgré cette spécialisation, le deuxième mouvement à l'œuvre voit la *Géographie* se normaliser et ressembler davantage aux autres domaines pour les entités détectées. La décroissance de sa densité en entités est probablement surestimée au moins en partie à cause des abréviations évoquées précédemment dans cette section p.141, très présentes dans les articles de *Géographie* de *LGE* et assez mal repérées par le modèle. Mais au-delà de ce seul indicateur, la différence de densités de `NP-Spatial` entre la *Géographie* et les autres domaines s'estompe également, la rapprochant des profils des autres disciplines. Les `ENE-Spatial` sont moins concentrées, elles forment des arbres moins profonds (c'est-à-dire moins d'`ENE-x` imbriquées comme pouvaient l'être celles de la figure 5.5 imbriquées sur quatre niveaux). L'`ENE-Spatial` la plus profonde trouvée dans *LGE* a ainsi une profondeur de quatre là où la profondeur maximale dans l'*EDdA* s'élevait à sept. La très grande majorité des ENE trouvées sont très plates et la profondeur moyenne dans ces encyclopédies est de l'ordre de grandeur de 10^{-2} mais celle hors de la classe *Géographie* dans l'*EDdA* est de 8×10^{-3} alors qu'elle s'élève à 1×10^{-1} en *Géographie* ; dans *LGE*, la profondeur moyenne hors *Géographie* est de 4×10^{-3} contre 9×10^{-3}. Les articles de *Géographie* dans *LGE* sont donc non seulement moins denses en entités que ceux de l'*EDdA* mais ils font l'objet d'une mutation plus profonde : les concentrations de noms propres (dont surtout des toponymes) s'effacent pour laisser se développer un discours plus structuré et plus similaire à celui des autres disciplines : la Géographie ne sert plus tant à nommer et à localiser des référents qu'à donner des informations sur eux.

5.1.2 Des articles qui changent de domaine

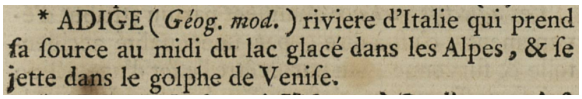
Les considérations précédentes fournissent des mesures «en gros» et se focalisent sur les groupes d'articles rassemblés par domaines en essayant de mesurer leurs variations de taille entre l'*EDdA* et *LGE*. La représentation sous-jacente correspond à des «essaims» d'articles qui grossiraient et se réduiraient au fil des ajouts ou des suppressions d'articles dans un domaine donné ainsi que des variations des volumes de textes consacrés aux différents articles de ce domaine. Cette sous-section se propose au contraire de suivre une approche relativement orthogonale en prenant un ensemble fixe d'entrées existant aux deux époques pour observer comment celles-là changent.

Constitution d'un sous-corpus

La première difficulté consiste donc à établir des correspondances entre les articles des deux encyclopédies, c'est-à-dire à définir une notion d'équivalence entre deux articles pourtant issus de deux encyclopédies distinctes et dont les différences seules nous intéressent. Étant donné le caractère qualitatif plus que quantitatif de cette étude (l'ensemble d'entrées suivies en détail n'a pas besoin d'être aussi grand que possible puisque de toute façon par construction il ne représentera pas la totalité des articles), l'algorithme le plus naïf est utilisé en constituant des paires d'articles de la manière suivante : une paire est constituée d'un article de l'*EDdA* et d'un article de *LGE* ayant la même vedette qui doit de plus se trouver être unique dans chacune des encyclopédies prise séparément.

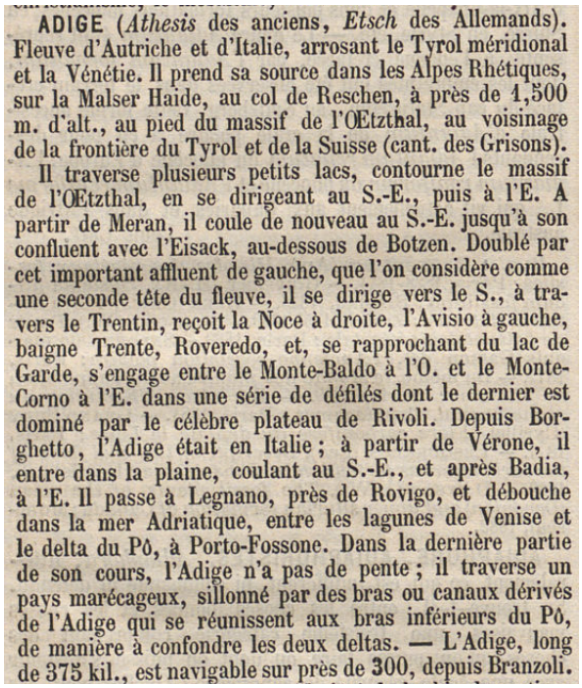
Les vedettes sont comparées en utilisant l'égalité usuelle des chaînes de caractères mais sans tenir compte de la casse étant donné que la typographie est un peu irrégulière dans l'*EDdA* et que les conventions typographiques varient de toute façon légèrement entre les deux encyclopédies. Cette règle relativement stricte n'est pourtant pas suffisante pour garantir que les paires sont toutes bien fondées : au lieu qu'elles renvoient bien au «même» objet (concept, lieu, etc.) il se pourrait qu'une entrée vienne à disparaître de la première encyclopédie pour être remplacée dans la suivante par une autre avec la même vedette mais sans lien sémantique avec l'entrée initiale. Si la vedette commune à une paire d'entrées est un nom propre, il s'agira de vérifier qu'elles décrivent bien la même entité ;

pour les vedettes relevant des autres catégories morphosyntaxiques, comme le proposent Lehmann & Martin-Berthet (2018, chap. 5, §1.3) l'étymologie complétée de critères sémantiques lorsqu'une évolution forte a éloigné les termes permettront de distinguer les polysémies, intéressantes à étudier, des simples homonymies et ce même si le sens utilisé dans une des encyclopédies est absent dans l'autre. Toute la difficulté de constitution de ce sous-corpus, nommé «Parallèle» et noté $\mathcal{P}$ dans ce qui suit, réside dans ce problème de bon appariement – distinguer les *vraies* paires des *fausses*, sources de bruit dans le corpus. Par construction, $\mathcal{P}$ comprend un nombre pair d'articles, autant de l'*EDdA* que de *LGE* puisqu'il est obtenu en sélectionnant des couples d'articles. Les figures 5.8a et 5.8b présentent un exemple de paire issue du sous-corpus Parallèle. Cette paire est une vraie paire car les articles, bien que très différents ne serait-ce que par leur longueur, renvoient bien au même cours d'eau.



* ADIGE (*Géog. mod.*) riviere d'Italie qui prend ſa ſource au midi du lac glacé dans les Alpes, & ſe jette dans le golphe de Veniſe.

(a) Article ADIGE dans l'*EDdA*, T1, p.138



ADIGE (*Athesis* des anciens, *Etsch* des Allemands). Fleuve d'Autriche et d'Italie, arrosant le Tyrol méridional et la Vénétie. Il prend sa source dans les Alpes Rhétiques, sur la Malser Haide, au col de Reschen, à près de 1,500 m. d'alt., au pied du massif de l'Œtzthal, au voisinage de la frontière du Tyrol et de la Suisse (cant. des Grisons).

Il traverse plusieurs petits lacs, contourne le massif de l'Œtzthal, en se dirigeant au S.-E., puis à l'E. A partir de Meran, il coule de nouveau au S.-E. jusqu'à son confluent avec l'Eisack, au-dessous de Botzen. Doublé par cet important affluent de gauche, que l'on considère comme une seconde tête du fleuve, il se dirige vers le S., à travers le Trentin, reçoit la Noce à droite, l'Avisio à gauche, baigne Trente, Roveredo, et, se rapprochant du lac de Garde, s'engage entre le Monte-Baldo à l'O. et le Monte-Corno à l'E. dans une série de défilés dont le dernier est dominé par le célèbre plateau de Rivoli. Depuis Borghetto, l'Adige était en Italie ; à partir de Vérone, il entre dans la plaine, coulant au S.-E., et après Badia, à l'E. Il passe à Legnano, près de Rovigo, et débouche dans la mer Adriatique, entre les lagunes de Venise et le delta du Pô, à Porto-Fossone. Dans la dernière partie de son cours, l'Adige n'a pas de pente ; il traverse un pays marécageux, sillonné par des bras ou canaux dérivés de l'Adige qui se réunissent aux bras inférieurs du Pô, de manière à confondre les deux deltas. — L'Adige, long de 375 kil., est navigable sur près de 300, depuis Branzoli.

(b) Article ADIGE dans *LGE*, T1, p.572

Figure 5.8 – Les articles sur le fleuve ADIGE, un exemple de paire issue du sous-corpus Parallèle

L'approche ci-dessus basée sur la vedette plutôt que le contenu des articles est en un sens trop restrictive et peut être critiquée pour plusieurs raisons. D'abord, des objets dont le nom aurait changé entre les XVIII^{ème} et XIX^{ème} siècles seront exclus du sous-corpus par la condition choisie alors que le renommage d'entrées pourrait aussi être un aspect intéressant à étudier. Ce cas de figure se rencontre par exemple avec la capitale impériale japonaise à l'ère Tokugawa, la ville dont le nom moderne est Kyōtō en trancription Hepburn. On trouve la ville dans l'*EDdA* à l'entrée «MÉACO ou MIACO» (L'Encyclopédie, T10, p.218) – correspondant au japonais みやこ, soit «miyako» en transcription Hepburn, un ancien nom de la ville signifiant simplement «capitale» – alors que dans *LGE* publiée après la restauration impériale de 1868 (l'année de début de publication de l'œuvre, 1885, correspond à l'an 18 de l'ère Meiji) la même ville est décrite par l'entrée «KIYOTO ou MIAKO» (La Grande Encyclopédie, T21, p.545). Outre la prise en compte nécessaire des graphies multiples (ce qui en théorie ne paraît pas revêtir un problème conceptuel particulier mais en pratique peut s'avérer une tâche infinie nécessitant sans cesse de raffiner la grammaire employée pour analyser les vedettes au fur et à mesure qu'elle bute sur de nouveaux cas particuliers), la variation orthographique aurait caché cette paire pourtant valide avec le parti pris de ne pas comparer le contenu des articles. Ensuite, une comparaison des contenus des articles permettrait de maintenir le bruit dans les paires à un niveau acceptable tout en utilisant une comparaison plus souple sur les vedettes, par exemple une

égalité basée sur un seuil pour une distance usuelle sur les chaînes de caractères comme la distance de Levenshtein ou le coefficient de Sørensen-Dice, ce qui aurait pu rattraper l'écart précédent entre «MIAKO» et «MIACO», illustration que l'orthographe change au cours du temps en particulier pour les transcriptions de noms de lieux étrangers. Enfin, il peut exister des groupes d'entrées homonymes dans une des encyclopédies au sein desquelles une entrée pourrait être mise en correspondance avec une entrée de l'autre encyclopédie, appartenant possiblement elle aussi à un groupe homonyme si la vedette était fortement polysémique aux deux époques. L'heuristique choisie les écarte plutôt que d'essayer de trouver comment les mettre en correspondance.

Ce choix se défend pourtant déjà en ce qu'il s'agit ici de favoriser la simplicité par rapport à l'exhaustivité. Il apparaît peu pertinent de rechercher à considérer précisément l'ensemble de tous les objets décrits à la fois dans l'*EDdA* et *LGE* comme s'il avait une qualité particulière qui ne tiendrait pas aux hasards éditoriaux des deux entreprises. En pratique, cet ensemble peut être remplacé par n'importe lequel de ses sous-ensembles pourvu qu'il soit assez large pour être représentatif. Mais surtout, imposer des restrictions trop sévères sur le contenu des articles empêcherait d'observer des changements majeurs dans la manière dont un même objet serait traité entre l'*EDdA* et *LGE*, par exemple un territoire qui changerait de pays et ne serait donc plus décrit en rapport aux mêmes entités administratives comme ce fut le cas de l'Alsace entre le milieu du XVIII^{ème} siècle et la fin du XIX^{ème} siècle, ou une évolution scientifique et technique qui ferait qu'on ne parle plus du tout d'un même sujet avec les mêmes termes. L'article SAVONE est particulièrement intéressant pour cette raison. D'un descriptif des ordres religieux de la ville, de son commerce moribond et de ses liens avec d'autres villes, italiennes exclusivement, qui est suivi d'une biographie du pape Jules II dans l'*EDdA* (L'Encyclopédie, T14, p.722), la ville paraît transformée dans *LGE* (La Grande Encyclopédie, T29, p.624), l'article bien plus bref insiste sur la force de son commerce notamment de fer et de pétrole, sur son aciérie, et la met en relation avec Marseille. Une comparaison basée sur le contenu aurait pu juger les deux articles trop différents.

Si la pertinence de la règle basée sur les vedettes a été justifiée, il n'en demeure pas moins que les objections levées précédemment démontrent l'intérêt d'une approche basée sur une mesure de similarités des contenus des articles. Il s'agit donc de trouver une manière de les combiner pour améliorer la précision de la règle de sélection initiale. Le problème majeur de l'approche basée sur le contenu des articles réside dans la part d'arbitraire qui demeure dans le choix d'un seuil au-delà duquel deux articles sont considérés comme différents et qui, sauf à trouver une métrique particulièrement bonne et dont l'existence n'est pas prouvée, excluera nécessairement des vraies paires tout en incluant des fausses. Pour cette raison, le sous-corpus $\mathcal{P}$ est donc constitué en appairant d'abord automatiquement les articles avec la règle simple basée sur les vedettes puis filtré de manière semi-automatique en utilisant une métrique basée sur le contenu seulement pour guider la validation, purement manuelle, des paires. Ce processus de sélection se base sur les données, sans à priori : le seuil est fixé en fonction des vraies et fausses paires trouvées parmi les articles.

L'application de la règle sur les vedettes génère une liste de 11 746 articles, c'est-à-dire 5 873 paires. Cette liste contient bien sûr des fausses paires, comme celle pour la vedette ABACO qui est donnée pour un synonyme désuet d'«arithmétique» dans l'*EDdA* (L'Encyclopédie, T1, p.6) mais désigne un architecte italien dans *LGE* (La Grande Encyclopédie, T1, p.11) – il y a en réalité trois entrées homonymes dans l'œuvre mais des problèmes de segmentation des articles de cette œuvre en ont masqué deux. Le nombre de paires trouvées est relativement faible au regard des 74 190 articles que comprend au total l'*EDdA* (moins de 8%) mais reste tout de même un peu trop élevé pour les évaluer manuellement. La stratégie composite définie au paragraphe précédent prend alors tout son intérêt en utilisant les similarités de contenus des articles pour les ordonner et accélérer la sélection. En effet, en appliquant une métrique de similarité cosinus basée sur le plongement vectoriel des textes avec *SentenceBERT* (Reimers & Gurevych, 2019), on constate une forte dépendance de la véracité des paires au score qu'elles obtiennent. Ainsi les cinq candidates les mieux notées selon cette mé-

trique traitent toutes des mêmes objets : CONTEMPLATION d'un concept théologique et mystique, ACUTANGLE d'un cas particulier de triangle, SYNE d'un mois éthiopien, TRIHEMITON d'un intervalle en musique grecque jusqu'à CONGRUE qui n'est qu'un renvoi dans les deux encyclopédies mais vers le même terme de Jurisprudence. Au contraire parmi les cinq les moins bien notées se trouvent PRESTON, une ville dans l'*EDdA* contre un renvoi biographique dans *LGE*, SERY respectivement un rongeur et une commune ardennaise et MARTIALE qui définit une fleur au XVIII^{ème} et une loi au XIX^{ème} (malgré leur étymon commun – le dieu romain Mars – il n'y a plus de lien sémantique entre les deux et la polysémie est devenue homonymie). Les deux autres sont une paire valide pour CAP-BRETON (mais au contenu peu comparable puisqu'il s'agit d'un simple renvoi dans l'*EDdA* mais d'une entrée géographique complète dans *LGE*) et l'autre, FILLES est due à une erreur de segmentation dans un article sur l'école dans *LGE*.

Ces exemples montrent que cette simple étude pourrait être énormément approfondie en travaillant sur la comparaison des vedettes (en incluant notamment une composante phonémique à la métrique utilisée pour valuer différemment l'écart entre MIACO et MIAKO et celui entre MIACO et MIANO) et en résolvant les renvois pour effectuer les comparaisons de similarité sur le contenu des articles complets. Mais dans le cadre de ces travaux où elle n'a pour but que de mettre en évidence des tendances d'évolution entre disciplines, il ne reste qu'à fixer le seuil de similarité pour retenir les paires et former $\mathcal{P}$. Et puisque les deux extrémités du classement ne sont pas violemment en contradiction avec cette hypothèse, on va supposer que le lien de dépendance entre qualité et score est relativement monotone (plus le score est élevé, plus la paire a de chance d'être vraie) et procéder par dichotomie sur la liste des paires candidates classées par ordre croissant de similarité.

Pour éviter d'accorder trop d'importance à une éventuelle paire trop bien ou trop mal classée, on procède en gardant une fenêtre pour la comparaison. En effet non seulement les cinq premières les mieux notées sont valides mais c'est également le cas des cinq suivantes. Une fenêtre de dix entrées est donc utilisée pour trouver le point au-delà duquel plus d'une paire sur dix est fausse, dans l'espoir d'atteindre ainsi une qualité d'au moins 90% de vraies paires dans $\mathcal{P}$. Puisque la plage de recherche initiale est de 5 873 paires, en considérant des fenêtres de dix éléments le processus s'arrête dans le pire des cas au bout de $\lceil log_2(5873/10) \rceil = 10$ étapes. L'évaluation d'une fenêtre peut s'arrêter dès que deux fausses paires ont été trouvées, mais dans le cas le plus défavorable (trouver exclusivement des vraies paires), il faudrait donc au pire 100 comparaisons. Le processus représente un compromis entre fiabilité statistique, qui requiert d'augmenter la taille de la fenêtre et demande davantage de comparaisons et rapidité qui est optimale quand on regarde seulement un article à la fois («fenêtre» de taille un) comme l'illustre la figure 5.9 (l'effet de «décrochage» est bien sûr dû à la fonction partie entière supérieure dans la formule précédente).

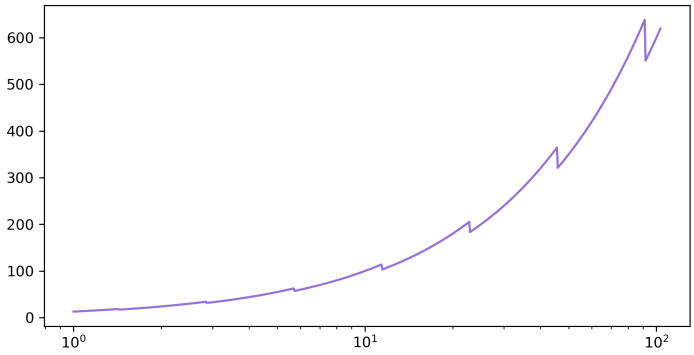


Figure 5.9 – Nombre total de comparaisons requis dans le pire des cas en fonction de la taille de la fenêtre utilisée

Au milieu des paires candidates classées par similarité, entre les rangs 2 931 et 2 940 on trouve une

unique fausse paire, celle pour SPIEGELBERG – une contrée allemande (L'Encyclopédie, T15, p.461) et un gynécologue (La Grande Encyclopédie, T30, p.388). Toutes les autres sont au moins en lien, l'acception moderne contenant l'acception ancienne comme c'est le cas pour BARRAGE où *LGE* contient une sous-section sur son acception commerciale dans l'ancien régime (La Grande Encyclopédie, T5, p.469) qui est la seule définie dans l'*EDdA* (L'Encyclopédie, T2, p.90) qui a été écrite à cette époque. Le taux de similarité rapporté avec la mesure choisie est supérieur à 56.9% (valeur pour la 2 931$^{\text{ème}}$ paire). On réitère donc le procédé entre les rangs 1 461 et 1 470 qui ont une similarité cosinus d'au moins 42% et contiennent au moins deux fausses paires. Puisqu'ils contiennent trop de bruit, on remonte donc le seuil en allant entre les rangs 2 194 et 2 203 (similarité > 50.3%) qui eux ne contiennent qu'une fausse paire et ainsi de suite. Le processus se poursuit jusqu'à ce que deux fenêtres se rejoignent ce qui arrive comme prévu à la dixième étape. On passe alors toutes les fausses paires jusqu'à la première vraie (pour ne pas ajouter délibérément des paires que l'on sait fausses à $\mathcal{P}$), celle de rang 2 167 qui avait obtenu un score de similarité de 50.0%. On les retient toutes depuis celle-ci jusqu'à la fin de la liste, ce qui donne $5872 - 2167 + 1 = 3706$ paires d'articles. Il est à noter qu'au cours du processus, une autre série de 10 paires sans aucune homonymie a été trouvée entre les rangs 2 171 et 2 180. Ce contre-exemple à l'hypothèse de monotonie de la qualité en fonction du score est toutefois rassurant sur la qualité des rangs sélectionnés et la pertinence de la limite.

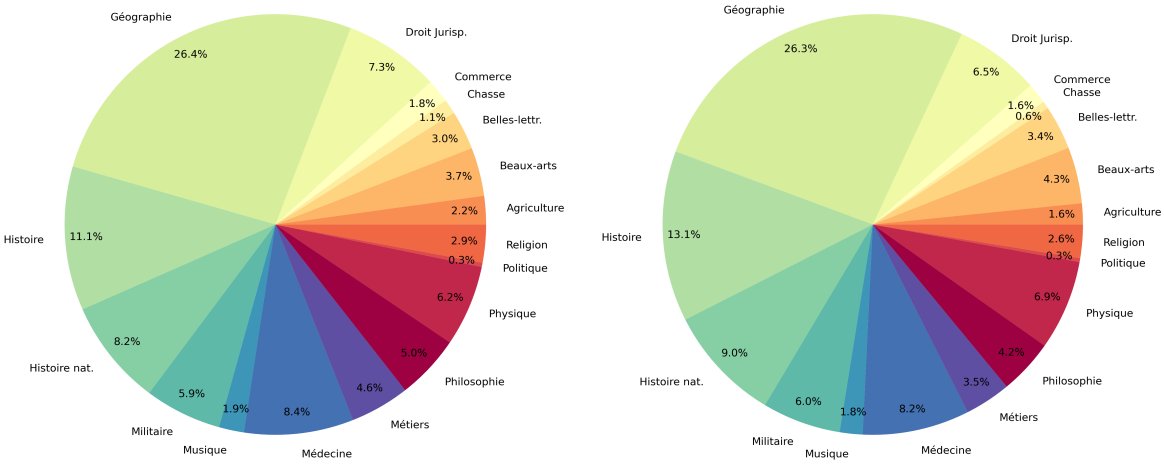
Analyses du sous-corpus Parallèle

Au cours du processus précédent de validation de fenêtres, quatre fenêtres positives ont été mises en évidence : deux sans aucune fausse paire, et deux avec une unique fausse paire. Il aurait bien évidemment été très tentant d'utiliser ces échantillons déjà validés pour estimer la qualité du sous-corpus sélectionné. C'est toutefois impossible pour des raisons méthodologiques car ces quatre fenêtres ne constituent pas un échantillon aléatoire d'éléments indépendants. En effet, la nature du processus de sélection même fait que les fenêtres ont été considérées en fonction des résultats de l'évaluation des autres déjà validées. Il est donc nécessaire de prélever un nouvel échantillon pour vérifier si la qualité a une chance d'être proche des 90% espérés initialement. Comme cette fois les n paires prélevées sont indépendantes, et puisqu'on évalue sur chacune une propriété booléenne (elles peuvent être soit vraies soit fausses), elles constituent elles aussi des variables aléatoires de Bernoulli, comme c'était le cas pour les articles de *LGE* étiquetés en domaine par le modèle *BERT* et dont il s'agissait de contrôler la qualité à la section 4.3.2 page 129. Le même cadre théorique s'applique donc et un raisonnement tout à fait semblable peut se développer. Le Théorème Central Limite peut à nouveau s'appliquer à condition que la taille de l'échantillon reste suffisamment faible devant celle de l'ensemble de la population : le rapport de 100 considéré comme suffisant impose donc que l'échantillon comporte moins de $3706/100 = 37.06$ paires. À cette condition, il est à nouveau possible d'utiliser la loi normale pour modéliser le comportement de ces variables. Cette contrainte impose une taille assez réduite et, pour maximiser la force de preuve de l'expérience il convient de ne pas examiner moins de paires.

Un échantillon de 37 paires a donc été extrait aléatoirement puis vérifié et ne comportait que deux fausses paires. La première est SPIEGELBERG discutée plus haut (qui s'est retrouvée par hasard dans l'échantillon). La seconde concerne les entrées PLUVIERS, toutes deux en *Géographie* mais ne désignant pas la même ville : il s'agit dans l'*EDdA* d'une ville de la Beauce connue maintenant sous le nom de Pithiviers (L'Encyclopédie, T12, p.805), mais de l'ancien nom d'une commune de la Dordogne désormais appelée Piégut pour *LGE* (La Grande Encyclopédie, T26, p.1146). De même que lors de l'échantillonnage précédent de la section 4.3.2, on utilise la moyenne empirique pour estimer la qualité q dans $\mathcal{P}$ valant $m = \frac{35}{37} \approx 0.946$. L'application numérique 5.1 de la borne inférieure de la formule 4.5 p.130 permet d'affirmer avec moins de 5% de risque d'erreur que la qualité réelle dans le sous-corpus Parallèle est d'au moins

$$m - z_{97.5\%} \times \sqrt{\frac{m \times (1-m)}{n}} = 87.3\% \tag{5.1}$$

Maintenant que la qualité de ce sous-corpus est établie, il est possible de se poser la question de sa représentativité. La figure 5.10 montre que la répartition du nombre d'articles par classe au sein du sous-corpus Parallèle est plutôt semblable à celle trouvée dans toute l'*EDdA* (voir la figure 4.20b p.127), ce qui est cohérent avec le fait que cette encyclopédie, la plus réduite en nombre d'articles joue en quelque sorte le rôle de «facteur limitant» dans cette réaction d'appariement. Les principales différences notables sont l'augmentation des parts de la *Géographie* et de la *Médecine* aux dépens surtout du *Droit Jurisprudence* et des *Métiers*. Ces changements sont tous conformes à la variation qui existe entre l'*EDdA* et *LGE* prises dans leur totalité ce qui peut s'interpréter comme une influence de cette dernière au cours du processus de sélection.



(a) Dans la moitié issue de l'*EDdA*

(b) Dans la moitié issue de *LGE*

Figure 5.10 – Répartition des nombres d'articles par domaines au sein du sous-corpus Parallèle

Au-delà leur forte ressemblance, il est à noter que les proportions diffèrent entre les deux moitiés du corpus, ce qui montre immédiatement qu'un même objet peut intéresser des disciplines différentes suivant l'époque à laquelle on lui consacre un article. Et le phénomène n'est pas du tout marginal puisque sur les 3 706 paires, 927 – soit tout juste plus d'un quart d'entre elles (25.01%) – ont des domaines différents, ce qui permet de rejeter l'explication qui attribuerait ces différences aux erreurs d'appariement (il y a moins de 5% de risque qu'il y en ait seulement 371[3] dans p, et la probabilité diminue fortement avec l'augmentation du nombre donc il serait d'autant plus improbable que ces erreurs expliquent ces différences presque deux fois et demie plus nombreuses).

La figure 5.11 présente ces changements : pour une paire donnée, le domaine en ordonnée est celui de l'article dans l'*EDdA*, alors qu'en abscisse est représenté celui dans *LGE*. Le premier élément visible est une diagonale relativement forte – mais pas homogène – qui matérialise les 74.99% de paires dont le domaine n'a pas changé. Sur cette diagonale, les points particulièrement colorés sont ceux des domaines *Géographie*, *Histoire-naturelle* et dans une moindre mesure *Beaux-arts*, *Droit Jurisprudence*, *Militaire*, *Musique* et *Médecine*. La colonne et la ligne du domaine *Géographie* sont remarquablement peu visibles hors de leur intersection, ce qui montre que ce domaine est celui dont statistiquement

3. soit 10% des articles de p, ce qui serait équivalent à dire que la qualité – mesurée empiriquement à 96% – serait en fait inférieure à 90%, ce qui d'après définition de $z_{97.5\%}$ a moins de 5% de chance d'être vrai

les articles ont le moins changé de catégorie entre les deux époques. Par contraste la colonne *Histoire*, très ténue mais nettement visible hors de la diagonale montre qu'un nombre conséquent d'articles d'autres domaines sont passés vers la catégorie *Histoire* dans *LGE* ce qui se manifestait dans la figure 5.10b par la nette augmentation de la proportion d'*Histoire* par rapport à la même figure pour l'*EDDA* (figure 5.10a).

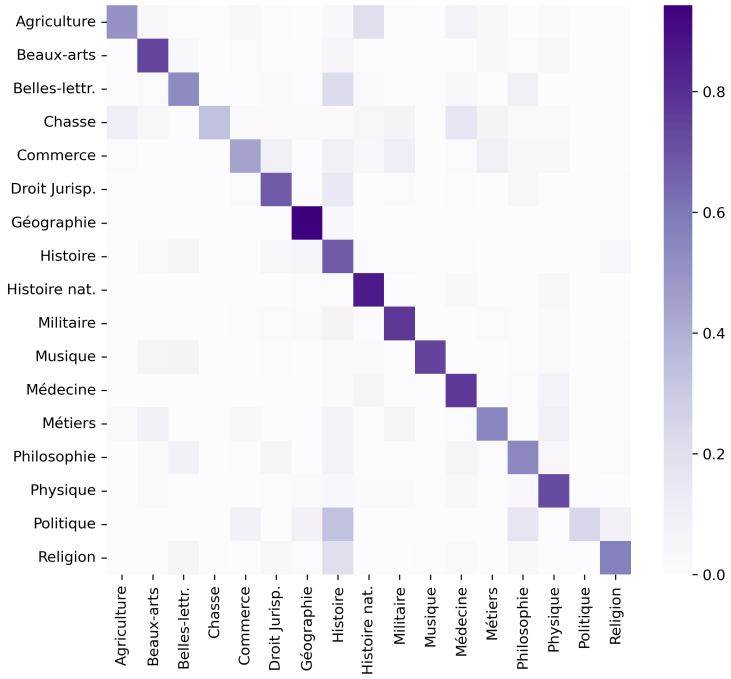
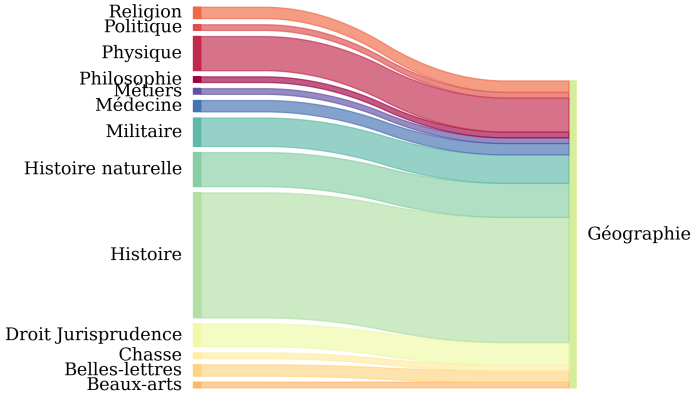


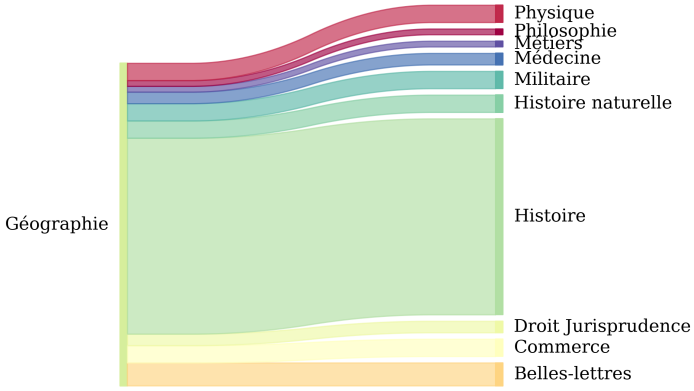
Figure 5.11 – Évolution des domaines des articles dans le sous-corpus Parallèle

En allant regarder dans le détail ce qui se passe pour les articles de *Géographie*, c'est-à-dire en s'intéressant à la ligne et la colonne de ce domaine, très pâles sur la figure 5.11, il ressort qu'il y a légèrement plus d'entrées qui étaient rattachées à cette discipline dans l'*EDdA* qui en changent dans *LGE* que l'inverse (56 contre 54). Leur nombre total relativement faible, 110, rend peu pertinente une étude statistique de leur répartition, mais permet d'en faire une revue qualitative exhaustive. La figure 5.12 montre que la plupart des changements de domaine pour des articles de *Géographie* concernent l'*Histoire*, mais dans les deux sens (aussi bien des articles de *Géographie* dans l'*EDdA* qui sont passé à l'*Histoire* dans *LGE* que le contraire).

Si l'ordre de grandeur du nombre de paires dont les domaines diffèrent prouve qu'elles ne peuvent toutes s'expliquer par les seules erreurs d'appariement, en revanche ce type d'erreur est évidemment très surreprésenté dans une sélection de paires d'articles aux domaines différents. Les erreurs apportent bien sûr peu d'information mais sont relativement peu nombreuses. On dénombre ainsi 15 paires dont les articles ne renvoient pas aux mêmes objets et qui n'auraient donc idéalement pas dûes être incorporées au sous-corpus $\mathcal{P}$ mais qui ont réussi à obtenir un score assez élevé pour être conservées par le processus précédent. À ces 15 paires s'ajoutent 12 autres dont la classe d'une des entrées a été mal prédite par le modèle. Parmi elles, six ne sont en fait que des renvois, pour lesquels la prédiction du domaine est évidemment très incertaine puisque le modèle ne peut se baser que sur la vedette de l'article pointé par le renvoi ce qui est rarement suffisant. Une troisième source d'erreurs est l'imperfection de la segmentation dans *LGE* (voir la sous-section 3.1.3 p.61) qui a par exemple mis l'entrée sur COMMUNAY (ville de l'Isère) avec l'entrée sur les Biens COMMUNAUX, faisant passer à *Géographie* dans *LGE* une vraie paire dont les deux entrées en correspondance relevaient du domaine *Droit Jurisprudence*. Ce type d'erreurs concerne cinq paires seulement.



(a) Part des domaines dans l'*EDdA* des articles passant à la *Géographie* dans *LGE*



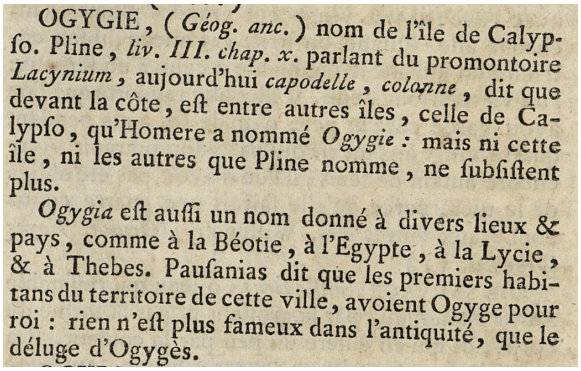
(b) Part des domaines dans *LGE* des articles issus de la *Géographie* dans l'*EDdA*

Figure 5.12 – Changements de domaines d'articles impliquant la *Géographie* dans le corpus Parallèle, de l'*EDdA* (à gauche) à *LGE* (à droite)

Ayant écarté ces 32 erreurs, les 78 autres paires se révèlent bien plus intéressantes. Elle peuvent essentiellement se répartir en trois groupes correspondant à trois mouvements distincts : la géographie abandonne d'abord les contenus imaginaires, elle acquiert ensuite des connaissances nouvelles – en particulier économiques et industrielles – sur les territoires qu'elle décrit et enfin cède à d'autres domaines des connaissances spécifiques qu'ils expliquent mieux grâce à de nouvelles théories ou de nouveaux faits importants, et qui relevaient en quelque sorte de la *Géographie* «par défaut» dans l'*EDdA*. Ces mouvements, décrits plus en détail dans les paragraphes suivants, peuvent conduire à des différences de domaine entre les entrées dans l'*EDdA* et dans *LGE* au sein de vraies paires, aussi bien au profit qu'au détriment de la *Géographie* selon que le domaine a été inféré par le modèle pour la moitié issue de l'*EDdA* ou bien décidé par ses éditeurs et selon que le contenu dans *LGE* en est resté proche ou non.

En ce qui concerne le premier de ces mouvements, la présence dans l'*EDdA* de désignants tels que «Géog. anc.» (ex. PEDUM, L'Encyclopédie, T12, p.238), «Géogr. & Myth.» (BATHOS, L'Encyclopédie, T2, p.141) et même «Géog. sacrée» (MASSADA, L'Encyclopédie, T10, p.176) est en soi assez édifiante. Le discours sur des lieux et des peuples historiques voire imaginaires relève dans ses pages de la *Géographie*, comme l'illustrent les articles HYPERBORÉENS (L'Encyclopédie, T8, p.405), JUTURNA (L'Encyclopédie, T9, p.102) ou OGYGIE (L'Encyclopédie, T11, p.429) reproduit à la figure 5.13a. Les

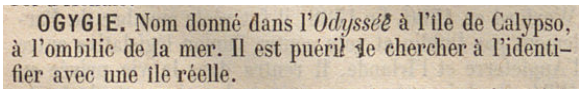
articles correspondants dans *LGE* (La Grande Encyclopédie, respectivement T20 p.481, T21 p.362 et T25 p.298 visible à la figure 5.13b) conservent un contenu proche et se retrouvent donc classés en *Histoire* ou en *Histoire Naturelle*. À l'inverse, des articles comme ERYMANTHE (La Grande Encyclopédie, T16, p.223) se concentrent sur les données factuelles de ces lieux réels et relèguent leur dimension mythologique à des remarques largement anecdotiques. Le modèle le classe donc à *Géographie*, alors que son traitement dans l'*EDdA* (L'Encyclopédie, T5, p.918) s'éloigne assez des articles de géographie type du XVIII[ème] siècle pour inclure des éléments mythologiques et recevoir du modèle la classe *Histoire*. On peut observer les signes de la même évolution pour des paires dont l'article dans l'*EDdA* est pourvu d'un désignant. Ainsi le traitement des entrées CLAIRETS (L'Encyclopédie, T3, p.500 et La Grande Encyclopédie, T11, p.528) et STONEHENGE (L'Encyclopédie, T15, p.535 et La Grande Encyclopédie, T30, p.520) est assez proche et, d'un point de vue moderne bien plus historique que géographique dans les deux œuvres. Les deux articles dans l'*EDdA* sont pourvus de désignants qui les classent en *Histoire* mais le modèle, trop entraîné à voir cette sorte de contenu historique dans des articles de *Géographie* a tout de même prédit cette classe sur leurs successeurs dans *LGE*.



OGYGIE, (*Géog. anc.*) nom de l'île de Calypſo. Pline, *liv. III. chap. x.* parlant du promontoire *Lacynium*, aujourd'hui *capodelle*, *colonne*, dit que devant la côte, eſt entre autres îles, celle de Calypſo, qu'Homere a nommé *Ogygie* : mais ni cette île, ni les autres que Pline nomme, ne ſubſiſtent plus.

Ogygia eſt auſſi un nom donné à divers lieux & pays, comme à la Béotie, à l'Egypte, à la Lycie, & à Thebes. Pauſanias dit que les premiers habitans du territoire de cette ville, avoient Ogyge pour roi : rien n'eſt plus fameux dans l'antiquité, que le déluge d'Ogygès.

(a) Article OGYGIE dans l'*EDdA*, T11, p.429



OGYGIE. Nom donné dans l'*Odyssée* à l'île de Calypso, à l'ombilic de la mer. Il est puéril de chercher à l'identifier avec une île réelle.

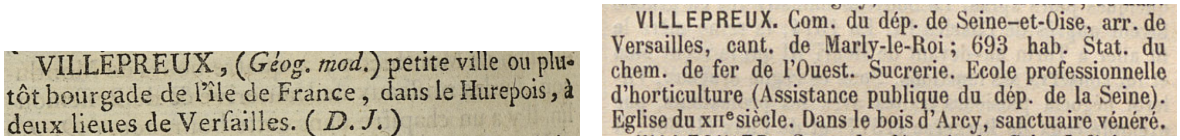
(b) Article OGYGIE dans *LGE*, T25, p.298

Figure 5.13 – Les articles sur l'île OGYGIE dans le sous-corpus Parallèle

Le deuxième mouvement concerne des paires pour lesquelles le modèle n'a pas reconnu une *Géographie* qui a trop évolué et intègre des informations trop nouvelles par rapport à celles qu'elle comportait au XVIII[ème] siècle, par leur forme ou leur contenu. Le groupe correspondant contient par exemple des paires comme HAMEAU dont l'article dans l'*EDdA* (L'Encyclopédie, T8, p.34) ne donne pas de définition mais se concentre sur l'étymologie du terme qu'il illustre par ses occurrences dans un grand nombre de toponymes ; son désignant le classe explicitement à *Géographie*. L'entrée a évolué dans *LGE* (La Grande Encyclopédie, T19, p.789) pour employer un certain nombre de termes formels comme «circonscription territoriale», «paroissial» ou «fraction de commune» et référencer par leurs numéros des articles du code civil et du code forestier si bien que le modèle ne reconnaît plus la discipline et propose la classe *Droit Jurisprudence*. Le même genre d'évolution semble être à l'œuvre dans le cas de la paire VILLEPREUX (L'Encyclopédie, T17, p.282 et La Grande Encyclopédie, T31, p.1007), visible à la figure 5.14, deux entrées très simples et «clairement» géographiques pour le lectorat contemporain. Le modèle échoue pourtant à reconnaître le domaine dans la moitié issue de *LGE*, sans doute à cause de la présence d'éléments anachroniques pour l'*EDdA* sur laquelle il a été entraîné comme «chem. de fer» ou l'adjectif «professionnelle» qui n'est attesté qu'à partir de 1842. Il est intéressant de remarquer que cet article débute par le motif des communes identifié à la section 4.3.2 p.128, mais ne fait pas partie des articles identifiés à cause de sa longueur [4], précisément due à ces phrases nominales en fin d'article qui mentionnent des infrastructures présentes sur la commune. Ces informations supplémentaires qui font que l'article n'a pas été repéré directement

4. pour rappel, un seuil de 50 tokens avait été imposé pour diminuer les faux-positifs et ne pas risquer d'annoter *Géographie* trop d'articles simplement amalgamés avec un article bref à cause des problèmes de segmentation de `soprano` (voir section 3.1.3 p.61)

comme un article de commune (et n'a donc pas été annoté *Géographie*) n'ont pas permis au modèle d'identifier de la géographie car elles mentionnent des realias trop différentes de celles existant au XVIII[ème] siècle.



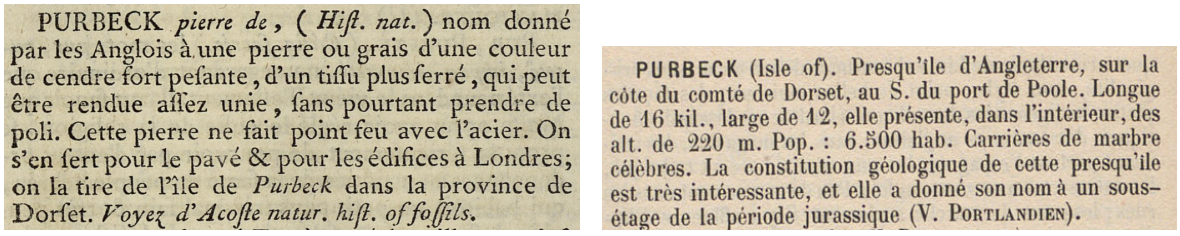
VILLEPREUX, (*Géog. mod.*) petite ville ou plutôt bourgade de l'île de France, dans le Hurepois, à deux lieues de Verſailles. (*D. J.*)

VILLEPREUX. Com. du dép. de Seine-et-Oise, arr. de Versailles, cant. de Marly-le-Roi ; 693 hab. Stat. du chem. de fer de l'Ouest. Sucrerie. Ecole professionnelle d'horticulture (Assistance publique du dép. de la Seine). Eglise du XII[e] siècle. Dans le bois d'Arcy, sanctuaire vénéré.

(a) Article VILLEPREUX dans l'*EDdA*, T17, p.282 (b) Article VILLEPREUX dans *LGE*, T31, p.1007

Figure 5.14 – Les articles VILLEPREUX dans le sous-corpus Parallèle

Enfin, l'évolution des connaissances ou l'arrivée de faits nouveaux a permis à d'autres disciplines de s'emparer de thématiques qui relevaient de la géographie dans l'*EDdA*. Là où l'entrée VOLCAN dans l'*EDdA* (L'Encyclopédie, T17, p.445) se cantonne par exemple à une description externe du concept avant de lister les plus connus à différents endroits du globe ce qui justifie pleinement son classement en *Géographie* par les éditeurs, l'article correspondant dans *LGE* (La Grande Encyclopédie, T31, p.1104) en fait une description bien plus structurelle, et introduit des termes comme «lave», «cheminée», «explosion», absents de la précédente ce qui conduit le modèle à le classer à *Physique*. Les trajectoires des paires PORPHYRITE (L'Encyclopédie, T13, p.127 et La Grande Encyclopédie, T27, p.328) et PURBECK (L'Encyclopédie, T13, p.576 et La Grande Encyclopédie, T27, p.964) sont assez semblables mais avec la classe *Histoire naturelle* et dans deux directions opposées. L'entrée PORPHYRITE de l'*EDdA* décrit le lieu d'origine d'une pierre particulière ; l'entrée correspondante dans *LGE* est un long article de géologie (désignant «Pétrogr.») sur la pierre elle-même. La figure 5.15 illustre le cas inversé de PURBECK : l'article est en *Histoire naturelle* dans l'*EDdA* où il décrit la pierre et mentionne l'endroit d'où elle provient dans le Dorset, alors que *LGE* couvre seulement la géographie de l'île dont elle est issue, et renvoie à un article séparé pour expliquer ses caractéristiques géologiques : le modèle le classe en *Géographie*. Il est aussi intéressant de voir qu'un certain nombre d'entrées géographiques assez brèves sur des fiefs dans l'*EDdA* – par exemple BRAGANCE (L'Encyclopédie, T2, p.392) ou INFANTADO (L'Encyclopédie, T8, p.697) – deviennent dans *LGE* des articles plus longs sur les familles possédant ces terres. Le choix éditorial de ne pas consacrer d'articles à des personnes (plus amplement discuté dans la section 5.2) suffit bien sûr à expliquer pourquoi ces articles portent sur les terres plutôt que sur les familles qui y règnent, mais leurs contreparties dans *LGE*, toutes deux classées en *Histoire*, sont l'occasion de décrire en détail la lignée de la maison de BRAGANCE (La Grande Encyclopédie, T7, p.959) en la mettant à jour et pour INFANTADO (La Grande Encyclopédie, T20, p.768) de raconter la vie d'un homme d'état espagnol qui s'est illustré pendant la première moitié du XIX[ème] siècle.



PURBECK *pierre de*, (*Hiſt. nat.*) nom donné par les Anglois à une pierre ou grais d'une couleur de cendre fort peſante, d'un tiſſu plus ſerré, qui peut être rendue aſſez unie, ſans pourtant prendre de poli. Cette pierre ne fait point feu avec l'acier. On s'en ſert pour le pavé & pour les édifices à Londres; on la tire de l'île de *Purbeck* dans la province de Dorſet. *Voyez d'Acoſte natur. hiſt. of foſſils.*

PURBECK (Isle of). Presqu'île d'Angleterre, sur la côte du comté de Dorset, au S. du port de Poole. Longue de 16 kil., large de 12, elle présente, dans l'intérieur, des alt. de 220 m. Pop. : 6.500 hab. Carrières de marbre célèbres. La constitution géologique de cette presqu'île est très intéressante, et elle a donné son nom à un sous-étage de la période jurassique (V. Portlandien).

(a) Article PURBECK dans l'*EDdA*, T13, p.576 (b) Article PURBECK dans *LGE*, T27, p.964

Figure 5.15 – Les articles PURBECK dans le sous-corpus Parallèle

Pour terminer, un dernier exemple éclairant de ce troisième mouvement concerne les deux paires ARCTIQUE et ANTARCTIQUE à la frontière entre *Géographie* et *Physique*. Dans l'*EDdA*, ANTARCTIQUE (L'Encyclopédie, T1, p.491) et ARCTIQUE (L'Encyclopédie, T1, p.621) sont tous deux signés

par d'Alembert, auteur (outre de nombreux articles de mathématiques évidemment) de plusieurs articles de *Géographie* construits sur une approche mathématique des concepts. La *Géographie* dans cette œuvre recouvre en effet encore les techniques géométriques utilisées pour mesurer la Terre, ce qui s'illustre bien par exemple dans l'article LATITUDE [5] (L'Encyclopédie, T9, p.302) du même auteur. L'entrée ANTARCTIQUE possède le double désignant «Astr. & Géog.», ce qui rappelle la place de la *Géographie* au sein de la *Cosmographie* dans le «Systême Figuré» (voir la section 3.1.2 à la page 58). De même, l'article ARCTIQUE est situé «en Astronomie» par son auteur qui définit la vedette en termes de cercles et de sphères. Les deux articles sont classés comme on peut s'y attendre à *Physique*.

La fin du XIX^{ème} siècle est marquée par de grandes explorations, entre autres vers les pôles qui deviennent peu à peu perceptibles en tant que territoire plutôt que comme de simples abstractions géométriques [6] ce qui conduit à un traitement très différent des deux mêmes entrées dans *LGE*. Si ANTARCTIQUE (La Grande Encyclopédie, T3, p.135) reste dans une certaine mesure similaire (on y parle d'«axe», de «cercle» et de «parallèle» dans un «I. ASTRONOMIE»), il y est tout de même question de «région», de «monde», et un «II. Géographie» qui consiste en un renvoi vers l'entrée OCÉAN suffit à ce que le modèle classe l'article à *Géographie*. La bascule est encore bien plus nette avec l'ARCTIQUE (La Grande Encyclopédie, T3, p.774) qui, après une phrase introductive pour définir l'Arctique en termes astronomiques, prend un tournant nettement géographique («En géographie») et utilise le point cardinal «Nord» avec et sans abréviations, ainsi que des termes comme «terre», «océan» qui permettent au modèle de le classer aussi en *Géographie*. Dans l'exemple de ces deux paires d'articles, on voit la géographie perdre ce qui relevait de la géométrie et se spécialiser pour garder ce qui lui est propre.

L'approche particulière de cette sous-section, en s'intéressant à des articles précis plutôt qu'à des domaines dans leur ensemble a permis de mener une étude plus qualitative. Les observations qui y ont été faites mettent en évidence des différences significatives dans les rôles et les méthodes de la géographie entre les articles dans l'*EDdA* et ceux dans *LGE*, dont le contenu s'individualise entre les deux époques par rapport à celui des autres disciplines.

En ajoutant les résultats de cette sous-section à ceux sur la place accordée aux discours géographiques obtenus au début de ce chapitre, une dynamique générale commence à apparaître dans les évolutions subies par la géographie. L'intervalle de temps est d'abord marqué par un fort accroissement de la volumétrie consacrée à la classe *Géographie*, avec un plus grand nombre d'articles, dont la plupart restent extrêmement brefs mais également une diversification des profils d'articles avec l'apparition d'articles bien plus longs que ceux consacrés à cette discipline dans l'*EDdA*. Parallèlement, elle se spécialise et gagne en technicité, utilisant au passage davantage de mots plus longs et structurant les informations au delà de simplement concentrer des noms de lieux et de personnes. Elle cède par ailleurs à d'autres disciplines des thématiques qui lui revenaient par défaut et donne bien plus d'informations plus détaillées sur celles qu'elle conserve.

5. Cet article contient dès le premier paragraphe l'expression «arc méridien», puis d'Alembert introduit des «cercles parallèles» dont il considère l'«intersection» dans le deuxième pour dans le troisième inviter son lectorat à considérer «un nombre infini de grands cercles».

6. De manière adjacente, l'article POLAIRE (La Grande Encyclopédie, T27, p.54) illustre bien cette tendance en consacrant plusieurs colonnes aux expéditions successives vers le pôle Nord ainsi qu'une carte en couleur sur une double page (un feuillet collé entre les pp.72 et 73) qui montre en détail toutes les îles et côtes de la zone et marque les points atteints par les différentes expéditions, jusqu'au tracé complet de l'expédition de Nansen en 1895. Le pôle Sud, encore bien moins accessible, est évoqué plus brièvement mais une carte lui est tout de même consacrée (p.59), moins précise et en noir et blanc mais reportant aussi les avancées majeures vers ce pôle.

5.2 La biographie cachée

Un parti pris éditorial remarquable de l'*EDdA* est de ne pas avoir inclus de biographies dans ses pages, choix assez emblématique du siècle des Lumières et qui correspond bien à l'idéal humaniste et démocratique de diffusion du savoir pour sa portée libératrice intrinsèque (D'Antuono, 2018, p. 99) plutôt que pour mettre en valeur des récits héroïques individuels. Les auteurs le justifient très simplement dans le Discours Préliminaire de l'*EDdA* en écrivant qu'un autre dictionnaire, celui de Moreri, contient déjà des notices sur les «noms de Rois, de Savans, & de Peuples» (L'Encyclopédie, T1, p.xlj).

Les pages de l'*EDdA* citent tout de même bien sûr des figures historiques autant qu'elles éclairent l'histoire des disciplines où elles se sont illustrées. Ainsi, l'article «MÉDECINS ANCIENS» (L'Encyclopédie, T10, p.276) offre un historique détaillé de la médecine illustré par les contributions de nombreux médecins célèbres depuis l'antiquité. Certains des paragraphes contiennent, en sus de la bibliographie et de la liste des apports de la personne qu'ils concernent, quelques éléments biographiques comme leur lieu et éventuellement année de naissance ou de mort. Mais de nombreuses biographies que l'on pourrait qualifier de «clandestines» ont également été rapportées dans la *Géographie* par exemple par Roe et al. (2016, p. 5) qui décrivent comment l'entrée sur le village de Wolstrope sert en fait de prétexte pour introduire une biographie de Newton. Particulièrement fréquentes sous la plume du chevalier de Jaucourt qui a été le contributeur le plus important de la *Géographie* dans l'*EDdA* signant 57.2% des articles du domaine d'après Laramée (2017, p. 166), de nombreuses autres biographies ont été rapportées dans des articles de *Géographie*, que ce soit plus loin par Laramée (2017, p. 169) ou par Vigier et al. (2022, p. 70).

Il est donc impossible d'envisager correctement la Géographie dans les pages de l'*EDdA* sans s'intéresser aussi à la place des biographies. Il faut d'abord déterminer si elles entretiennent une relation particulière avec les articles de géographie, ou si ces inclusions relèvent simplement des pratiques encyclopédiques de l'époque. Deux attitudes peuvent s'envisager face à la présence de ces textes : on peut ou bien les considérer comme des îlots de contenu non disciplinaires à l'intérieur des articles et les détournant de leur vocation première [7] ou bien considérer au contraire que puisqu'ils font partie des articles qui ont été écrits, ils sont constitutifs des discours de l'époque. Quoi qu'il en soit, une conséquence importante des choix éditoriaux des encyclopédistes pour les présents travaux est que le modèle utilisé n'a pas pu apprendre à reconnaître les biographies car elles ne sont pas identifiées explicitement dans les pages de l'*EDdA* et constituent donc un angle mort important du modèle et par conséquent des résultats qu'il est possible d'observer en l'utilisant.

Et la question est d'autant plus pertinente dans une perspective diachronique, puisque *LGE* comporte un très grand nombre de notices biographiques explicites sous forme d'entrées dont la vedette est le nom de famille de la personne décrite. La figure 5.16 le rappelle, la figure 3.7 l'illustrait déjà à la page 56. Faut-il donc penser que la Biographie aurait pris son indépendance de la Géographie entre le XVIII^{ème}s. et le XIX^{ème}s. ou plutôt y voir la preuve que la biographie a une existence propre et n'aurait jamais dû envahir l'espace consacré à cette science ?

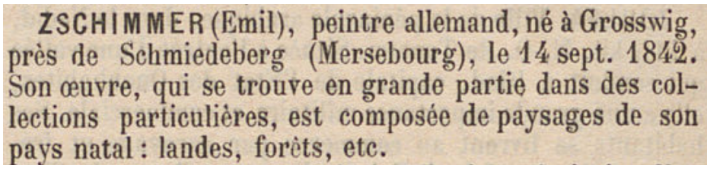

ZSCHIMMER (Emil), peintre allemand, né à Grosswig, près de Schmiedeberg (Mersebourg), le 14 sept. 1842. Son œuvre, qui se trouve en grande partie dans des collections particulières, est composée de paysages de son pays natal : landes, forêts, etc.

Figure 5.16 – Article ZSCHIMMER dans *LGE*, T31, p.1338

De plus, l'existence de la biographie comme domaine de connaissance en soi au même titre que les

7. et donc, dans le cas précis de la *Géographie*, comme autant de pommes de la discorde entre Diderot et Jaucourt

Mathématiques, la Médecine ou la Géographie ne relève pas de l'évidence. En revenant comme à la section 3.1.2 (voir p. 49) à l'avant-propos de *LGE* (La Grande Encyclopédie, T1, p.XI) on constate qu'il n'est fait aucune mention de la Biographie parmi les sciences citées alors qu'elles sont tellement omniprésentes qu'une page ouverte au hasard en contient statistiquement au moins une – le plus souvent même plusieurs – pourvu qu'elle ne soit pas à l'intérieur d'un article-essai s'étendant sur plusieurs pages.

On pourrait s'attendre à voir une biographie rattachée au domaine dans lequel s'est illustrée la personne à laquelle elle est consacrée, comme c'est par exemple le cas de celle sur SHAKESPEARE (La Grande Encyclopédie, T29, p.1142) pour laquelle le modèle prédit la classe «Belle-lettres», mais on constate en pratique qu'un grand nombre d'entre elles ont reçu la classe «Histoire». Ce choix n'est pas dénué de sens si l'on considère la forte composante narrative des biographies, écrites au passé, comportant fréquemment des références à des dates d'événements et à d'autres personnages historiques que l'on retrouve habituellement dans les articles d'Histoire. Cette remarque apporte d'ailleurs un premier éclairage possible sur le constat dressé précédemment à la section 5.1.1 sur la disproportion des articles d'«Histoire et Géographie» dans *LGE* d'après le modèle par rapport aux autres domaines. En effet, le très grand nombre de biographies et la très forte tendance du modèle à les considérer comme des articles d'Histoire pourrait être une bonne piste pour expliquer leur nombre anormalement élevé. Un corollaire amusant de ces constats est qu'une façon efficace de sélectionner une biographie au hasard dans *LGE* consiste à choisir le premier article classé «Histoire» dont le titre n'est ni un nom commun ni un nom de roi, de pape ou de personnage mythologique.

Cette section ne propose pas de solution définitive à la question de la place des biographies dans les articles de Géographie mais tente de formuler assez précisément le problème pour commencer à y réfléchir et suggère quelques critères simples basés sur une combinaison d'observations qualitatives et quantitatives pour aller au-delà du seul constat : «Jaucourt a piraté la Géographie de l'*EDdA* pour écrire des Biographies».

5.2.1 Spécificité des lemmes «naître» et «mourir» en *Géographie*

Les réflexions préliminaires de cette section amènent d'abord à examiner les relations entre la biographie et les différents domaines de connaissance. L'enjeu est de déterminer si les contenus biographiques sont réellement plus présents dans la Géographie ou si ce n'est qu'une impression donnée par le fait que les exemples connus de contenus biographiques dans l'*EDdA* apparaissent tous dans cette discipline. En partant du cas de WOLSTROPE (L'Encyclopédie, T17, p.630), qui utilise sa ville de naissance pour parler de Newton, il est naturel de prendre la notion de ville comme point de départ. L'étude des cooccurrents syntaxiques introduits à la section 2.5.2 (voir p.39) offre la possibilité d'observer les constructions dans lesquelles un terme comme «ville» est employé.

Concrètement, cette technique permet de trouver les mots les plus fréquemment associés à un motif donné à partir de considérations statistiques. En se basant sur les liens de syntaxe plutôt que sur l'ordre des mots dans la phrase, elle s'affranchit de la notion de fenêtre de recherche présente habituellement dans ce type de mesure et peut détecter des cooccurrents pertinents n'importe où dans les phrases où le motif apparaît. Dans le Lexicoscope (Kraif, 2016), une telle requête sur lemme «ville» étiqueté `NOUN` (nom commun) se formule en langage TQL [8] (voir le code source 5.1).

Code source 5.1 – Requête TQL sur le lemme «ville» annoté `NOUN` (nom commun)

```
<l=ville, c=NOUN,#1>
```

La figure 5.17 représente le lexicogramme des dix mesures les plus élevées obtenues pour cette requête. Sur cette figure, le déterminant «ce» apparaît comme cooccurrent principal de «ville», devant

8. http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_beta/doc/Reference%20TQL.fr.pdf

l'adjectif «natal». Deux autres mots outils figurent parmi les cinq premiers cooccurrents : le déterminant «le» (intéressant par comparaison avec «ce») ainsi que la préposition «dans», qui suggère l'emploi du concept de ville comme cadre circonstanciel dans lequel se produisent les événements qui intéressent les articles. Un deuxième adjectif, «petit», obtient également un score proche : l'asymétrie que crée cette observation par rapport au contraire «grand» semble suggérer qu'on qualifie bien plus volontiers une ville de «petite» que de «grande» dans les pages des encyclopédies du corpus.

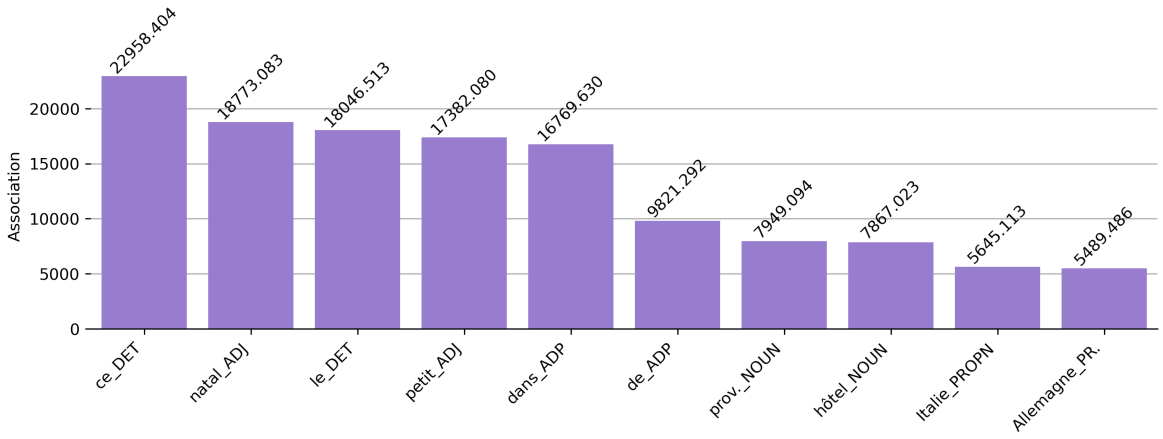


Figure 5.17 – Les 10 cooccurrents syntaxiques principaux du lemme «ville» annoté comme `NOUN` (nom commun)

Le Lexicoscope permet ensuite de générer automatiquement une autre requête qui intègre un des cooccurrents parmi ces résultats. Cette fonctionnalité correspond à une étape du procédé itératif permettant de construire progressivement un ALR, équivalent de la méthode des «segments répétés» pour les cooccurrents syntaxiques. La nouvelle requête obtenue pour le premier résultat, le déterminant «ce», revient à chercher les plus forts cooccurrents du motif «cette ville». L'extrait de code 5.2 montre cette requête formulée en TQL.

Code source 5.2 – Requête TQL sur la mise en relation syntaxique quelconque du lemme «ce» annoté comme `DET` et du lemme «ville» annoté comme `NOUN`

```
<l=ce,c=DET,#2>&&<l=ville,c=NOUN,#1>::(.*,1,2)
```

Son exécution sur le même corpus aboutit à la figure 5.18 qui retrouve la préposition «dans», mais cette fois avec la plus forte mesure d'association. Déjà présente sur la figure 5.17, la requête 5.1 ne pouvait par construction que rendre compte des intéractions de «ce» et de «dans» avec «ville» de manière séparée. Cette nouvelle mesure montre qu'en réalité les trois éléments apparaissent fréquemment ensemble dans le syntagme «dans cette ville». La préposition «de» suit une trajectoire semblable puis, après l'adjectif «dernier» avec une mesure d'association de 2 662, déjà bien plus faible, les autres valeurs chutent très rapidement pour un ensemble de verbes et de noms, à moins de 10% de celle obtenue par la préposition «dans».

En suivant ce nouveau cooccurrent «dans» on construit la requête 5.3, dont un équivalent sous forme d'arbre syntaxique en dépendances est visible à la figure 5.19 pour plus de clarté. Les astérisques sur les deux arêtes partant du nœud `ville_NOUN` représentent l'absence de contrainte sur les relations qui relient «ville» respectivement à «dans» et à «ce». En pratique, on peut s'attendre à ce que ces relations soient réalisées par un `case` pour «dans» et par un `det` pour «ce» [9] mais le motif demeure flexible sur ce point. De plus, travaillant au niveau de la syntaxe et pas de la réalisation de surface des mots, la requête inclut des résultats comme «dans ces deux villes» à l'article ARABIE (L'Encyclopédie,

9. https://universaldependencies.org/u/dep/index.html

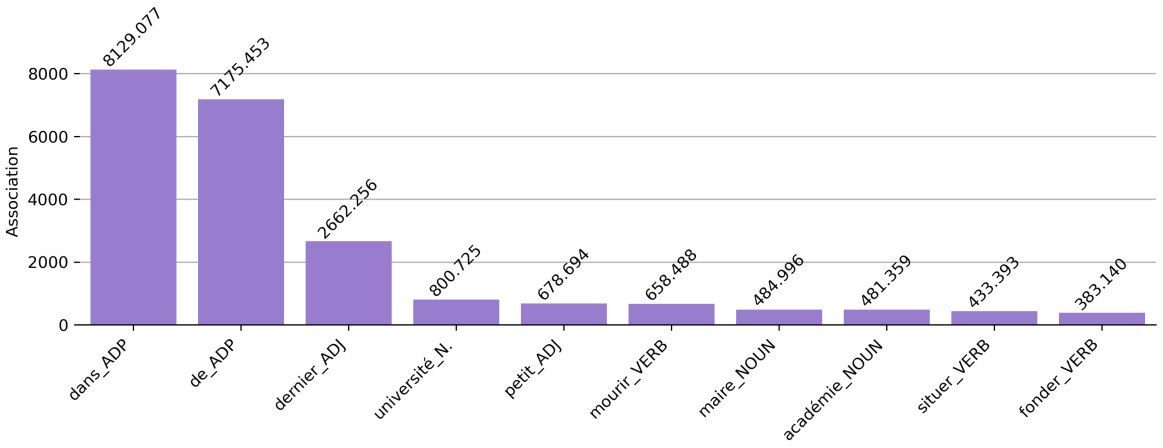


Figure 5.18 – Les 10 cooccurrents syntaxiques principaux du motif formé d'une relation syntaxique quelconque entre le lemme «ce» annoté comme DET et le lemme «ville» annoté comme NOUN

T1, p.570) ou «dans cette dernière ville» à l'article DAEHLING (La Grande Encyclopédie, T13, p.749). L'usage de ce motif est répandu dans le corpus d'étude comme en atteste la dispersion des résultats : utilisé par au moins 27 contributeurs distincts (26 dans l'*EDdA* et les métadonnées disponibles pour les articles de *LGE* ne comportent pas d'auteur), il apparaît dans 1 674 articles.

Code source 5.3 – Requête TQL sur le motif syntaxique «dans cette ville»

```
<l=dans,c=PREP,#3>&&<l=ce,c=DET,#2>&&<l=ville,c=NOUN,#1>::(.*,1,2)(.*,1,3)
```

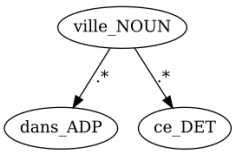


Figure 5.19 – Représentation sous forme d'arbre de syntaxe en dépendance du motif défini par la requête 5.3

En recherchant ce troisième motif, on obtient le lexicogramme de la figure 5.20 qui représente les dix cooccurrents de «dans cette ville» avec la plus forte mesure d'association. Sur cette figure, il n'y a quasiment plus que des verbes et des adjectifs. Le phénomène de remontée d'un lemme, déjà observé au cours du processus, se produit à nouveau avec l'adjectif «dernier» qui n'obtenait que la troisième place sur la figure 5.18 mais devient le cooccurrent principal de «dans cette ville». De la même manière, le verbe «mourir» prend le deuxième rang, alors qu'il n'était que sixième pour «cette ville». De plus, le troisième rang est occupé par le verbe «naître», absent jusqu'ici des résultats mais de sens contraire à «mourir» et proche de l'adjectif «natal» observé lors de l'analyse de la figure 5.17. À travers ce vocabulaire susceptible d'apparaître dans des contenus biographiques, on retrouve le procédé utilisé dans l'article WOLSTROPE à l'origine de cette étude. Toutefois, en étant parti seulement du terme «ville», le fait d'observer ce thème émerger du fait d'associations syntaxiques statistiquement significatives montre que le phénomène ne se limite pas à ce seul article mais est assez général dans le corpus.

Ces deux lemmes inattendus se présentent donc comme des objets incontournables pour poursuivre l'exploration. Dans TXM (Heiden et al., 2010), on commence par se restreindre au sous-corpus ne contenant que les articles de l'*EDdA*, qu'on partitionne par domaine de connaissance. L'outil «Index»

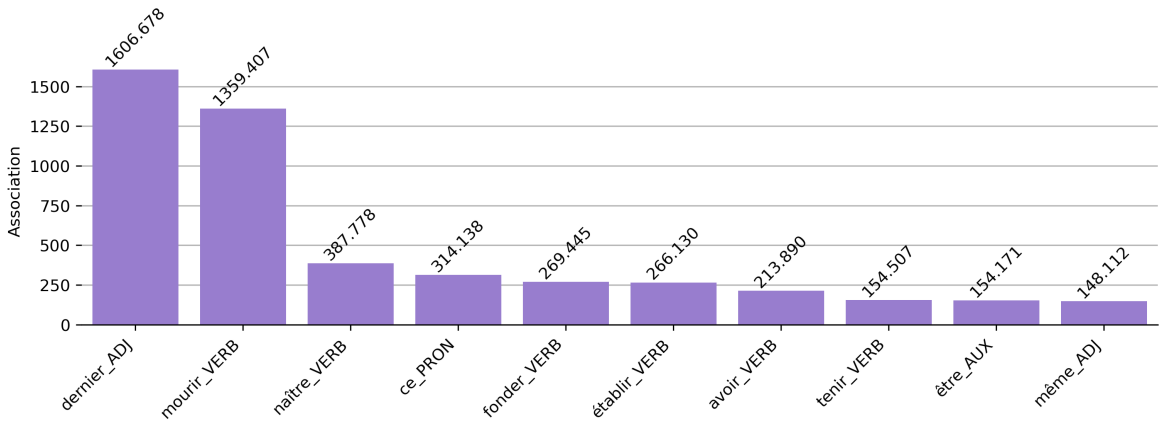


Figure 5.20 – Les 10 cooccurrents syntaxiques principaux de «dans» (ADP), «ce» (DET), et «ville» (NOUN)

paramétré pour compter les lemmes et appliqué sur cette partition avec la requête CQL dont le code est visible à l'extrait 5.4 permet de compter les occurrences des deux verbes «naître» et «mourir» dans le corpus. Un nombre d'occurrences n'est pas pertinent en soi sans être rapporté à la taille des différentes partitions. Le calcul des spécificités (voir la section 2.5.2 p.38) implémenté dans TXM opère cette traduction pour révéler la quantité d'information contenue dans la distribution de formes observées (à quel point elle est susceptible d'apparaître spontanément seulement du fait du hasard ou témoigne au contraire d'un phénomène).

Code source 5.4 – Requête CQL sur tout mot dont le lemme est «naître» ou «mourir»

```
[lemma="(naître|mourir)"]
```

Sur la figure 5.21 qui montre le résultat de ces calculs, les deux lemmes apparaissent comme extrêmement spécifiques du domaine *Géographie*. Le verbe «mourir» obtient une spécificité supérieure à 318 ce qui exclut en pratique toute formation spontanée de cette distribution du vocabulaire. La surreprésentation du lemme «naître» dans la même partition, certes moindre que le précédent avec un score de «seulement» 84, ne peut pas non plus s'expliquer seulement par le hasard. En ce qui concerne les autres partitions, le profil est tout autre : les deux lemmes sont très sous-représentés dans toutes les autres domaines sauf *Philosophie* (mais où seul «naître» se distingue, «mourir» reste peu significatif avec une spécificité de 1), *Histoire Naturelle* (où «mourir» est même sous-spécifique) et *Beaux-arts* («mourir» et «naître» y obtiennent des scores respectifs de 13 et 19, assez élevés pour justifier éventuellement une étude séparée, mais tout de même nettement inférieurs à ceux observés en *Géographie*). Les contenus biographiques semblent donc bel et bien entretenir une relation particulière avec la *Géographie* dans les pages de l'*EDdA*.

L'étape suivante du raisonnement consiste naturellement à se demander si cette observation est propre à l'*EDdA* ou si elle se généralise à d'autres discours encyclopédiques. Pour répondre à cette question, on procède de la même façon en partitionnant par domaine de connaissance le sous-corpus ne contenant que les articles de *LGE* avant de calculer les spécificités des deux mêmes lemmes. Les spécificités obtenues suivent un profil bien plus tranché, de nombreux domaines obtenant pour les deux verbes des scores «saturés» (voir la section 2.5.2 p.39 sur ce point), en positif ou en négatif, c'est-à-dire que les deux y sont «infiniment» spécifiques ou sous-spécifiques. Partout ailleurs, les spécificités sont négatives ou très faibles (seul «naître» obtient un score de 8 pour la *Musique*). En ce qui concerne la *Géographie*, les deux lemmes sont très fortement sous-représentés avec des spécificités de -1000. Le phénomène détecté dans l'*EDdA* semble donc avoir disparu 130 ans plus tard dans *LGE*.

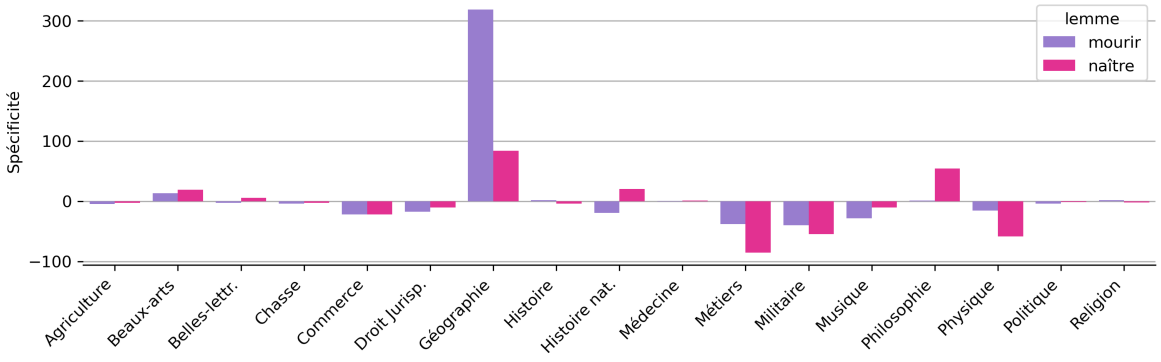


Figure 5.21 – Spécificités des lemmes «naître» et «mourir» par superdomaine dans l'*EDdA*

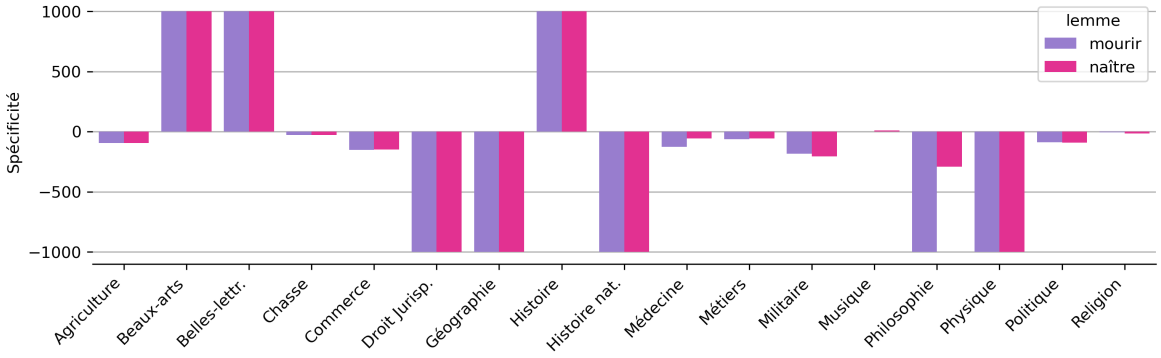


Figure 5.22 – Spécificités des lemmes «naître» et «mourir» par superdomaine dans *LGE*

Pour cerner tout à fait l'emploi de ces verbes révélateurs de contenus biographiques dans les entrées de *Géographie*, il ne reste plus qu'à déterminer dans quelle mesure ils dépendent du style personnel des contributeurs de l'*EDdA*. En effet, le nom de Jaucourt revient fréquemment à propos de ce type d'inclusions auxquelles Diderot s'est opposé et dont il semble s'être parfois agacé (Laramée, 2017, p. 169). Pour achever cette étude, on partitionne donc cette fois le sous-corpus de l'*EDdA* suivant deux critères, pour croiser à la fois les notions d'auteur et de domaine de connaissance. Pour la première, il faut en effet pouvoir distinguer les écrits de Jaucourt, ceux de Diderot, ceux des autres auteurs et ceux restés anonymes (pour lesquels il n'est pas en tout rigueur possible d'écarter la possibilité qu'ils aient été écrits par Jaucourt ou Diderot). Puisque certains articles résultent de la collaboration entre plusieurs auteurs, il est nécessaire de traiter à part les quatre articles écrits à la fois par Jaucourt et Diderot –les entrées CHANVRE (L'Encyclopédie, T3, p.147), CHAUSSURE (L'Encyclopédie, T3, p.259), Ere Philippique (L'Encyclopédie, T5, p.902) et JALOUSIE (L'Encyclopédie, T8, p.439). La notion de domaine demande moins de finesse car, la surreprésentation écrasante des lemmes «naître» et «mourir» en *Géographie* ayant déjà été établie, il suffit pour chacune des configurations possibles d'auteur(s) de comparer leurs emplois de ces deux verbes dans et hors de la *Géographie*. Comme aucun des quatre articles écrits par Jaucourt et Diderot ensemble ne porte sur la *Géographie*, la partition correspondant à cette configuration reste vide.

Pour rendre compte des deux dimensions d'analyse simultanées, la figure 5.23 comporte deux distributions de spécificités, mesurées en *Géographie* et hors de cette classe pour les mêmes auteurs. Le trait principal qui ressort de cette figure est la prépondérance des deux verbes chez le chevalier de Jaucourt en *Géographie*, très au-dessus du seuil de banalité fixé par convention à 3. Ensuite, l'emploi du verbe «naître» reste assez élevé dans les autres articles du même auteur (spécificité de 86, contre 138 en *Géographie*) ce qui souligne l'importance du style de cet auteur, outre le domaine de la *Géo-*

graphie dans l'emploi de ce verbe. La différence d'ordre de grandeur entre les différentes valeurs mesurées empêche toute analyse plus fine car les petites mesures et le seuil de banalité lui-même se confondent avec l'axe des abscisses.

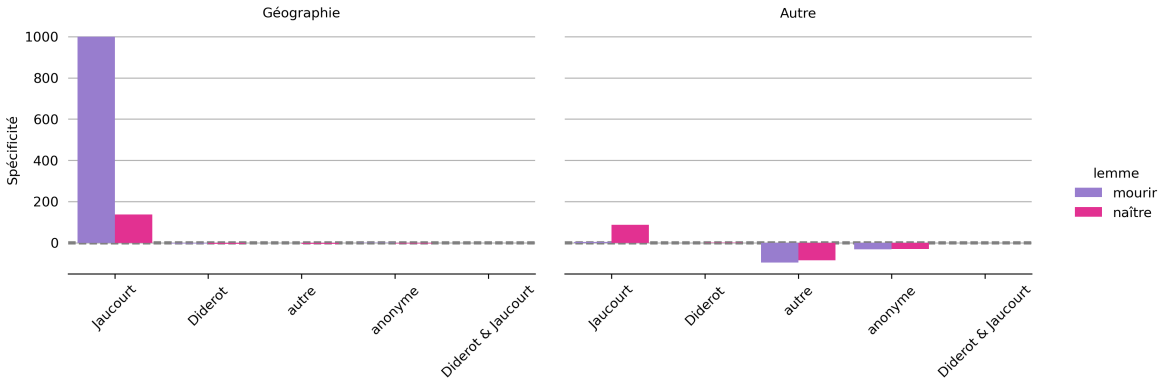


Figure 5.23 – Spécificités des lemmes «naître» et «mourir» par auteur, en *Géographie* et hors de ce domaine

Pour aller plus loin, il faut masquer les colonnes correspondant à Jaucourt pour obtenir la figure 5.24. Cette dernière figure montre que les deux lemmes de l'étude obtiennent des spécificités fortement négatives hors de la *Géographie* pour les autres auteurs que Diderot (c'est-à-dire qu'ils sont sous-représentés de manière statistiquement très significative dans les partitions correspondantes). En négatif, cette observation révèle deux choses. La première est que les autres auteurs utilisent tout de même ces deux verbes dans les articles de *Géographie*, à un niveau assez moyen mais donc supérieur au sous-emploi qui en est fait hors de ce domaine. Ces deux lemmes et les contenus biographiques qu'ils permettent d'écrire ne sont donc pas exclusifs à Jaucourt. La seconde réside dans le fait que Diderot ne semble quant à lui pas sujet au même régime : son emploi de «naître» et «mourir» reste dans les seuils de banalité hors de la *Géographie* et même très légèrement positif alors qu'il reste négatif sur les articles de cet auteur sur le même domaine (le lemme «naître» y est même assez rare avec une spécificité de -4.6). Diderot apparaît donc plus constant dans son style d'écriture et utilise davantage les deux verbes hors de la *Géographie*.

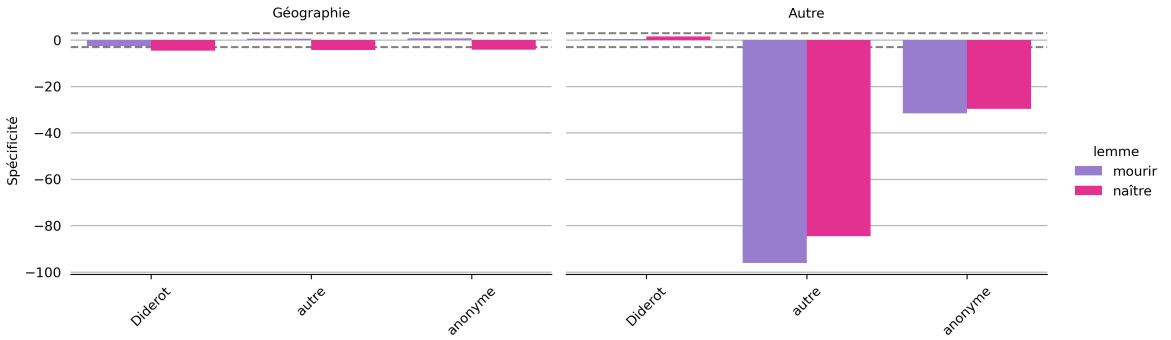


Figure 5.24 – Spécificités des lemmes «naître» et «mourir» par auteur à l'exception de Jaucourt, en *Géographie* et hors de ce domaine

Cette étude révèle donc une répartition nuancée des contenus biographiques dans l'*EDdA*. Il existe certes un «effet Jaucourt» qui concentre ces contenus, en particulier dans la *Géographie*, mais également un signal de fond qui démontre l'emploi de ces éléments par les autres auteurs pour traiter le domaine, comme le fait d'Holbach à l'article ISLANDE (L'Encyclopédie, T8, p.916).

5.2.2 L'influence des domaines

La section précédente a permis de quantifier l'importance des contenus biographiques en *Géographie* mais, ce faisant, a aussi mis en lumière l'emploi de verbes associés à ces contenus dans d'autres domaines. C'est ainsi que la figure 5.21 présentait les *Beaux-arts* et surtout la *Philosophie* comme des pistes intéressantes pour la recherche de biographies. Malgré le choix éditorial de ne pas mettre en avant de figures individuelles, le Discours Préliminaire de l'*EDdA* (L'Encyclopédie, T1, p.xlj) fait en effet mention de nombre de «génies» qui ont façonné les domaines où ils se sont illustrés. L'attention particulière qui leur est portée dans ce Discours Préliminaire – et le fait que Newton dont on sait qu'il y a une biographie dans les pages de l'*EDdA* soit parmi eux – laisse à penser qu'ils font de bons candidats pour trouver d'éventuelles biographies.

En cherchant leurs noms parmi les vedettes de l'*EDdA*, on trouve celui de Locke à l'article «LOCKE, Philosophie de» (L'Encyclopédie, T9, p.625). Contrairement à ce que pourrait laisser penser sa vedette, l'article ne se limite pas à la philosophie de Locke et, sur toute sa première moitié, donne un récit de la vie du philosophe de sa naissance à sa mort, riche en dates, lieux et mentions d'autres personnages historiques qui lui sont liés. Il apparaît tout à fait comparable à l'article WOLSTROPE qui, bien que trois fois plus long environ (7 277 mots contre 2 659) contient lui aussi environ une moitié d'éléments biographiques, le reste de l'article concernant directement les travaux de Newton, ses théories et jusqu'à l'exposition des travaux de ses précurseurs comme le paragraphe sur les lois de Kepler.

En plus de constituer un premier exemple de biographie individuelle hors de la Géographie (le désignant de l'article est «Hist. de la Philosoph. moder» et le modèle le classe à *Philosophie*), cette entrée donne une piste d'explication pour le pic observé dans ce domaine à la figure 5.21 et dévoile une stratégie pour en trouver d'autres : on peut trouver des informations biographiques sur une personne dans un article motivé par la discussion de sa philosophie. Et cette stratégie s'avère payante puisque l'on en trouve neuf de cette manière, présentés dans le tableau 5.2. Certains contiennent davantage de récits et de détails personnels (LÉIBNITZIANISME en est particulièrement riche), d'autres sont des attaques assez virulentes contre le philosophe qu'ils décrivent (surtout CAMPANELLA) mais tous contiennent à la fois des éléments biographiques et des détails sur ses travaux. On remarque de plus qu'aucun article n'est de Jaucourt, et que les articles JORDANUS BRUNUS, LÉIBNITZIANISME, LOCKE, PLATONISME et PYTHAGORISME sont tous les cinq signés par Diderot. Ce fait est d'autant plus remarquable quand on considère que la «sécheresse» du style de Diderot dans les articles de Géographie – dont il a dû se défendre dans l'article ENCYCLOPÉDIE [10] – est souvent opposée à celui de Jaucourt, fourmillant et érudit, quand il s'agit de décrire la «contrebande» biographique de ce dernier. De plus, ayant publié une biographie de Leibnitz en 1734 (Ferenczi, 2017, p. 87), le chevalier de Jaucourt était particulièrement bien placé pour l'écrire dans l'*EDdA* ; il se contentera pourtant à l'article LEIPSIC (L'Encyclopédie, T9, p.380) de le nommer et de citer Voltaire à son sujet avant de renvoyer son lectorat à l'article LÉIBNITZIANISME. Ces observations montrent donc non seulement qu'il y a des biographies individuelles hors des articles de Géographie (dans la Philosophie), mais qu'en plus il n'y avait pas non plus de «tabou» biographique que Jaucourt aurait enfreint seul et de son propre chef.

Tableau 5.2 – Coordonnées d'articles de philosophie de l'*EDdA* contenant des éléments biographiques

Philosophe	Vedette	Tome	Page	#article
Bacon	BACONISME ou PHILOSOPHIE DE BACON	2	8	92

10. voir sur ce point la section 1.1 page 8

Philosophe	Vedette	Tome	Page	#article
Campanella	CAMPANELLA	2	576	5039
Jérôme Cardan	CARDAN	2	675	5624
René Descartes	CARTÉSIANISME	2	716	5851
Giordano Bruno	JORDANUS BRUNUS, Philosophie de	8	881	3608
Leibnitz	LÉIBNITZIANISME ou PHILOSOPHIE DE LÉIBNITZ	9	369	1730
John Locke	LOCKE, Philosophie de	9	625	2934
Platon	PLATONISME ou Philosophie de Platon	12	745	3325
Pythagore	PYTHAGORISME ou Philosophie de Pythagore	13	614	2470

Puisque les biographies de l'*EDdA* ne sont pas identifiées explicitement, il n'est pas possible d'en dresser une liste exhaustive et de dire avec certitude qu'une biographie est absente de ses pages, mais il semble bien plus difficile d'en trouver ailleurs qu'en Philosophie. En reprenant les noms du Discours Préliminaire, on remarque que pour chaque philosophe de la liste ci-dessus, l'article ayant le plus grand nombre d'occurrences de ce nom (ou au moins un des plus grands comme pour Bacon où il arrive «seulement» au 3ème rang) contiennent l'article qui traite conjointement de sa philosophie et de sa vie. En appliquant la même heuristique à Boërhaave, Boyle et Huyghens les résultats sont différents. L'article ayant le plus d'occurrences de Boërhaave (plus précisément de `[word="Bo[eë]rhaave"]`, la requête CQL utilisée pour ramener ces résultats [11]) est l'article de Médecine CRISE (L'Encyclopédie, T4, p.471) qui ne le mentionne que pour ses contributions sur le sujet (avec 29 occurrences), mais le deuxième avec le plus de mentions est un article de Géographie signé par Jaucourt : l'article VOORHOUT (L'Encyclopédie, T17, p.468) qui offre un nouvel exemple de biographie «complète» dans un article prétexte comme l'était WOLSTROPE. Le nom de Boërhaave apparaît dès la première phrase en termes laudatifs et l'intégralité de l'article se partage entre les détails de sa vie personnelle et professionnelle. Au lieu que sa vie soit racontée dans un article sur ses contributions principales au domaine comme c'était le cas pour les philosophes, sa biographie est «déguisée» en article de Géographie.

Pour Boyle et Huyghens, les articles avec le plus grand nombre d'occurrences de leurs noms sont purement techniques et portent exclusivement sur leurs thématiques de recherches, respectivement la Chimie – AIR (L'Encyclopédie, T1, p.225) et CHYMIE (L'Encyclopédie, T3, p.408) – et la Physique et les Mathématiques – «Figure de la Terre» (L'Encyclopédie, T6, p.749), CYCLOIDE (L'Encyclopédie, T4, p.590), PENDULE (L'Encyclopédie, T12, p.293) et «Ressort spiral» (L'Encyclopédie, T14, p.188). Mais, en sélectionnant dans la liste des résultats non plus les articles ayant le plus d'occurrences mais ceux en Géographie, émergent alors les articles LISMORE (L'Encyclopédie, T9, p.574) – ville d'origine de Boyle – et «HAIE (la)» (L'Encyclopédie, T8, p.23) – celle de Huyghens. Les deux sont également de Jaucourt mais à la différence de VOORHOUT il s'agit plutôt de biographies «partielles». L'article sur LISMORE contient un paragraphe entier sur la ville, son importance politique (faible) et sa position, et même la biographie de Boyle dans le second paragraphe est introduite par rapport au contexte historique de la ville, ce qui fait que Boyle occupe strictement moins que la moitié du texte. Quant à celui sur La Haye, en plus de détailler l'étymologie du nom de la ville, son histoire et sa place contemporaine dans le pays et l'Europe, il cite non seulement Huyghens mais également Jacques Golius, Jean Meursius, Frédéric Ruysch, Albert Henri de Sallengre et Jean Second. L'article s'achève même par une «biographie-mystère» d'un «monarque célèbre» mais pas nommé, dans laquelle il

11. l'orthographe des noms propres peut parfois varier dans les pages de l'*EDdA* surtout quand il y a des diacritiques

n'est pas difficile de reconnaître Guillaume III d'Orange-Nassau. Une fois encore, si ces trois exemples ne constituent pas une démonstration définitive que les savants hors de la Philosophie n'ont pas leur biographie dans l'article disciplinaire décrivant leur doctrine, ils semblent toutefois bien le suggérer, et si des biographies d'autres savants moins mis en avant dans le Discours Préliminaire et moins mentionnés en général dans les pages de l'*EDdA* venaient à être mise au jour dans des articles de spécialité, il s'agirait surtout de comprendre pourquoi Boërhaave, Boyle et Huyghens n'ont pas eu le même honneur.

En admettant cette inférence, il semble donc y avoir un régime différent pour les philosophes par rapport aux autres savants. Cela pourrait s'expliquer par les différences entre disciplines : là où les physiciens travaillent généralement sur des problèmes variés, les théories philosophiques qui par nature tendent à décrire le monde dans son ensemble sont plus enclines à former des «systèmes» caractéristiques de leur auteur et difficiles à séparer d'eux. Il est sans doute malaisé de bien distinguer aussi nettement les disciplines à une époque de polymathie encore assez générale mais il semble tout de même y avoir une importance particulière accordée à la Philosophie. Alors que Newton est surtout considéré actuellement comme un physicien, sa contribution est particulièrement valorisée dans les pages de l'*EDdA* et mise en regard de celles de Descartes et de Leibnitz qu'il a affronté sur certains problèmes. Les trois savants se retrouvent dans des contextes similaires, et ont des cooccurrents très proches : des mots comme «analyse», «philosophie», «calcul», «principes» et «démontrés» et même plus directement leurs noms (par exemple `"Newton"` et `"Descartes"` sont dans les cinq cooccurrents les plus forts de `"Leibnitz"` et il en va de même pour les deux autres combinaisons qu'on peut former de leurs trois noms). Et on constate même qu'à la différence des autres physiciens il possède lui aussi un article à son nom : NEWTONIANISME (L'Encyclopédie, T11, p.122), reformulé en «Philosophie Newtonienne». Il semble que ce terme, bien qu'il soit explicité de plusieurs façons («physique expérimentale», «mécanique & mathématique», «partie de la Physique»), soit surtout utilisé pour souligner l'ampleur du saut conceptuel que représente la démarche de Newton par rapport à ses prédécesseurs. L'article est très différent des articles de doctrines philosophiques discutés plus haut ne serait-ce que par son lexique : des mots comme «âme» et «Dieu» qui apparaissent fréquemment chez les autres philosophes sont absents de l'article NEWTONIANISME, et le mot «corps» par exemple qui était opposé à âme et discuté dans un rapport à soi-même chez eux y devient un terme de physique : «corps physique», «corps céleste», «corps pesans» (sic). Ces différences sont particulièrement visibles sur le diagramme d'AFC de la figure 5.25 représentant les 65 lemmes les plus fréquents parmi les textes de doctrines philosophiques (ceux du tableau 5.2, le nuage de points rouges sur la figure) auxquels est rajouté NEWTONIANISME (le point `EDdA_11_679` très nettement à part sur la figure). Newton est donc mis au même niveau que les grands philosophes alors que le contenu de ses travaux est tout à fait différent. La vraie question qui se pose n'est donc pas «pourquoi les biographies sont-elles cachées dans la Géographie ?» car elles ne le sont pas toutes, mais bien plutôt «pourquoi celle de Newton n'a-t-elle pas pu figurer dans son propre article ?». Faut-il y voir un lien avec les reproches faits à Newton mentionnés dans le Discours Préliminaire et la manière dont il y est opposé à Descartes, jusqu'à suggérer qu'il était difficile en France de s'en déclarer partisan, comme a pu le faire Maupertuis «le premier qui ait osé parmi nous se déclarer ouvertement Newtonien» ?

5.2.3 Différents profils de passages biographiques

Proportion de contenu biographique

Si on revient à l'article WOLSTROPE qui est en grande partie à la base de toutes les réflexions de cette section, la situation est limpide car Jaucourt ne fait pas le moindre effort pour déguiser sa biographie : le bourg n'est cité qu'en relation à Isaac Newton dès la première phrase et les deux seules informations, lapidaires, que l'on pourrait qualifier de géographiques à son propos tiennent dans les huit premiers mots, en un syntagme nominal et un complément circonstanciel : «bourg

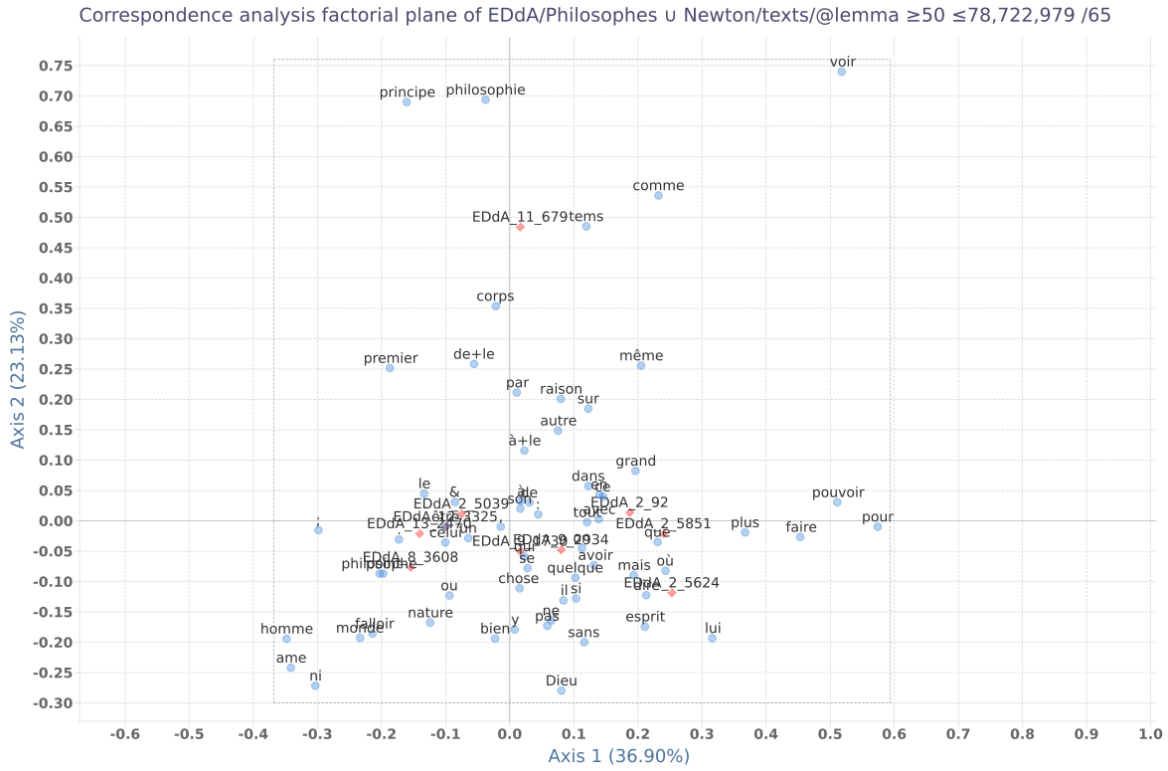


Figure 5.25 – Analyse Factorielle des Correspondances pour les articles de systèmes philosophiques dans l'*EDdA*, y compris NEWTONIANISME

d'Angleterre, dans le comté de Lincoln». Aucune coordonnée, aucune distance ni même direction à partir d'une autre ville ou point remarquable du territoire anglais.

Mais la situation est quand même toute autre avec l'article LODEVE (L'Encyclopédie, T9, p.627). L'article commence par un paragraphe entier qui en plus de donner une précision d'ordre administrative («un évêché suffragant») dresse un rapide historique de la ville depuis son nom latin attesté par Pline jusqu'à donner le gentilé de ses habitants pour se conclure par un concentré d'informations sur le paysage et le climat («sec et stérile»), l'activité économique de la ville («manufactures de draps & de chapeaux») et sa situation par rapport aux cours d'eau et reliefs («sur la Lergue, au pié des Cévennes») avant de la situer en distance par rapport à cinq villes différentes classées par ordre croissant de distance et d'importance dont les deux dernières avec en prime une direction sur la rose des vents. Et de se conclure, de surcroît, par des coordonnées en longitude et latitude, basées vraisemblablement sur le méridien de l'île de Fer (longitude ≥ 20 pour un lieu en France métropolitaine). C'est quand même se donner beaucoup de peine pour seulement parler du cardinal de Fleury. D'autant plus que, non seulement la biographie du cardinal occupe déjà moins de lignes que le seul premier paragraphe de l'article, mais en plus il n'est pas le seul à être mentionné. Une phrase présente deux cardinaux nés à Lodève, et chacun a droit a son propre paragraphe. On est donc manifestement en présence de deux phénomènes très différents.

Parmi les autres articles cités par Vigier et al. (2022), SOULIERS (L'Encyclopédie, T15, p.406) est bien plus proche du modèle de WOLSTROPE quoique la première phrase réussisse à s'achever avant de citer Antoine Arena. L'article VALOGNE (L'Encyclopédie, T16, p.824) en revanche, bien qu'il accorde à Jean de Launoi une place plus importante que celle qu'occupe le cardinal de Fleury dans l'article LODEVE, commence tout de même par un paragraphe géographique complet. Dans le même cas, on peut également citer l'article NICE (L'Encyclopédie, T11, p.131) qui accorde une très large place à Cas-

sini mais seulement après avoir parlé de la ville elle-même. Ces différents exemples semblent donc s'inscrire dans un spectre continu de possibilités entre les deux tendances extrêmes en termes de proportion de contenu que représenteraient WOLSTROPE – biographie «complète» – et LODEVE – biographie «partielle» – qui constitue un premier critère pour caractériser les biographies de l'*EDdA*.

Motivation de l'article

Les biographies «cachées» dans la Géographie rendent la navigation particulièrement difficile puisqu'elles sont inaccessibles à moins de connaître une «clef» particulière : le lieu de naissance de la personne recherchée. Pour en trouver, une stratégie consiste à penser à des personnages historiques dont on espère qu'il pourrait y avoir une biographie, à trouver leur lieu de naissance dans une autre source et à consulter l'article du lieu en question. On trouve ainsi par exemple l'article Ferté-Milon (L'Encyclopédie, T6, p.556) qui est en fait le lieu de naissance de l'auteur Racine (figure 5.26). En plus de faire étrangement écho à la remarque ci-dessus sur la prédominance de la Philosophie – on reproche quand même à demi-mot à Racine de n'avoir été «que» poète et pas philosophe – l'article très bref signé par Jaucourt pose un défi d'interprétation. En effet, l'auteur semble dire que le seul intérêt de la ville – et par là même de l'article – réside dans la mention de Racine («uniquement remarquable par»). Mais il est dit bien peu de choses sur le tragédien : sa date de mort et l'âge qu'il avait alors, le domaine dans lequel il a exercé et le succès qu'il y a rencontré. Même ces 4 lignes qui précèdent la remarque lui reprochant de ne pas avoir fait de philosophie ne parviennent pas à lui être entièrement consacrées puisqu'il y est également fait mention de Corneille. Sur les sept lignes et demie que comprend l'article, cela ne fait donc qu'une grosse moitié du volume de l'article réellement dédiée à Racine. Le critère portant sur la proportion de l'article dédiée à la biographie est donc d'une utilité très limitée à l'échelle d'un article aussi court. Jaucourt parvient quand même à glisser des coordonnées géographiques avant sa signature, on pourrait presque se demander s'il lui importe de donner de l'information sur le lieu, malgré sa remarque initiale dépréciative sur l'intérêt de ce dernier.

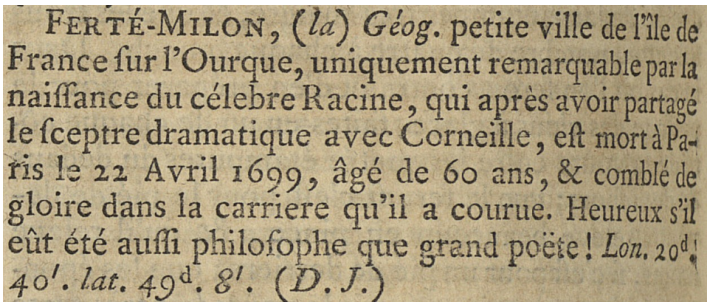

FERTÉ-MILON, (*la*) *Géog.* petite ville de l'île de France sur l'Ourque, uniquement remarquable par la naissance du célebre Racine, qui après avoir partagé le sceptre dramatique avec Corneille, est mort à Paris le 22 Avril 1699, âgé de 60 ans, & comblé de gloire dans la carriere qu'il a courue. Heureux s'il eût été auſſi philoſophe que grand poëte! *Lon.* 20^d. 40'. *lat.* 49^d. 8'. (*D. J.*)

Figure 5.26 – Article "Ferté-Milon" *EDdA*, T6

Ce trait dépréciatif se rencontre d'ailleurs dans plusieurs biographies cachées dans la Géographie par exemple dans VOORHOUT «village de Hollande [...] mais village illustré» (l'opposition avec la conjonction de coordination «mais» semble montrer que village revêt ici un caractère négatif) ou dans Villa-nueva (L'Encyclopédie, T17, p.274) «qui n'est connu que pour avoir donné la naissance à Michel Servet» dans une autre biographie complète cachée, qui ne parle en fait pas d'autre chose que de ce personnage historique. D'ailleurs, les recherches avec la stratégie illustrée ci-dessus pour Racine montrent assez vite qu'il est d'autant plus probable de trouver une biographie complète que le lieu de naissance de la personne est petit. Une ville de très grandes dimensions comme Paris est bien sûr le lieu de naissance de nombreux personnages historiques importants, mais ce fait ne permet pas d'accéder à leurs biographies, car l'article Paris pourtant long de 23 272 mots n'en contient aucune ! Après avoir écarté l'évidence dans son deuxième paragraphe «Elle a produit plus de grands personnages [...] que toutes les autres villes de France [...]», l'article passe sous silence le fait qu'il

n'en donnera pas. Une explication parfois avancée pour la présence de biographies dans les articles de Géographie est une tendance de Jaucourt à laisser son érudition prendre la place : c'est pourtant lui qui signe cet article, et qui y consacre le dernier quart (600 lignes !) pour dresser un parallèle entre Paris et l'Athènes antique. L'érudition, le style personnel de Jaucourt et la simple longueur des articles ne suffisent donc pas à prédire la présence de passages biographiques.

En essayant de trouver un juste milieu entre la Ferté-Milon et Paris, pour regarder des villes de taille intermédiaire, on trouve alors très facilement un grand nombre de biographies : l'entrée pour TOURS (L'Encyclopédie, T16, p.490) en contient 8 dont deux frères qui sont décrits indépendamment, celle pour TROYES (L'Encyclopédie, T16, p.719) en contient 9 sans liens entre les personnes, ROUEN atteint le total de 24. À chaque fois le motif est semblable : une suite de biographies succintes sans liens entre elles et présentée implicitement comme un passage obligé des articles («je passe aux simples hommes de lettre natifs de», «l'abondance m'oblige de m'arrêter à cette liste», «je ne me propose que d'indiquer ici les principaux»), ce qui est renforcé par l'usage fréquent de l'ordre alphabétique. Ce que disent en négatif ces énumérations, c'est qu'elles sont là parce que le lectorat de ces articles s'attendent à les trouver, plus que pour l'intérêt de la vie individuelle de chaque personne. Ce qui ne signifie pas qu'elles doivent être anecdotiques : les exemples de la section 5.2.2 contiennent des savants majeurs de leurs domaines cités brièvement dans les villes qui les ont vu naître : même Leibnitz, dont la vraie biographie est à l'article LÉIBNITZIANISME, est mentionné dans l'article LEIPSIC, en même temps qu'une remarque éclairante sur la motivation profonde de ces biographies : «Leibnitz seul auroit suffi pour donner du relief à Leipsic sa patrie». Ces énumérations sont là pour l'anecdote, elles ornent, donnent davantage d'intérêt aux lieux traités. En cela elles font partie du discours géographique de l'époque.

Toutes ces remarques convergent donc pour dégager un dernier critère à appliquer aux biographies trouvées dans les articles de Géographie : celui de la motivation «biographique» ou «intrinsèque» de l'article, c'est-à-dire à quel point il ne doit son existence qu'à la personne dont il contient la biographie ou bien au contraire à quel point le passage biographique ne fait qu'apporter une information à propos d'un endroit sur lequel il y aurait eu de toute façon une entrée. La figure 5.27 illustre les deux critères dégagés ci-dessus en plaçant les articles les plus emblématiques discutés au fil de cette section. Chaque axe de la figure correspond à un des critères : celui des abscisses à la proportion de contenu biographique dans l'article calculée comme le rapport du nombre de lignes consacrées à des personnages historiques sur le nombre total de lignes de l'article ; celui des ordonnées à une note quelque peu arbitraire tâchant de rendre compte des arguments discutés pour chaque article, notamment de la quantité et de la précision du contenu géographique, des remarques dépréciatives éventuelles des auteurs, du nombre de biographies présentes dans le même article et de l'importance relative de la ville dont il est possible de se faire une idée par exemple en cherchant le nombre d'occurrences dans l'ensemble du texte de l'*EDdA*.

Il ressort donc de ce qui précède que la Géographie dans l'*EDdA* a effectivement servi dans une certaine mesure à abriter pour diverses raisons des textes purement biographiques. En revanche cette étude révèle également que dans le contexte de l'époque, une certaine part de contenu biographique était au moins acceptable sinon attendue dans un article de Géographie. C'est d'ailleurs peut-être ce fait qui a rendu possible l'opération de dissimulation des biographies : en effet, comment celles-ci auraient pu être cachées dans les articles de Géographie si elles avaient choqué par leur présence ?

La situation est tout autre à la fin du XIX^{ème} siècle dans les pages de *LGE* où les biographies ont leur place attitrée. Isaac Newton a alors son propre article «NEWTON (Sir Isaac)» (La Grande Encyclopédie, T24, p.1048) dans lequel il est simplement fait mention de son lieu de naissance, désormais orthographié «Whoolstorpe» et sans article. Un autre «article-prétexte», celui de Boërhaave, suit le même chemin et fait l'objet d'une entrée (La Grande Encyclopédie, T7, p.42) qui mentionne «Woorhout», sans entrée associée. Pour les autres articles, on constate que VALOGNES (La Grande Encyclopédie,

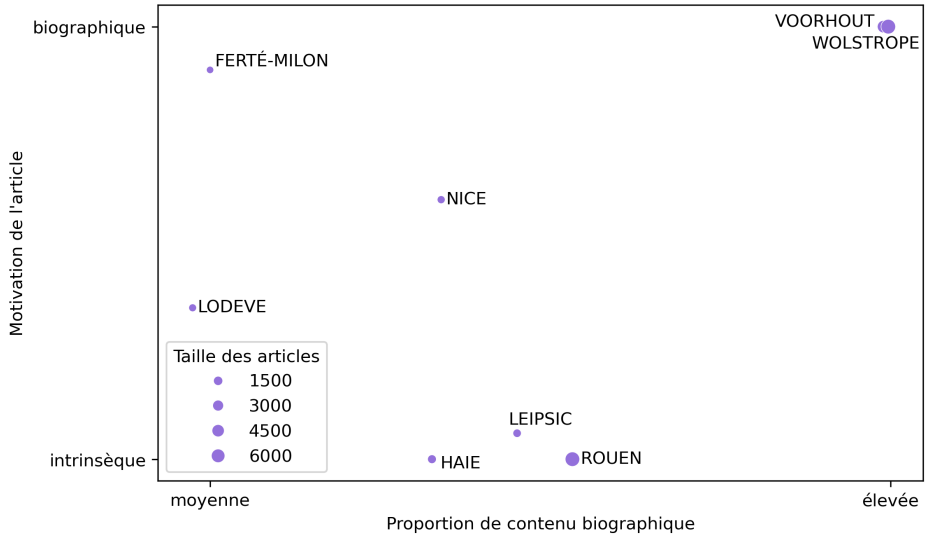


Figure 5.27 – Visualisation d'articles de l'*EDdA* discutés dans la section suivant les deux critères identifiés

T31, p.683) ne cite plus Jean de Launoi alors que que LODÈVE (La Grande Encyclopédie, T22, p.404) retient une simple mention du cardinal de Fleury à l'issu d'un paragraphe sur les bâtiments religieux de la ville, sans date ni caractère narratif. Pour une ville de taille intermédiaire comme TROYES (La Grande Encyclopédie, T31, p.432), une section Grands Hommes est présente, au même niveau que les sections Monuments, Histoire et Évêques de Troyes. Cette partie se trouve à la toute fin de l'article, juste avant la bibliographie, confirmant son caractère plus pittoresque. Elle consiste en une énumération de noms et d'années de décès, parfois sans même de titre ni de profession. Évoquer les personnes qui ont vécu à un endroit donné reste donc une des fonctions attendues de la Géographie dans *LGE*, tandis qu'une frontière plus nette se dessine entre la biographie en tant que telle et ce qui n'est plus qu'anecdote.

Chapitre 6

Conclusion

6.1 Synthèse des résultats 171
6.2 Perspectives 174

6.1 Synthèse des résultats

Les trois chapitres précédents convergent, avec des approches très différentes mais complémentaires, vers une réponse à la problématique initiale consistant à bâtir une méthodologie permettant d'étudier les discours géographiques en diachronie dans les encyclopédies (voir p.10). En consignant les enseignements tirés de la préparation et de la manipulation du corpus, en montrant la pertinence des techniques de classification automatique sur ce type de texte tout en tirant parti de leurs limites, puis en mobilisant des outils statistiques pour la linguistique, ils démontrent la possibilité d'identifier ces discours pour observer leurs évolutions.

Identification des discours géographiques

La solution proposée se caractérise d'abord par son pragmatisme et sa souplesse face à des contraintes difficiles à anticiper imposées par les données et les outils utilisés. Cela se manifeste par exemple dans la décision de croiser les sources de l'ARTFL et de l'ENCCRE pour associer un domaine de connaissance aux textes des articles, ou encore sur des modifications en cours de projet de la structure des métadonnées – dont la clef primaire des textes (voir la section 3.3.2 page 84 à ce sujet). L'ensemble de la chaîne de traitement repose donc sur la composition d'éléments individuellement simples plutôt que sur un bloc monolithique qui aurait nécessité une connaissance préalable parfaite de toutes les difficultés potentielles contenues dans le corpus. Ce choix de conception a favorisé un travail itératif, la chaîne restant facile à relancer à partir de n'importe quel point. À son tour, cette façon de travailler a fortement bénéficié de l'application d'outils reproductibles permettant de contrôler les paramètres modifiés d'une itération à l'autre. En plus de contribuer à une science ouverte et transparente, de telles techniques facilitent donc les approches adaptatives.

Les différentes étapes de traitement suivent une approche progressive en termes de granularité des objets manipulés. L'extraction de l'ensemble du texte de chaque tome des deux œuvres pour le segmenter en articles représente la partie la plus complexe du système. L'encodage de *LGE* en particulier a demandé un effort important de traitements géométriques et de repérage de motifs dans des

chaînes de caractères avec du bruit causé par l'OCR. Au niveau des articles eux-mêmes, l'essentiel des annotations et analyses est réalisé par des outils existants, et des scripts servent à mettre en forme les données, charger des librairies ou extraire leurs résultats. L'avantage majeur de la démarche suivie réside dans la modularité qu'elle offre : en séparant les différentes représentations des textes et les métadonnées qu'elle leur associe, elle permet une grande liberté pour les combiner et formater les données en vue de nouvelles analyses, parfois propres seulement à un sous-corpus. Enfin, au niveau des discours qui composent les articles, la spécificité de la méthode utilisée se situe dans la mise en regard d'études quantitatives et qualitatives, là encore en utilisant des logiciels répandus en analyse textuelle des données.

Dans l'ensemble, le code produit dans le cadre de cette thèse remplit donc essentiellement trois fonctions : il modélise la structure du corpus pour faciliter la manipulation des fichiers, il convertit des formats pour servir de liant entre les différents logiciels utilisés et il produit des visualisations pour l'analyse comme celles présentes dans ce manuscrit. Il se répartit principalement entre sept dépôts de code (voir la section 3.3.3 p.89) dont la taille totale avoisine 200 fichiers pour 10 000 lignes de code écrites en quatre langages différents. La plus large partie est en haskell, qui a servi à implémenter le programme `soprano` et la librairie de modélisation `ghc-geode` mais aussi une partie des scripts d'analyse et de conversion. Le reste consiste en des scripts, surtout en python mais aussi en bash pour quelques uns. La description des paquets et de leurs dépendances est écrite en guile et occupe un peu plus de 10% du volume de code pour assurer le contrôle des environnements logiciels à l'aide de `guix`.

L'usage de cet outil pour la reproductibilité représente la seule brique logicielle complexe de la méthodologie suivie. Ce choix restreint nécessairement un peu la portabilité mais même cette contrainte est moins forte qu'il n'y paraît. D'abord, `guix` peut être utilisé seulement comme gestionnaire de paquet et installé sur de nombreux systèmes linux, pas seulement sur le système d'exploitation `guix`. Ensuite, d'autres implémentations comme `nix` donnent accès aux mêmes techniques de reproductibilité et ont déjà été portées sur davantage de plateformes. En outre, ces technologies nouvelles permettent de générer des environnements sous de nombreux formats comme des archives portables ou des conteneurs qui peuvent être déployés sur la plupart des plateformes existantes, sans que `guix` ou `nix` eux-mêmes n'aient besoin d'y être installés.

Enfin, les formats aussi suivent cette logique de réemploi et de modularité. Les représentations intermédiaires des textes utilisent des standards ouverts comme les dialectes du XML, le JSON, le texte encodé en UTF-8 ou encore le TSV pour lesquels des outils existent sur la plupart des plateformes et qui restent faciles à implémenter sinon. Le choix du TSV pour stocker les métadonnées plutôt que de véritables bases de données s'est avéré payant en termes de flexibilité grâce à un petit outillage vaste et polyvalant capable d'exploiter ce format (des programmes de recherche et de remplacement comme `grep`, `sed` ou même `csvsql` pour écrire des requêtes d'une complexité intermédiaire). Cet aspect de la méthodologie suivie accroît son utilité potentielle pour de futurs travaux, en ne rajoutant pas de nouvel outil tout-en-un difficile à supporter sur un grand nombre de plateformes mais en encourageant au contraire à utiliser ce qui est disponible.

Les facettes de la disciplinarisation

L'étude diachronique des changements survenus dans les discours géographiques joue un double rôle dans ce manuscrit. Elle revêt d'abord une valeur illustrative, essentielle pour montrer la pertinence des méthodes proposées. Cependant, elle possède également une valeur intrinsèque dans la perspective du projet GEODE qui s'intéresse aux changements subis par la Géographie dans les encyclopédies entre le XVIII^{ème} et le XIX^{ème} siècle. Les observations faites ont permis de mettre en lumière différents aspects des transformations qui se sont produites sur cette période, et notamment la constitution de la Géographie en une véritable discipline scientifique.

Cette disciplinarisation se traduit d'abord par une complexification des articles de géographie en termes de richesse du vocabulaire, constitué en moyenne de mots plus longs (voir la section 5.1 p.133). Ce phénomène est observable malgré une haute concentration d'abréviations dans la plupart des entrées de la classe *Géographie* qui décrivent succintement un lieu sans donner plus de précision. Mais ce seul fait ne doit pas occulter la finesse des mouvements contraires et simultanés auquels la Géographie est sujette.

En premier lieu, la ressemblance entre la *Géographie* et les autres classes s'accentue entre l'*EDdA* et *LGE*. Cela se manifeste par exemple au niveau de la distribution du nombre de mots par article, bien plus similaire à celle des autres domaines de connaissance dans la partie du corpus correspondant à *LGE* (voir section 5.1.1 p.135). Des articles plus longs apparaissent dans le domaine, particulièrement bref par contraste dans l'*EDdA*. Une autre conséquence de ces rapprochements peut s'observer en considérant les occurrences d'EN comme le fait la section 5.1.1 (p.141). Alors que les philosophes des Lumières emploient surtout les EN dans les articles de la classe *Géographie*, le profil d'utilisation des différents types relevés dans l'annotation géo-sémantique des articles s'harmonise entre les disciplines. Les noms propres de lieu en particulier (NP-Spatial) se propagent dans une certaine mesure aux autres domaines.

Simultanément, la Géographie semble aussi se cloisonner. La très large majorité des articles du domaine dans l'*EDdA* étaient extrêmements brefs et constitués d'une longue phrase nominale. Cette tendance, propre au domaine, s'accroît dans *LGE*, où ces articles se font plus nombreux et où le recours presque systématique aux abréviations se double d'un «figement» des différents compléments qui composent la phrase nominale, jusqu'à l'obtention de données pour ainsi dire tabulaires. Par ailleurs, d'un domaine central et en quelque sorte «refuge» pour de nombreuses entrées peu scientifiques (mythologie et phénomènes naturels inexpliqués entre autres), les discours géographiques s'épurent et s'affirment par contraste comme la description de territoires.

Les passages de biographie ne disparaissent toutefois pas entièrement des discours géographiques. Si les biographies clandestines telles que celles écrites par Jaucourt n'ont plus lieu d'être dans les pages de *LGE*, des articles de *Géographie* traitant de villes conservent quelques mentions de personnages célèbres qui y ont vécu.

Contributions

Version numérique structurée de *LGE*

Le premier résultat apporté par cette thèse est la définition d'un schéma d'encodage XML-TEI pour représenter une encyclopédie. Le schéma qui se restreint au module *core* sans utiliser les éléments du module *dictionaries* tient compte des particularités de ces œuvres par opposition aux dictionnaires et permet de représenter des développements arbitrairement longs et fortement structurés. Cet encodage a été partiellement appliqué à *LGE* pour représenter la structure de ses articles au-dessus du niveau des paragraphes.

Cette contribution se présente concrètement sous la forme d'un ensemble de dépôts de code source et d'archives sur l'entrepôt de données Nakala (Vigier & Brenon, 2021) contenant les fichiers représentant les articles dans différents formats, accompagnés de leurs métadonnées. Le logiciel central dans le processus d'encodage est `soprano` [1] qui opère la segmentation du texte de *LGE* en articles et applique l'encodage défini. Au cours de cette thèse, le logiciel a reçu plusieurs améliorations qui augmentent notamment la précision de son algorithme de segmentation. Il s'accompagne d'autres dépôts comme `Processing LGE` [2] qui contient des métadonnées et des scripts pour faciliter le traitement

1. https://gitlab.inria.fr/abrenon/soprano
2. https://gitlab.huma-num.fr/disco-lge/processinglge

de l'œuvre dans son ensemble et `LGE Meta`[3] qui présente sous un format tabulaire les métadonnées présentes au début de l'œuvre dont la liste des abréviations utilisées afin de faciliter la prise en main des données à de futures équipes de recherche.

Modèle de classification en domaines de connaissance

La comparaison de modèles d'AA en vue de classer les articles par domaine de connaissance a apporté trois résultats principaux. Le premier confirme la supériorité de l'architecture Transformers et des modèles *BERT* préentraînés en particulier en termes de qualité des résultats pour cette tâche de classification. Le deuxième montre la pertinence de méthodes classiques qui obtiennent des performances proches – et supérieures à celles des méthodes d'AP autres que *BERT* – pour un coût en énergie bien moindre. L'application des prédictions d'un modèle *BERT*, rectifiées pour un ensemble d'articles de *Géographie* courts de *LGE* (identifiés comme traitant de communes et regroupés au sein d'un sous-corpus), a non seulement rendu possibles les analyses conduites au chapitre 5 mais a surtout permis d'enrichir la version numérique de *LGE* d'une métadonnée associant un domaine de connaissance à chaque entrée.

D'autre part, l'analyse des erreurs d'une des architectures classiques ayant obtenu d'assez bons résultats (celle associant un *TF-IDF* pour la vectorisation et *SGD* pour la classification elle-même) a permis de visualiser les liens entre la *Géographie* et les autres domaines dans l'*EDdA*. Les contenus géographiques infusent de manière sélective dans certains autres domaines au centre desquels la *Géographie* fait en quelque sorte office de passerelle. L'étude a en outre montré que ces ressemblances entre domaines n'étaient pas seulement lexicales. L'ensemble de ces résultats a fait l'objet d'une publication (Brenon et al., 2022).

Biographies et discours géographiques

La troisième contribution présentée dans cette thèse concerne la relation privilégiée qui existe entre les biographies et les discours géographiques. Les spécialistes de l'*EDdA* connaissent bien les biographies déguisées en articles de *Géographie* en utilisant le lieu de naissance de la personne décrite comme point d'entrée. La section 5.2 dresse un état des lieux du phénomène, ce qui permet ensuite de mettre en évidence un autre filon de contenus biographiques dans l'*EDdA* au travers des articles de *Philosophie*. En comparant le traitement de la philosophie de Newton et de celles des autres philosophes, l'étude renverse la perspective sur l'article WOLSTROPE en montrant que la question posée par cette entrée n'est pas tant l'existence d'une biographie du savant anglais que le choix de son lieu de naissance pour l'introduire.

Enfin la section apporte des éléments nouveaux pour comprendre la place des biographies dans les articles de *Géographie* en distinguant entre différents types de contenus biographiques dont la valeur illustrative varie d'un article à l'autre. Après avoir défini deux critères pour caractériser ces contenus, elle montre la survivance au XIX^{ème} siècle de ceux qui relèvent de l'anecdote et sont principalement utilisés pour colorer la manière dont une ville est présentée.

6.2 Perspectives

Au-delà des résultats discutés dans la section précédente, les présents travaux ont également permis de dégager des pistes prometteuses pour de futures recherches. Il s'agit pour une part d'améliorations techniques car les logiciels peuvent bien sûr toujours être perfectionnés davantage mais ces nouveaux chemins à explorer ne se limitent pas à des détails d'implémentation. Les méthodes utilisées dans le cadre de cette thèse pourraient trouver des applications sur d'autres corpus et, de plus, les données produites pourraient intéresser des études sur d'autres types de discours.

3. Le fichier correspondant à ces données est distribué dans le dépôt https://gitlab.liris.cnrs.fr/geode/lge-meta.git

Des améliorations techniques

La recherche d'exemples dans des articles, nécessaire pour illustrer différents points du propos, a été grandement facilitée par les interfaces de consultations de l'*EDdA* mises en place par l'ARTFL et l'ENCCRE. De telles interfaces permettent de naviguer rapidement dans les tomes de l'œuvre, de rechercher des articles et d'opérer un retour au texte en les visualisant sur des photographies des pages. Pour ces raisons, elles sont utiles d'une manière générale tant au public qu'aux équipes de recherche. Les photos des pages numérisées et le texte brut de *LGE* sont consultables sur le site de la BnF [4] mais si ce dernier autorise la recherche de texte de manière indiscriminée, il ne permet pas de trouver une vedette spécifique ni même de naviguer par article. L'état actuel de la segmentation de *LGE* et la précision des métadonnées de navigation qui comprennent le numéro des pages ont rendu faisable le même type de recherches sans trop de friction lors de la rédaction de ce manuscrit mais l'interface expérimentale déployée à cet effet n'est pas encore assez pratique pour être utile hors du projet GEODE. Quelques efforts supplémentaires de développement restent nécessaires pour la rendre plus intuitive, pour y implémenter la recherche par vedette et pour trouver une solution pérenne pour l'héberger.

Il a aussi été envisagé durant cette thèse d'étudier les variations thématiques et discursives au sein des articles en utilisant les méthodes de classification à l'échelle des paragraphes. La segmentation en paragraphes est disponible pour l'*EDdA* mais pas encore assez robuste pour *LGE* (voir la section 3.2.2 p.76). Il est à souhaiter que les prochains mois permettent de régler les derniers problèmes et de réaliser un nouvel encodage de cette œuvre qui intègre une telle segmentation. Malgré ces lacunes, quelques efforts ont déjà été entrepris pour définir un ensemble de fonctions que les paragraphes remplissent dans les textes encyclopédiques. À partir d'une observation empirique d'articles de l'*EDdA* jusqu'à saturation des données, huit grand rôles type ont été identifiés. En tout, 1 722 paragraphes ont été annotés manuellement pour entraîner un modèle capable de reconnaître ces fonctions. Des F-mesures légèrement supérieures à 0.70 pour l'ensemble des classes ont été obtenues assez rapidement mais il s'est avéré impossible d'améliorer ces scores, en particulier pour certaines classes restant à des valeurs nettement inférieures. Cet échec relatif amène plusieurs pistes de réflexion pour de futurs travaux. Un travail théorique plus poussé permettrait de vérifier la pertinence des classes choisies. Il faudrait sans doute également s'assurer de la représentativité des exemples annotés pour l'instant, ou tenter simplement de poursuivre l'annotation en espérant atteindre un seuil qui permettrait des gains significatifs grâce à la seule masse de données.

Outre les besoins propres à cette étude pour l'instant mise en pause, d'autres améliorations de la qualité de l'encodage de *LGE* paraissent encore réalisables. La détection des vedettes reste à ce jour rudimentaire et bénéficierait d'un travail en profondeur pour comprendre la grammaire qui permettrait de les décrire. En effet celles-ci se composent souvent de plusieurs mots, dont certains peuvent employer des minuscules ou des petites majuscules, jusqu'à pouvoir s'étendre sur plusieurs lignes alors que l'encodage actuel ne conserve que le premier mot en majuscules. Un travail similaire d'identification des motifs présents à la fin des articles pourrait donner accès à une métadonnée précieuse comme les auteurs des articles, souvent mentionnés et parfois différents d'une partie à l'autre d'un article (et qu'il pourrait donc être intéressant d'associer au niveau des paragraphes quand celui-ci sera disponible). Un tel effort devrait idéalement prendre aussi en compte l'important travail de bibliographie réalisé dans *LGE*, pour pouvoir encoder ces contenus correctement et les traiter séparément. Il y a en effet peu d'intérêt à les annoter en syntaxe (ce qui est malheureusement fait dans l'état actuel de la chaîne), alors qu'à l'inverse ils pourraient constituer des métadonnées précieuses pour comprendre le paysage académique de référence de l'époque. Enfin, la résolution des renvois représente certainement le problème le plus délicat pour plusieurs raisons. L'imprécision de l'OCR d'abord rend un certain nombre de «V.», abréviation de «voir», comme la séquence «l'» ce qui com-

4. https://catalogue.bnf.fr/ark:/12148/cb41651490t

plique leur détection. Ensuite, les vedettes formées de plusieurs mots augmentent les risques que la séquence utilisée pour un renvoi diffère légèrement de celle utilisée à l'article cible lui-même – sans compter l'usage d'expressions référentielles en lieu et place de la vedette explicite telles que «voir ce mot». À ces difficultés s'ajoute la distance entre un renvoi et sa cible, potentiellement présente dans un tome distinct alors que le traitement actuel par `soprano` considère les tomes séparément. La résolution de ces problèmes et l'ajout d'une fonctionnalité pour suivre les renvois entre articles apporterait pourtant encore bien plus d'intérêt à une interface web de consultation de l'œuvre.

Enfin la conduite des analyses contrastives a révélé certains défauts récurrents dans l'annotation des textes en morphosyntaxe et en syntaxe avec Stanza. Certaines graphies sans accent du XVIII^{ème} siècle ont gêné la lemmatisation, ce qui est d'autant plus regrettable que ces questions ont été précédemment résolues à la conception de la chaîne de traitement du projet PRESTO (Diwersy et al., 2017) qui en revanche ne gère que l'annotation en morphosyntaxe. L'étape de reconnaissance des lemmes semble également avoir souffert de la forte concentration d'abréviations dans *LGE* laissées telles quelles par l'outil. La position de substantifs en tête des phrases nominales par lesquelles débutent de nombreux articles constitue une dernière source d'erreur facilement détectable dans les sorties de Stanza. Les majuscules à l'initiale de ces noms communs les font passer pour des noms propres et leur typographie a été conservée dans le lemme associé au token. Ces imprécisions diminuent l'intérêt de l'étape de lemmatisation, de nombreuses variantes graphiques devant être introduites dans les requêtes textométriques de la même façon que si elles portaient directement sur la forme des mots plutôt que sur leur lemme. Des progrès restent donc à accomplir dans la qualité de ce type d'analyses, et ce malgré le nombre important d'outils très performants déjà développés. Il semble donc quelque peu illusoire d'espérer régler tous les problèmes en créant encore une nouvelle chaîne d'annotation. Trouver des moyens pour pouvoir réutiliser en partie les solutions déjà disponibles et les composer paraît plus prometteur.

Les discours disciplinaires d'autres encyclopédies

La fin de cette thèse coïncide environ avec celle du projet GEODE. Cela signifie qu'il ne reste guère de temps pour prolonger les travaux présentés ici dans le cadre du projet, mais les thématiques étudiées restent pertinentes en elles-même et il est à souhaiter que de futurs efforts soient entrepris pour caractériser les évolutions des discours géographiques, dans ces encyclopédies ou dans d'autres. En outre, si la Géographie est au centre des attentions dans cette thèse, la très large majorité des méthodes et outils utilisés ne dépendent pas de la discipline considérée.

Pour commencer, deux sous-corpus construits pour les besoins des présentes analyses pourraient constituer des objets d'étude intéressants pour de futurs travaux. Le sous-corpus Parallèle d'abord (voir la section 5.1.2 p.146) se prêterait aussi bien à des améliorations techniques qu'à une exploitation dans d'autres contextes. L'algorithme d'appariement utilisé reste pour l'instant assez naïf et il y a sans doute des gains significatifs à obtenir en définissant des méthodes de désambiguïsation plus fines pour permettre de considérer des vedettes proches mais pas rigoureusement identiques, voire pour trouver des paires parmi des groupes d'articles homonymes. La taille du sous-corpus Parallèle ne jouait pas un rôle déterminant dans l'utilisation qui en est faite dans cette étude, mais augmenter sa couverture ouvrirait la porte à des études plus quantitatives. Il pourrait aussi s'avérer utile pour étudier les variations diachroniques au sein d'autres domaines de connaissance.

Ensuite, le corpus des Communes, mis au jour presque accidentellement en évaluant la qualité des prédictions du modèle BERT sur les articles de *LGE* (voir la section 4.3.2 p.126) possède plusieurs qualités notables. Sa taille d'abord est remarquable : avec plus de 20 000 articles, il contient davantage d'articles qu'il n'y en a en tout dans la classe *Géographie* de l'*EDdA* (même plus d'une fois et demie puisqu'il y en a environ 13 000). Composé d'articles extrêmement brefs (50 tokens ou moins) et tous construits sur le modèle d'une phrase nominale dense en abréviations, sa grande régularité invite à

une extraction automatique des données.

De plus, les changements mis en évidence dans cette thèse ne concernent que les deux premières œuvres du corpus de GEODE. Il apparaît naturel de se demander ce que deviendraient les constats qui y sont faits au regard des deux autres œuvres versées au corpus de GEODE, *Universalis* et *Wikipédia*, voire dans d'autres encyclopédies encore. Les considérations diachroniques des présents travaux, qui devaient initialement inclure *Wikipédia*, ont dû se limiter au premier intervalle du corpus faute de temps. Il serait donc profitable de prolonger les méthodes utilisées dans cette thèse pour regarder si les tendances découvertes se poursuivent ou bien se modifient. Les approches algébriques et celles basées sur la théorie des graphes pour déterminer les influences entre domaines et leurs proximités lexicales devraient trouver sans difficulté des applications à d'autres encyclopédies ou d'autres domaines de connaissance et pourraient être mises à profit pour réaliser une cartographie des domaines de connaissance d'une œuvre donnée et la comparer avec les intentions déclarées par ses auteurs telles que le «Système figuré» des Lumières.

LGE, un sujet riche pour les HN

Enfin, il est difficile de travailler aussi longtemps au contact d'une œuvre aussi fascinante que *LGE* sans apercevoir des passages qui éveillent la curiosité et feraient probablement des sujets de recherche passionnants.

Les biographies occupent une part substantielle de l'œuvre mais sont à peine mentionnées dans l'Avant-Propos (La Grande Encyclopédie, T1, p.XIII). Classées dans ces travaux en *Histoire* ou bien en fonction de la discipline où la personne décrite s'est illustrée, elles restent à ce jour non répertoriées. La présence du prénom de la personne en lieu et place du désignant (voir la section 3.1.2) fournit un indice précieux pour les identifier et pourrait sans doute également permettre de discerner le genre de la personne décrite. Si ces articles traitent dans leur large majorité d'hommes célèbres, certains sont consacrés à des femmes comme l'entrée sur LOVELING évoquée à la même section (p.56). Un sous-corpus des biographies classées par genre permettrait à n'en pas douter des analyses fructueuses : il suffit de se saisir des premiers mots de la biographie de LOVELING pour trouver un motif sûrement fécond. On trouve beaucoup de biographies de «femme de lettres», mais aussi quelques occurrences où une femme est décrite par rapport à son époux ou son père. Ainsi, les premiers mots sur lady FRANKLIN (La Grande Encyclopédie, T18, p.66) la présentent comme la «seconde femme de sir John Franklin», et sa vie n'est racontée dans le reste de l'article qu'en relation à celle de son mari. Pour certains couples célèbres, les deux personnes ont leur entrées. Alors que l'article sur ABAILARD (La Grande Encyclopédie, T1, p.14) ne mentionne Héloïse qu'à la 22ème ligne et s'étend sur 2 pages, celui sur HÉLOISE (La Grande Encyclopédie, T19, p.1043) la présente d'emblée comme «amante d'Abailard» (avec un renvoi vers l'article précédent qui, lui, n'en possède pas) et ne consiste qu'en un bref paragraphe de 6 lignes. Mais cette asymétrie en faveur des hommes n'est pas systématique comme le démontre le traitement du couple HOLLAND (La Grande Encyclopédie, T20, p.192). Lady Holland est certes présentée d'emblée comme «femme du suivant» (renvoi implicite à l'article sur son mari) alors que son mari est d'abord un «célèbre homme politique» mais son entrée est plus brève et renvoie explicitement vers celle de son épouse. La place des femmes dans *LGE* en général et plus particulièrement cette opposition «femme de lettres / femme de» mérite certainement d'être examinée en détail et de manière quantitative, ce que permet la version numérique de l'œuvre obtenue à l'issue de cette thèse.

Un autre thème ressort nettement des pages de *LGE* et en particulier des entrées géographiques. La construction du second empire colonial français impulsée en 1830 par Charles X en Algérie est déjà bien avancée sous la III^{ème} république au moment de la publication de *LGE* : la présence française est établie en Afrique subsaharienne (en Côte d'Ivoire depuis 1843 ou au Gabon depuis 1839 entre autres), en Asie (Viêt Nam depuis 1858, première pierre de ce qui deviendra l'Indochine en 1887

–2 ans après le début de la parution de *LGE*) et dans le Pacifique (dont Kanaky à partir de 1853). L'exploitation des territoires, souvents décrits en termes des ressources qu'ils peuvent fournir dans des sections réservées à la démographie et à l'économie des articles de *Géographie* semble à première vue concerner autant des pays d'Europe (dont la France elle-même) que des colonies mais la nature des statistiques rapportées pourrait différer suivant les lieux. En revanche, la description des populations colonisées se distingue de manière bien plus nette et révèle une vision du monde basée sur une notion de «races» hiérarchisées. Les nombres d'habitants sont souvent divisés suivant ce critère, par exemple aux articles PHILIPPEVILLE (La Grande Encyclopédie, T26, p.676) –ancien nom de la ville de Skikda– et ZANZIBAR (La Grande Encyclopédie, T31, p.1305). D'autres articles comme CANAQUES (La Grande Encyclopédie, T8, p.1195) ou BAMBARA (La Grande Encyclopédie, T5, p.192) prennent résolument le parti de caractériser les populations par des qualités physiques ou psychologiques supposées inhérentes à tous les individus d'un groupe ethnique. L'adhésion implicite aux forces coloniales transparaît dans l'emploi de la première personne du pluriel, par exemple «notre protectorat» à l'article BAMAKOU (La Grande Encyclopédie, T5, p.192) alors que le recours à cette personne est très rare dans *LGE* par rapport à l'*EDdA*. Il s'agirait indubitablement d'un travail pénible du fait de la violence de certains textes, mais une étude lexicale précise basée sur des considérations textométriques permettrait sans nul doute de rendre compte des représentations et des théories de la fin du XIX^{ème} siècle et de montrer leur articulation avec l'esprit républicain de l'époque qui transparaît par ailleurs dans le reste de l'œuvre.

Il faut donc espérer que d'autres s'empareront des myriades de questions qui subsistent dans les pages de *LGE*. Ces recherches rendront assurément les efforts déployés au centuple, en émerveillement et en découvertes, comme elles l'ont fait pour cette thèse.

Table des figures

1.1 Échantillonnage d'un processus sinusoïdal par un ensemble de points répartis aléatoirement . . . 12
1.2 Exemple schématique de rebrassage des domaines de connaissance . . . 13

2.1 Quelques membres éminents de la famille SGML . . . 18
2.2 Des balises SGML . . . 19
2.3 Un court extrait d'un document XML-ALTO . . . 20
2.4 La triple relation qu'entretiennent le TAL et les œuvres lexicographiques . . . 25
2.5 Deux analyses syntaxiques d'une phrase de l'article SYMPATHIQUE, *LGE*, T30, p.759 35

3.1 Les différents constituants possibles d'un article d'encyclopédie . . . 48
3.2 Les différents éléments péritextes en haut et en cours de pages . . . 50
3.3 Les différents éléments péritextes en bas de page . . . 50
3.4 Le «Systême figuré des connoissances humaines», une taxonomie au cœur de l'*EDdA* 52
3.5 Article A A A dans l'*EDdA*, T1, p.5 . . . 53
3.6 Tableau de répartition du volume de texte prévisionnel de *LGE* entre les différents domaines de connaissance . . . 55
3.7 Article LOVELING dans *LGE*, T22, p.699 . . . 56
3.8 Extrait de l'Article CHEMIN dans *LGE*, T10, p.1025 . . . 57
3.9 Trois entrées de l'*EDdA* sur la vedette AIR . . . 57
3.10 Le contexte XML est dupliqué pour chaque article . . . 60
3.11 Dernière page du premier tome de *LGE* . . . 61
3.12 Traduction sous forme d'arbre XML de l'existence d'une arête `a` → `b` sur le graphe . 63
3.13 Le sous-graphe du module *dictionaries* . . . 65
3.14 Article EUROPE dans *LGE*, T16 s'étendant de la p.782 à la p.846 . . . 66
3.15 Article BALAGRUS dans *LGE*, T5, p.48 . . . 66
3.16 Énumération dans l'article EUROPE, *LGE*, T16, p.783 . . . 67
3.17 Extrait du début de la section III. Jurisprudence de l'article CONSTRUCTION dans *LGE*, T12, p.734 . . . 68
3.18 Début de l'article SANJO SANETOMI dans *LGE*, T29, p.441 . . . 70
3.19 Extrait de l'article ACTION dans *LGE*, T1, p.495 . . . 70
3.20 Article CATHÈTE dans *LGE*, T9, p.849 . . . 71
3.21 L'élément conteneur `div` pour l'article CATHÈTE . . . 71
3.22 L'encodage de la vedette et du désignant de l'article CATHÈTE . . . 72
3.23 La structure vide permettant de représenter le seul sens du mot CATHÈTE . . . 72
3.24 Un encodage complet de l'article CATHÈTE . . . 72
3.25 Figure de l'article BOUMERANG dans *LGE*, T7, p.704 . . . 73
3.26 Encodage de la figure dans l'article BOUMERANG et de sa légende . . . 73
3.27 Article GELOCUS dans *LGE*, T18, p.699 . . . 73
3.28 Encodage des renvois dans l'article GELOCUS . . . 74

3.29 Extrait de l'article ALCALA-DE-HÉNARÈS dans *LGE*, T1, p.1200 74
3.30 Encodage d'un début de page dans l'article ALCALA-DE-HÉNARÈS 74
3.31 Encodage actuel de l'article CATHÈTE produit par `soprano` 76
3.32 Article ALPAM dans *LGE*, T2, p.437 79
3.33 Deux entrées de «Géographie moderne» dans l'*EDdA* 83
3.34 Distribution des nombres d'articles par *domaine regroupé* 83

4.1 F-scores obtenus par BERT Multilingual et CamemBERT sur chaque classe 96
4.2 F-mesures obtenues par *Naive Bayes + TF-IDF* sur chaque classe avec trois échantillonnages différents 97
4.3 F-mesures obtenues par *SGD* sur chaque classe avec trois vectorisations différentes et sans échantillonnage. 97
4.4 F-mesures par classe obtenus sur le jeu de test par les combinaisons *SGD + TF-IDF*, *BiLSTM + FastText* and *BERT* Multilingue. 99
4.5 Matrice de confusion matrix pour la combinaison *SGD+TF-IDF* sur le jeu de test. . . 99
4.6 Distributions normalisées des nombres de clusters par classe. 101
4.7 Nombre d'articles de chaque classe par cluster 101
4.8 Les paramètres en jeu lors de la classification automatique d'un article 103
4.9 Articles étiquetés *Géographie* par l'ENCCRE mais pas par le modèle (faux négatifs) 105
4.10 Article ROCHER, l'*EDdA*, T14, p.313 106
4.11 Articles étiquetés *Géographie* par le modèle mais pas par l'ENCCRE (faux positifs) . 107
4.12 Représentation schématique de l'équivalence entre matrice et graphe 108
4.13 Matrice des Plus Proches Voisins pour la méthode SGD+TF-IDF sur le jeu de test. . 110
4.14 Graphe des Plus Proches Voisins pour la méthode SGD+TF-IDF sur le jeu de test. . 111
4.15 Similarités lexicale des 100 2-grammes les plus fréquents 115
4.16 Graphe des Plus Proches Voisins pour la similarité *cardinale* mesurée sur les 100 2-grammes les plus fréquents de chaque classe. 118
4.17 Des itérées de la matrice de confusion du modèle, C 122
4.18 La distribution par classe de mesures de centralité du graphe du modèle 122
4.19 Matrices de confusion des modèles entraînés sur les superdomaines 124
4.20 Nombre d'articles par domaine dans l'*EDdA* 127
4.21 Article BOULAINCOURT, *LGE*, T7, p.655 127
4.22 Article HEYRIEUX, *LGE*, T20, p.60 128
4.23 Nombre d'articles par domaine dans *LGE* 131

5.1 Nombre de mots par domaine dans l'*EDdA* 136
5.2 Distribution des nombres de mots par article au sein des différents domaines de l'*EDdA* 137
5.3 Nombre de mots par domaine dans *LGE* 138
5.4 Distribution des nombres de mots par article au sein des différents domaines de *LGE* 139
5.5 Annotations de l'article Strongoli (L'Encyclopédie, T15, p.547) 142
5.6 Densité d'entités dans l'*EDdA* par domaine et par type 143
5.7 Densité des ENE dans *LGE* par domaine et par type d'ENE 145
5.8 Les articles sur le fleuve ADIGE, un exemple de paire issue du sous-corpus Parallèle 147
5.9 Nombre total de comparaisons requis dans le pire des cas en fonction de la taille de la fenêtre utilisée 149
5.10 Répartition des nombres d'articles par domaines au sein du sous-corpus Parallèle . 151
5.11 Évolution des domaines des articles dans le sous-corpus Parallèle 152
5.12 Changements de domaines d'articles impliquant la *Géographie* dans le corpus Parallèle, de l'*EDdA* (à gauche) à *LGE* (à droite) 153
5.13 Les articles sur l'île OGYGIE dans le sous-corpus Parallèle 154
5.14 Les articles VILLEPREUX dans le sous-corpus Parallèle 155

5.15 Les articles PURBECK dans le sous-corpus Parallèle 155
5.16 Article ZSCHIMMER dans *LGE*, T31, p.1338 157
5.17 Les 10 cooccurrents syntaxiques principaux du lemme «ville» annoté comme NOUN (nom commun) 159
5.18 Les 10 cooccurrents syntaxiques principaux du motif formé d'une relation syntaxique quelconque entre le lemme «ce» annoté comme DET et le lemme «ville» annoté comme NOUN 160
5.19 Représentation sous forme d'arbre de syntaxe en dépendance du motif défini par la requête 5.3 160
5.20 Les 10 cooccurrents syntaxiques principaux de «dans» (ADP), «ce» (DET), et «ville» (NOUN) 161
5.21 Spécificités des lemmes «naître» et «mourir» par superdomaine dans l'*EDdA* 162
5.22 Spécificités des lemmes «naître» et «mourir» par superdomaine dans *LGE* 162
5.23 Spécificités des lemmes «naître» et «mourir» par auteur, en *Géographie* et hors de ce domaine 163
5.24 Spécificités des lemmes «naître» et «mourir» par auteur à l'exception de Jaucourt, en *Géographie* et hors de ce domaine 163
5.25 Analyse Factorielle des Correspondances pour les articles de systèmes philosophiques dans l'*EDdA*, y compris NEWTONIANISME 167
5.26 Article "Ferté-Milon" *EDdA*, T6 168
5.27 Visualisation d'articles de l'*EDdA* discutés dans la section suivant les deux critères identifiés 170

Liste des tableaux

3.1 Clefs primaires des entrées COMMERCY et EVIAN dans l'*EDdA* 85
3.2 Métadonnées de navigation pour les entrées COMMERCY et EVIAN dans l'*EDdA* . 85
3.3 Quatre notions différentes de la reproductibilité 88

4.1 Détail du nombre d'articles dans le corpus entier, le corpus prétraité et les jeux d'entraînement et de test. 94
4.2 F-mesures moyennes pour les différents modèles entraînés sur un échantillon de 500 articles maximum par classe (1), 1500 maximum par classes (2) et tous les articles disponibles de chaque classe (3) appliqués sur le jeu de test. 95
4.3 F-mesures par classe obtenues sur le jeu de test par les combinaisons *SGD + TF-IDF* (1), *BiLSTM + FastText* (2) et *BERT* Multilingue (3). 98
4.4 Part de la classe la plus représentée dans chaque cluster. 102
4.5 Les 3-grammes les plus fréquents communs aux classes *Histoire* et *Religion* avec leurs comptes d'occurrences dans chacune 116
4.6 Coût en énergie des différentes phases de l'entraînement des modèles *SGD+TF-IDF* et *BERT* 125
4.7 Scores de précision par domaine obtenus par le modèle *BERT* pour la classification en superdomaines sur l'échantillon de *LGE* 130

5.1 Longueur moyenne des mots par domaine pour chacune des deux encyclopédies étudiées 140
5.2 Coordonnées d'articles de philosophie de l'*EDdA* contenant des éléments biographiques 164

Glossaire

AA *Apprentissage Automatique* est un domaine scientifique visant à programmer un ordinateur pour découvrir automatiquement les paramètres clefs dans la résolution d'un problème, plutôt que de traduire dans son langage des règles expertes issues de considérations théoriques préalables sur le problème à résoudre. C'est un sous-domaine de l'IA.

AFC *Analyse Factorielle des Correspondances* est une technique de visualisation développée par le mathématicien Jean-Paul Benzécri qui permet d'observer des mots et les textes dont ils sont issus dans un même espace à deux dimensions pour observer les proximités thématiques entre eux. Cette représentation est obtenue à l'aide d'une décomposition matricielle de l'écart entre la distribution des mots dans les textes observée empiriquement et une distribution homogène.

ALR *Arbres Lexico-syntaxiques Récurrents* désigne une technique d'analyse statistique du texte qui procède par itérations en sélectionnant à chaque étape un cooccurrent syntaxique important d'une expression pour extraire progressivement une tournure caractéristique.

AP *Apprentissage Profond* est un sous-domaine de l'AA comprenant toutes les architectures de réseaux de neurones dont le grand-nombre de couches permet l'émergence de comportements complexes utiles pour résoudre des problèmes tels que le traitement du signal (en particulier l'analyse d'images) et de la langue naturelle (pour laquelle les *Transformer* dominent l'état de l'art).

ARTFL l'American and French Research on the Treasury of the French Language («Recherche américaine et française sur les trésors de la langue française») est une collaboration du laboratoire ATILF, du CNRS et de l'Université de Chicago. Le projet donne accès à un vaste corpus de textes anciens en français. https://artfl-project.uchicago.edu/.

CNN *Convolutional Neural Network* est une architecture de réseaux de neurones dans laquelle l'information se déplace dans une seule direction, des entrées vers les sorties, et avec une interconnexion peu dense entre les premières couches qui privilégie la localité des intéractions pour aggréger progressivement l'information. Cette caractéristique les rend particulièrement efficaces pour traiter des images.

CPL *Caractères Par Ligne*, désigne en typographie le nombre de caractère en fonte de largeur fixe que peut afficher un dispositif mécanique ou électronique comme une machine à écrire ou un télétype. Ce nombre, dont la valeur est généralement voisine de 80, joue un rôle clef pour comprendre les choix de mise en forme du texte sur des supports numériques.

CQL *Corpus Query Langage*, est un langage de requête développé à l'université de Stuttgart et implémenté dans le Corpus Query Processor (CQP) par l'outil *CWB* (Corpus Workbench, https://cwb.sourceforge.io/). Il permet de représenter des contraintes sur du texte lemmatisé et annoté en morphosyntaxe en définissant des critères sur différentes dimensions des tokens (forme, lemme, partie de discours) qui peuvent être ensuite combinées logiquement et séquentiellement de manière analogue à ce que permettent les expressions régulières.

EN *Entité Nommée* (*Named Entity* en anglais) désigne une unité d'information dans un texte, souvent un mot ou un groupe de quelques mots qui renvoient à un objet unique dans le

contexte. Plus large que la notion de nom propre, il englobe par exemple les noms de personne, d'organisation ou d'endroit, les dates ou les grandeurs pourvues d'une unité.

ENCCRE l'*Édition Numérique Collaborative et CRitique de l'Encyclopédie*, projet français étudiant l'*EDdA* http://enccre.academie-sciences.fr/encyclopedie/.

ENE *Entité Nommée Étendue* désigne un groupe de mots structuré en une construction qui dépasse le cadre des EN pour inclure les éléments présents autour : principalement des noms communs et des relations. Les ENE se bâtissent récursivement, pouvant inclure d'autres ENE, pour constituer de petits arbres autours des entités nommées.

Ainsi, «le kiosque dans le parc Sainte-Marie» peut s'analyser comme une ENE de lieu, constituée d'un nom commun de lieu qui lui donne son type (c'est un kiosque), d'une relation spatiale «dans» (traduisant l'inclusion), et d'une ENE : «le parc Sainte-Marie». À son tour, celle-ci se décompose en un nom commun de lieu (c'est un parc), et un nom propre, «Sainte-Marie», qui sans l'apport des ENE aurait probablement été la seule EN simple retenue pour toute l'expression. Il est intéressant de voir que le nom propre apporte pourtant peu d'information en lui-même, et qu'il n'est possible de percevoir sa dimension spatiale qu'avec l'ajout du nom commun «parc» (il existe peut-être par exemple une «communauté Sainte-Marie» qui pourrait être une organisation et pas un lieu) ce qui souligne l'utilité des ENE pour la désambiguïsation et les tâches de compréhension automatique.

HN *Humanités Numériques* ou *Digital Humanities* en anglais, sont la rencontre d'approches et de méthodes issues des technologies de l'information et des problématiques traditionnellement réservées aux sciences dites «humaines et sociales», aux arts et aux lettres. Le terme intègre le souhait que cette rencontre soit, plus qu'une simple application technique, un véritable dialogue entre disciplines.

IA *Intelligence Artificielle*, terme parapluie traduisant davantage un projet qu'une réalité mesurable objectivement, est l'ensemble des techniques permettant à un ordinateur de produire un comportement assez complexe pour résoudre des tâches traditionnellement liées aux capacités intrinsèques de l'espèce humaine : langage, reconnaissance d'image, généralisation et adaptation à de nouvelles situations.

LDA *Latent Dirichlet Allocation*, procédé de réduction de dimension permettant de mettre en évidence des classes de ressemblances présentes dans les données d'entrée. C'est une technique de Topic Modeling.

Naive Bayes ou *classification bayésienne «naïve»* désigne une approche probabiliste des tâches de classification consistant à apprendre une estimation de la probabilité qu'un vecteur d'entrée donnée reçoive une certaine classe, en opérant la simplification consistant à supposer que les coordonnées du vecteur sont indépendantes entre elles. Le classifieur obtenu associe ensuite à chaque vecteur d'entrée la classe ayant la plus haute probabilité.

OCR *Optical Character Recognition*, reconnaissance optique de caractères, est le procédé par lequel un logiciel extrait du texte, c'est-à-dire une suite de caractères compréhensibles par la machine et traitables ensuite par des moyens automatiques, à partir d'une image.

OLR *Optical Layout Recognition*, reconnaissance optique de la disposition de la page, est le procédé par lequel un logiciel découpe une page en blocs géométriques, afin de pouvoir distinguer les parties de texte, les images et les données tabulaires, permettant à terme d'inférer un ordre de lecture. Cette étape est préalable à l'OCR, qu'elle «guide» en quelque sorte vers les zones de la page où il est susceptible de trouver du contenu à reconnaître. Le problème fondamental sous-jacent que tente de résoudre un système d'OLR est celui du passage de la page – objet bidimensionnel qu'un humain peut parcourir à loisir dans le sens qu'il souhaite – au fichier – abstraction intrinsèquement linéaire des systèmes informatiques et dont toutes portions distinctes peuvent être ordonnées, l'une étant *avant* l'autre.

partition est un concept fondamental pour les linguistiques de corpus. Il consiste à regrouper les

textes d'un corpus en fonction de critères comme la valeur d'une métadonnée (typiquement année, auteur, genre littéraire) en vue de mettre en évidence des différences statistiquement significatives entre les groupes dans le cadre d'une étude contrastive. Les groupes doivent être disjoints (aucun article n'est dans deux partitions à la fois) et leur union égale au corpus entier (aucun article n'est dans aucun groupe). Le terme est utilisé aussi bien pour désigner une telle division (conjointement avec «partitionnement» dans ce cas) qu'un des groupes obtenus.

POS *Part Of Speech*, couramment abrévié en POS correspond au français «Partie de Discours» qui est une autre façon de désigner la catégorie morphosyntaxique des mots d'une phrase. C'est un concept essentiel en textométrie, tant pour formuler des requêtes que dans l'agrégation des résultats.

PPV le problème des *K Plus Proche Voisin* (*KNN* pour *K Nearest Neighbors* en anglais) est un problème général d'IA consistant à trouver les points les plus proches d'un point donné dans un espace. D'une grande complexité dans le cas général, de nombreuses techniques permettent d'accélérer sa résolution, qui trouvent des applications des réseaux à la vision par ordinateur. Dans cette thèse, la notion de *PPV* est utilisée pour simplifier des matrices afin de tracer des graphes exploitables visuellement.

RNN *Recurrent Neural Network* est une architecture de réseaux de neurones présentant une boucle de rétroaction entre ses entrées et ses sorties ce qui lui permet de conserver une mémoire des premiers éléments rencontrés et la rend relativement indépendante du paramètre temporel. Cette caractéristique la rend particulièrement efficace à traiter des entrées sérielles.

SDDL *Système Dynamique Discret Linéaire* : un système est *dynamique* si une fonction décrit son évolution dans le temps. Il est *discret* s'il évolue par états successifs qui peuvent être indicés par des nombres entiers (par opposition à *continu* où le temps est représenté par une variable réelle). La résolution d'un système continu impose parfois de le discrétiser pour le simplifier et pouvoir approcher son évolution en la modélisant par des petites étapes pendant lesquels les paramètres sont fixés. Enfin, il est *linéaire* si sa fonction d'évolution l'est.

En un sens, un *SDDL* est donc la généralisation d'une suite géométrique ($u_{n+1} = r \times u_n$) à un système qui ne peut pas être représenté par un seul nombre mais nécessite d'en observer plusieurs simultanément, dans un vecteur U_n dont chaque étape s'obtient par un produit matriciel (opération linéaire) $U_{n+1} = M \cdot U_n$.

TAL *Traitement Automatique des Langues*, est la partie de l'IA qui s'intéresse au langage. Il consiste à utiliser un ordinateur pour effectuer des tâches sur des textes écrits en langues naturelles comme par exemple leur traduction dans une autre langue, l'écriture de résumé ou, d'une manière générale l'extraction d'information sémantique comme les chaînes de coréférence, les entités nommées ou les émotions exprimées par un texte.

TEI la *Text Encoding Initiative*, est un consortium fondé en 1994 qui développe et maintient un ensemble ouvert de recommandations pour encoder des textes numériques. Ces recommandations définissent un standard, le XML-TEI, qui est un dialecte du méta-langage de balises XML.

Topic Modeling en français «modélisation de thème/sujet» est un ensemble de techniques d'apprentissage non supervisé qui permettent de faire émerger des thématiques d'un corpus en opérant des regroupements entre les différents textes qui le composent.

TSV *Tab-Separated Values*, «valeurs séparées par des tab(ulation)s» est un format de données tabulaires (où chaque donnée trouve sa place à l'intersection d'une colonne et d'une ligne). Comme les autres langages SSV (*Something-Separated Values*), il possède un caractère d'échappement pour pouvoir inclure le séparateur (ou ce caractère d'échappement lui-même) dans le contenu de n'importe quelle cellule.

En pratique tant que les données ne comportent pas le caractère tabulation (ce qui est le cas pour des métadonnées brèves et souvent numériques) le décodage de ce format est trivial,

nécessitant seulement de découper les lignes suivant les tabulations. Il faut en revanche se méfier de l'opération inverse, consistant à produire des fichiers en séparant seulement les cellules par des tabulations sans prendre en compte le caractère d'échappement (guillemets doubles d'après la RFC 4180) : si une cellule contient ce caractère, alors le fichier obtenu ne serait plus un TSV valide et ne serait pas lu correctement par les outils pour ce format (typiquement, des tableurs).

UD *Universal Dependencies*, sont une convention d'annotation grammaticale comprenant des jeux d'étiquettes à plusieurs niveaux (morphosyntaxe, morphologie et syntaxe en dépendance) conçus pour être commun à un grand nombre des langues humaines les plus étudiées, ce qui permet notamment de pouvoir travailler sur des corpus multilingues (https://universaldependencies.org/).

Bibliographie

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a treebank for french. In A. Abeillé (Ed.), *Treebanks : Building and using parsed corpora* (pp. 165–187). Springer Netherlands. https://doi.org/10.1007/978-94-010-0201-1_10

Adriyanto, T., Ramadhani, R. A., Helilintar, R., & Ristyawan, A. (2022). Classification of dog and cat images using the CNN method. *Ilkom Jurnal Ilmiah*, *14*(3), 203–208.

Aghion, P., & Bouverot, A. (2024). *IA : Notre ambition pour la france*. Commission de l'Intelligence Artificielle. https://www.vie-publique.fr/rapport/293444-ia-notre-ambition-pour-la-france?xtor=RSS-719

Allen, T., Gladstone, C., & Whaling, R. (2013). PhiloLogic4 : An Abstract TEI Query System. *Journal of the Text Encoding Initiative*, *Issue 5*. https://doi.org/10.4000/jtei.817

ALTO : Technical Metadata for Layout and Text Objects (Standards, Library of Congress) (Version 4.4). (2023). https://www.loc.gov/standards/alto/

Barron, A. T. J., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, *115*(18), 4607–4612. https://doi.org/10.1073/pnas.1717729115

Basnage de Beauval, H. (1702). *Dictionnaire universel, contenant généralement tous les mots français tant vieux que modernes et les termes de toutes les sciences et des arts* (pp. 1–1116). Arnoud & Reinier Leers. http://catalogue.bnf.fr/ark:/12148/cb304709589

Bauman, S. (2019). A TEI customization for writing TEI customizations. *Journal of the Text Encoding Initiative*, *12*(Issue 12).

Beaudouin, V. (2016). Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données. *JADT 2016*, 17–27. https://hal.science/hal-01376938

Benazha, H., Ayache, S., & Artières, T. (2024, June). Measuring Hallucination in Disentangled Representations. *International Joint Conference on Neural Networks (IJCNN)*. https://hal.science/hal-04532666

Bender, E. (2019). The #BenderRule : On naming the languages we study and why it matters. *The Gradient*.

Benveniste, É. (1966). *Problèmes de linguistique générale . [i]*. Gallimard.

Benzécri, J.-P., Bellier, L., Bénier, B., & Benzécri, F. (1973). *L'Analyse des données... , 2 : L'Analyse des correspondances*. Dunod.

Bewley, A., & Upcroft, B. (2013). Advantages of exploiting projection structure for segmenting dense 3D point clouds. *IEEE International Conference on Robotics and Automation*.

https://api.semanticscholar.org/CorpusID:3691407

Blair, A. (2010). *Too much to know : Managing scholarly information before the modern age*. Yale University Press.

Blanchard, G., & Olsen, M. (2002). Le système de renvois dans l'Encyclopédie : Une cartographie des structures de connaissances au XVIIIe siècle. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *31-32*, 45. https://doi.org/10.4000/rde.122

Blumenthal, P. (2017). D'une encyclopédie à l'autre. L'ascension des locutions prépositionnelles : *Langages*, *N° 206*(2), 123–138. https://doi.org/10.3917/lang.206.0123

Bohbot, H., Frontini, F., & Faucher, A. (2019, July). Le projet Nénufar, nouvelle édition numérique du Petit Larousse illustré (1906-1948). *XXIXe congrès international de linguistique et de philologie romanes*. https://hal.science/hal-03262248

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bonfante, G., Guillaume, B., & Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing* (Vol. 1, p. 272). ISTE Wiley. https://inria.hal.science/hal-01814386

Boussuge, E., & Launay, F. (2019). La description des arts dans l'Encyclopédie. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *54*, 181–253. https://doi.org/10.4000/rde.6270

Bowker, G. C., & Star, S. L. (2000). *Sorting things out : Classification and its consequences*. MIT Press.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergau, F. (Eds.). (2008). *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. https://www.w3.org/TR/REC-xml/

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brenon, A. (2024). *Chapter 3 encoding the specificities of encyclopedias* (pp. 36–62). Brill. https://doi.org/10.1163/9789004702660_004

Brenon, A., Moncla, L., & McDonough, K. (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. *Data and Knowledge Engineering*, 102098. https://doi.org/10.1016/j.datak.2022.102098

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, 152–155. https://doi.org/10.3115/974499.974526

Brugidou, M., & Labbé, D. (2000). *Le discours syndical français contemporain (CFDT, CGT, FO en 1996-98)* (pp. 150 p.). Cerat. https://shs.hal.science/halshs-00493321

Brunet, E. (2009). Muller le lexicomaître. In *Mélanges offerts à Charles Muller pour son centième anniversaire* (pp. 99–119). Conseil International de la langue française. https://hal.univ-cotedazur.fr/hal-01362724

Brunet, E. (2014). La lexicométrie française : Naissance, évolution et perspectives. *Revue de l'Université de Moncton*, *45*(1-2), 13–33. https ://doi.org/https://doi.org/10.7202/1038899ar

Brunet, É. (2002, April). Hugométrie. *Colloque Victor Hugo*. https://hal.science/hal-01562323

Brunet, É., & Luong, X. (1989). Computer processing and quantitative text analysis : HYPERBASE, an interactive software for large corpora. In I. E. Diday (Ed.), *Data analysis, learning symbolic and numeric knowledge* (pp. 207–214). Nova Science Publishers. https://hal.science/hal-01482906

Buchholz, S., & Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In L. Màrquez & D. Klein (Eds.), *Proceedings of the tenth conference on computational natural language learning (CoNLL-x)* (pp. 149–164). Association for Computational Linguistics. https://aclanthology.org/W06-2920

Candito, M., & Seddah, D. (2012, June). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*. https://inria.hal.science/hal-00698938

Cernuschi, A. (2006). Des désignants combinés ou vers une dimension opératoire des articles de l' Encyclopédie. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *40-41*, 93–106. https://doi.org/10.4000/rde.335

Chagué, A., Clérice, T., & Romary, L. (2021, November). HTR-United : Mutualisons la vérité de terrain ! *DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*. https://hal.science/hal-03398740

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews : Comparative Analysis. *Journal of Medical Internet Research*, *26*, e53164. https://doi.org/10.2196/53164

Chevalier, J.-P. (1997). Quatre pôles dans le champ de la géographie ? *Cybergeo : European Journal of Geography*. https://doi.org/10.4000/cybergeo.6498

Chevalier, J.-P. (2013). La géographie scolaire et l'idée de nation en France au XIXe siècle. In B. Falaize, C. Heimberg, & O. Loubes (Eds.), *L'école et la nation : Actes du séminaire scientifique international. Lyon, Barcelone, Paris, 2010* (pp. 129–136). ENS Éditions. https://books.openedition.org/enseditions/2356

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113–124. https://doi.org/10.1109/TIT.1956.1056813

Chomsky, N. (1965). *Syntactic structures...* Mouton & C°.

Christianini, N., & Shawe-Taylor, J. (2000). Support vector machines and other kernel-based learning methods. *Cambridge UP*.

Church, K. (2011). A Pendulum Swung Too Far. *Linguistic Issues in Language Technology*, *6*. https://doi.org/10.33011/lilt.v6i.1245

Clark, J., & Murata, M. (Eds.). (2001). *RELAX NG specification*. OASIS Committee Specification. https://www.oasis-open.org/committees/relax-ng/spec-20011203.html

Cori, M., David, S., & Léon, J. (2008). Présentation : éléments de réflexion sur la place des corpus en linguistique. *Langages*, *171*(3), 5–11. https://doi.org/10.3917/lang.171.0005

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, *134*(3), 321. https://doi.org/10.2307/2344237

Courtès, L. (2021). Reproduire les environnements logiciels : Un maillon incontournable de la recherche reproductible. *1024*, *18*, 15–22. https://doi.org/10.48556/SIF.1024.18.15

D'Antuono, G. (2018). Démocratie et éducation du peuple dans le discours de Diderot et ses héritages en Italie (XVIII e -XIX e siècles): *Lumières*, *N° 32*(2), 95–111. https://doi.org/10.3917/lumi.032.0095

Dang, Q. (2020). *Brouillard de pollution en Chine. Analyse sémantique différentielle de corpus ins-*

titutionnels, médiatiques et de microblogues (Theses No. 2020INAL0009, Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O'). https://theses.hal.science/tel-03019809

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Di Donato, F., Monachini, M., Eskevich, M., Pohle, S., Moranville, Y., & Dumouchel, S. (2020). Social sciences and humanities pathway towards the European open science cloud. In D. Broeder, M. Eskevich, & M. Monachini (Eds.), *Proceedings of the workshop about language resources for the SSH cloud* (pp. 5–9). European Language Resources Association. https://aclanthology.org/2020.lr4sshoc-1.2/

Dijkstra, Edsger. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*, 269–271. https://ir.cwi.nl/pub/9256/9256D.pdf

Diot-Parvaz, B., & Gianola, L. (2020, June). Hétérogénéité des genres textuels produits par la justice : étude textométrique. *Journées d'Analyse Statistique des Données Textuelles (JADT 2020)*. https://hal.science/hal-03667766

Diwersy, S., Falaise, A., Lay, M.-H., & Souvay, G. (2017). Ressources et méthodes pour l'analyse diachronique : *Langages*, *N° 206*(2), 21–44. https://doi.org/10.3917/lang.206.0021

Doig, K. (1990). Notices sur les auteurs des quatre volumes de « discours » du supplément à l'encyclopédie. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *9*(1), 157–170. https://www.persee.fr/doc/rde_0769-0886_1990_num_9_1_1082

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, *29*(2), 103–130.

Doucet, A., Gasteiner, M., Granroth-Wilding, M., Kaiser, M., Kaukonen, M., Labahn, R., Moreux, J.-P., Muehlberger, G., Pfanzelter, E., Thérenty, M.-È., Toivonen, H., & Tolonen, M. (2020, July). NewsEye : A digital investigator for historical newspapers. *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020*. https://hal.science/hal-03029072

Dyer-Witheford, N., Kjøsen, A. M., & Steinhoff, J. (2019). *Inhuman power : Artificial intelligence and the future of capitalism*. Pluto Press. http://www.jstor.org/stable/j.ctvj4sxc6

Ehrmann, M., Nouvel, D., & Rosset, S. (2016). Named entity resources - overview and outlook. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 3349–3356). European Language Resources Association (ELRA). https://aclanthology.org/L16-1534

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Fellbaum, C. (Ed.). (2007). *Idioms and collocations : Corpus-based linguistic and lexicographic studies*. Continuum.

Ferenczi, T. (2017). Le chevalier de Jaucourt, un combattant des Lumières : *Le Philosophoire*, *n° 47*(1), 77–133. https://doi.org/10.3917/phoir.047.0077

Firth, J. R. (1957). *A synopsis of linguistic theory, 1930-1955*. https://api.semanticscholar.org/CorpusID:

208093066

Foucault, M. (1970). *The order of things : An archaeology of the human sciences*. Tavistock.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software*, *3*(3), 209–226. https://doi.org/10.1145/355744.355745

Fukushima, K. (1980). Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Fürnkranz, J. (1998). *A study using n-gram features for text categorization*. https://api.semanticscholar.org/CorpusID:14854253

Gaio, M., & Moncla, L. (2017, March). Extended Named Entity Recognition Using Finite-State Transducers : An Application To Place Names. *The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2017)*. https://hal.science/hal-01492994

Galleron, I., & Williams, G. C. (2022). Tenir la promesse du Dictionnaire universel : L'esprit encyclopédique d'Henri Basnage de Beauval. *Langue Francaise*, *214*(2), 27–42. https://www.cairn.info/revue-langue-francaise-2022-2-page-27.htm

Gao, S., Sperberg-McQueen, C. M., Thompson, H. S., Mendelsohn, N., Beech, D., & Maloney, M. (Eds.). (2012). *W3C XML schema definition language (XSD) 1.1 part 1 : structures*. https://www.w3.org/TR/xmlschema11-1/

García, M., Maldonado, S., & Vairetti, C. (2021). Efficient n-gram construction for text categorization using feature selection techniques. *Intelligent Data Analysis*, *25*(3), 509–525. https://doi.org/10.3233/IDA-205154

Gavari Bami, H., Moharamkhani, E., Zadmehr, B., Najafpoor, V., & Shokouhifar, M. (2022). Detection of zero-day attacks in computer networks using combined classification. *Concurrency and Computation : Practice and Experience*, *34*(27), e7312. https ://doi.org/https://doi.org/10.1002/cpe.7312

Grabus, S., Greenberg, J., Logan, P., & Boone, J. (2019). Representing aboutness : Automatically indexing 19th-century encyclopedia britannica entries. *NASKO*, *7*(1), 138–148. https://doi.org/10.7152/nasko.v7i1.15635

Grobol, L., & Crabbé, B. (2021). Analyse en dépendances du français avec des plongements contextualisés. In P. Denis, N. Grabar, A. Fraisse, R. Cardon, B. Jacquemin, E. Kergosien, & A. Balvet (Eds.), *Traitement Automatique des Langues Naturelles* (pp. 106–114). ATALA. https://hal.science/hal-03265893

Groote, P. de. (2001). Towards abstract categorial grammars. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 252–259. https://doi.org/10.3115/1073012.1073045

Groult, M. (Ed.). (2003). *L'encyclopédie ou la création des disciplines*. CNRS Éditions.

Groult, M. (2019a). *Introduction* (pp. 7–21). Classiques Garnier. https://classiques-garnier.com/panckoucke-et-l-encyclopedie-methodique-ordre-de-matieres-et-transversalite-introduction.html

Groult, M. (2019b). *Ordre encyclopédique et ordre méthodique* (pp. 119–133). Classiques Garnier. https://classiques-garnier.com/panckoucke-et-l-encyclopedie-methodique-ordre-de-matieres-

et-transversalite-ordre-encyclopedique-et-ordre-methodique.html

Guillaume, B., Marneffe, M.-C. de, & Perrier, G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL : Traitement Automatique Des Langues*, *60*(2), 71–95. https://inria.hal.science/hal-02267418

Guillaume, B., & Perrier, G. (2015). Dependency parsing with graph rewriting. *Proceedings of the 14th International Conference on Parsing Technologies*, 30–39. https://doi.org/10.18653/v1/W15-2204

Guldi, J. (2019). Parliament's debates about infrastructure : An exercise in using dynamic topic models to synthesize historical change. *Technology and Culture*, *60*(1), 1–33. https://doi.org/10.1353/tech.2019.0000

Guo, F., Zhou, W., Wang, Z., Ju, C., Ji, S., & Lu, Q. (2023). A link prediction method based on topological nearest-neighbors similarity in directed networks. *Journal of Computational Science*, *69*, 102002. https://doi.org/10.1016/j.jocs.2023.102002

Gutehrlé, N., & Atanassova, I. (2022). Processing the structure of documents : Logical layout analysis of historical newspapers in french. *Journal of Data Mining and Digital Humanities*, *NLP4DH*(Digital humanities in...).

Habert, B. (2005). *Instruments et ressources électroniques pour le français / benoît habert*. Ophrys.

Hagen, T., Ketzan, E., Jannidis, F., & Witt, A. (2020). Twenty-two historical encyclopedias encoded in TEI : A new resource for the digital humanities. In S. DeGaetano, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 4th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature* (pp. 112–120). International Committee on Computational Linguistics. https://aclanthology.org/2020.latechclfl-1.13

Haiman, J. (1980). Dictionaries and encyclopedias. *Lingua*, *50*(4), 329–357. https://doi.org/10.1016/0024-3841(80)90089-3

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. Longman.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1 : Long papers)* (pp. 1489–1501). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141

Hardesty Doig, K. (2006). Le désignant « Histoire » dans l'Encyclopédie et dans l'Encyclopédie d'Yverdon. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *40-41*, 77–91. https://doi.org/10.4000/rde.344

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In S. Bolasco, I. Chiari, & L. Giuliano (Eds.), *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010* (Vol. 2, pp. 1021–1032). Edizioni Universitarie di Lettere Economia Diritto. https://shs.hal.science/halshs-00549779

Hilpert, M., & Gries, S. Th. (2016). Quantitative approaches to diachronic corpus linguistics. In M. Kytö & P. Pahta (Eds.), *The cambridge handbook of english historical linguistics* (pp. 36–53). Cambridge University Press. https://doi.org/10.1017/CBO9781139600231.003

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–

1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holmberg, L. (2017). *The maurists' unfinished encyclopedia*. Voltaire Foundation.

Horton, R., Morrissey, R., Olsen, M., Roe, G., & Voyer, R. (2009). Mining eighteenth century ontologies : Machine learning and knowledge classification in the encyclopédie. *Digital Humanities Quarterly 3.2*, *3*, 1–15.

Huitfeldt, C. (1994). Multi-dimensional texts in a one-dimensional medium. *Computers and the Humanities*, *28*(4), 235–241. https://doi.org/10.1007/BF01830270

Humbel, M., Nyhan, J., Vlachidis, A., Sloan, K., & Ortolja-Baird, A. (2021). Named-entity recognition for early modern textual documents : A review of capabilities and challenges with strategies for the future. *Journal of Documentation*, *77*(6), 1223–1247. https://doi.org/10.1108/JD-02-2021-0032

Ide, N. (1995). *Encoding standards for linguistic corpora.* https://api.semanticscholar.org/CorpusID:59692744

Ide, N., & Sperberg-Mcqueen, C. M. (1995). *The text encoding initiative : Its history, goals, and future development.*

Ide, N., & Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities*, *29*(2), 167–179. https://doi.org/10.1007/BF01830710

Ide, N., & Véronis, J. (1998). *Background and context for the development of a corpus encoding standard.*

Ide, N., & Véronis, J. (2000). *Extracting knowledge bases from machinereadable dictionaries : Have we wasted our time ?* 10.

Jacquet-Pfau, C. (2015). Élaboration et destinée d'une encyclopédie la fin du XIXe siècle : Les trente-et-un volumes de *la grande encyclopédie*, inventaire raisonné des sciences, des lettres et des arts par une société de savants et de gens de lettres. *Éla. Études de Linguistique Appliquée*, *177*, 85–100. https://doi.org/10.3917/ela.177.0085

Jacquet-Pfau, C. (2022). Actualiser l'Encyclopédie de Diderot et d'Alembert au XIXe siècle : Le Grand Dictionnaire universel (1866-1890) et La Grande Encyclopédie (1885-1902): *Langue Française*, *N° 214*(2), 95–109. https://doi.org/10.3917/lf.214.0095

Jaquet, C. (2010). *Bacon et la promotion des savoirs* (p. 304). Presses Universitaires de France. https://www.cairn.info/bacon-et-la-promotion-des-savoirs--9782130568575.htm

Jelliffe, R. (n.d.). *Information technology –Document Schema Definition Languages (DSDL) –Part 3 : Rule-based validation, Schematron* (International {Standard} {ISO}/{IEC} No. 19757 ; Version 3). https://schematron.com/

Jentsch, P., & Porada, S. (2020). From Text to Data : Digitization, Text Analysis and Corpus Linguistics. In S. Schwandt (Ed.), *Digital Humanities Research* (1st ed., Vol. 1, pp. 89–128). Bielefeld University Press / transcript Verlag. https://doi.org/10.14361/9783839454190-004

Ji, Y., Yujing, J., & Tutin, A. (2018). Les routines de topicalisation dans les écrits scientifiques en français. In M. M. T. Agnès (Ed.), *11es journées du réseau Lexicologie, Terminologie, Traduction à Grenoble "Lexique(s) et genre(s) textuel(s) : approches sur corpus " : Vols. Lexique(s) et genre(s) textuel(s) : approches sur corpus. Actes de la conférence LTT 2018 - 11es journées du réseau Lexicologie, Terminologie, Traduction à Grenoble.* Editions des Archives Contemporaines. https://doi.org/10.17184/eac.9782813003454

Jong, F. de, Maegaard, B., Fišer, D., Uytvanck, D. van, & Witt, A. (2020). Interoperability in an in-

frastructure enabling multidisciplinary research : The case of CLARIN. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 3406–3413). European Language Resources Association. https://aclanthology.org/2020.lrec-1.417/

Joshi, A. K., & Schabes, Y. (1997). Tree-adjoining grammars. *Handbook of Formal Languages*. https://api.semanticscholar.org/CorpusID:417746

Kafker, F. A., & Loveland, J. (2016). André-François Le Breton, initiateur et libraire en chef de l'Encyclopédie. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *51*, 107–125. https://doi.org/10.4000/rde.5390

Kay, M. (2017). *XSL transformations (XSLT) version 3.0* [{W3C} Recommendation]. W3C.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. Springer.

Kordjamshidi, P., Otterlo, M. van, & Moens, M.-F. (2011). Spatial role labeling : Towards extraction of spatial relations from natural language. In *ACM Transactions on Speech and Language Processing* (No. 3 ; Vol. 8). Association for Computing Machinery, Inc.

Kordjamshidi, P., Van Otterlo, M., & Moens, M.-F. (2010). Spatial role labeling : Task definition and annotation scheme. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/846_Paper.pdf

Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de Lexicologie*, *1*(108), 91–106. https://doi.org/10.15122/isbn.978-2-406-06281-3.p.0091

Kraif, O., Novakova, I., & Sorba, J. (2016). Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction. *Lidil*, *53*, 143–159. https://doi.org/10.4000/lidil.3976

Krishnamurthy, R. (2008). Corpus-driven Lexicography. *International Journal of Lexicography*, *21*(3), 231–242. https://doi.org/10.1093/ijl/ecn028

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts : A survey. *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397. https://aclanthology.org/C18-1117

La Cotardière, P. de. (2019). *Les descendants de charles-joseph panckoucke et leurs alliances* (pp. 81–115). Classiques Garnier. https://classiques-garnier.com/panckoucke-et-l-encyclopedie-methodique-ordre-de-matieres-et-transversalite-les-descendants-de-charles-joseph-panckoucke-et-leurs-alliances.html

Labbé, C., & Labbé, D. (2000). La distance intertextuelle et l'attribution d'auteur. Corneille et Molière. *Journal of Quantitative Linguistics*, *8*(3), 213–231. https://shs.hal.science/halshs-00137675

Labbé, C., & Labbé, D. (2010). La diachronie dans le discours politique. Le général de Gaulle. In D. BANKS (Ed.), *Aspects diachroniques du texte de spécialité* (pp. 129–148). L'Harmattan. https://shs.hal.science/halshs-00117147

Labbé, D. (1983). *François mitterrand : Essai sur le discours*. 191.

Laboulais, I. (2003). La géographie dans les arbres encylcopédiques de la seconde moitié du XVIIIe

siècle. *Visages et usages de la géographie au moment de l'émergence des sciences humaines*, 63–93. https://hal.science/hal-02932786

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *CoRR*, *abs/1910.09700*. http://arxiv.org/abs/1910.09700

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les Langages Du Politique*, *1*(1), 127–165. https://doi.org/10.3406/mots.1980.1008

Lafon, P. (1981). Analyse lexicométrique et recherche des cooccurrences. *Mots. Les Langages Du Politique*, *3*(1), 95–148. https://doi.org/10.3406/mots.1981.1041

Lance, G. N., & Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies : 1. Hierarchical Systems. *The Computer Journal*, *9*(4), 373–380. https://doi.org/10.1093/comjnl/9.4.373

Language and machines : Computers in translation and linguistics. (1966). [Computer software]. The National Academies Press. https://doi.org/10.17226/9547

Laramée, F. D. (2017). La production de l'espace dans l'Encyclopédie. *Document Numerique*, *20*(2), 159–177. https://www.cairn.info/revue-document-numerique-2017-2-page-159.htm

Larrousse, N., Baude, O., Pouyllau, S., & Jouguet, H. (2023, October). Huma-Num, 10 years of building an infrastructure in SSH : a revolution in practices or an evolution ? *SciDataCon-IDW Salzburg 2023*. https://shs.hal.science/halshs-04333382

Le Cun, Y. (1986). Learning Process in an Asymmetric Threshold Network. In E. Bienenstock, F. F. Soulié, & G. Weisbuch (Eds.), *Disordered Systems and Biological Organization* (pp. 233–240). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-82657-3_24

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Le Guern, M. (1983). Le « dictionnaire » de trévoux (1704). *Cahiers de l'AIEF*, *35*(1), 51–68. https://doi.org/10.3406/caief.1983.2402

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning*, 1188–1196.

Lebart, L., Pincemin, B., & Poudat, C. (2019). *Analyse des données textuelles / ludovic lebart, bénédicte pincemin et céline poudat*. Presses de l'Université du Québec.

Leca-Tsiomis, M. (1995). *Écrire l'encyclopédie : Diderot, de l'usage des dictionnaires à la grammaire philosophique* (pp. 2 vol. (XXIII–583 f.)) [PhD thesis]. http://www.theses.fr/1995PA070071

Leca-Tsiomis, M. (2006). Une tentative de conciliation entre ordre alphabétique et ordre encyclopédique. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *40-41*, 55–66. https://doi.org/10.4000/rde.347

Leca-Tsiomis, M. (2009). Des dictionnaires comme vecteurs du savoir : De Furetière à l'Encyclopédie. In L. Andries (Ed.), *La Construction des savoirs* (pp. 29–42). Presses universitaires de Lyon. https://doi.org/10.4000/books.pul.11795

Leca-Tsiomis, M. (2023). Cyclopædia ? Encyclopédie ?Des « titres fastueux et nouveaux ». *Dix-Huitième Siècle*, *55*(1), 437–452. https://doi.org/10.3917/dhs.055.0437

Lehmann, A., & Martin-Berthet, F. (2018). *Lexicologie. Sémantique, morphologie et lexicographie : Vols.*

5ème éd. (p. 352). Armand Colin. https://www.cairn.info/lexicologie--9782200622374.htm

Léon, J. (2002). Le CNRS et les débuts de la traduction automatique en France. *La Revue Pour l'histoire Du CNRS*, *6*. https://doi.org/10.4000/histoire-cnrs.3461

Léon, J. (2008). Aux sources de la « corpus linguistics » : Firth et la london school. *Langages*, *171*(3), 12–33. https://doi.org/10.3917/lang.171.0012

Léon, J. (2015). *Histoire de l'automatisation des sciences du langage*. ENS Éditions. https://doi.org/10.4000/books.enseditions.3733

Loiseau, S., Gréa, P., & Magué, J.-P. (2011). Dictionnaires, théorie des graphes et structures lexicales. *Revue de Sémantique Et Pragmatique*, *27*, 51. https://shs.hal.science/halshs-00801694

Lojkine, S. (2013).“Et l'auteur anonyme n'est pas un lâche ...” Diderot, l'engagement sans le nom. *Littératures Classiques*, *2013/1*(80), 249–263. https://hal.archives-ouvertes.fr/hal-01081454

Longrée, D., Mellet, S., & Luong, X. (2008). *Les motifs : Un outil pour la caractérisation topologique des textes*. https://www.semanticscholar.org/paper/Les-motifs-%3A-un-outil-pour-la-caract%C3%A9risation-des-Longr%C3%A9e-Mellet/314f002c7df84b24720bbe8152040d29f85c2220

Lowthorp, J. (1665). *Philosophical transactions of the Royal Society of London : Vols. v.1=no.1-22 (1665-1666)*. Royal Society of London. https://www.biodiversitylibrary.org/item/183299

Macary, J. (1973). Les dictionnaires universels de furetière et de trévoux, et l'esprit encyclopédique moderne avant "l'encyclopédie". *Diderot Studies*, *16*, 145–158.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. https://aclanthology.org/J93-2004

Maurer, Y., Schneider, P., & Marschall, R. (2023). Nautilus : An end-to-end METS/ALTO OCR enhancement pipeline. *LIBER Quarterly*, *33*(1), 1–19.

Mayaffre, D. (2000). *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres. Maurice Thorez, Léon Blum, Pierre-Etienne Flandin et André Tardieu (1928-1939)*. Honoré Champion. https://hal.science/hal-01364904

Mayaffre, D., Pincemin, B., & Poudat, C. (2019). Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse de discours. *Langue Française*, *203*(3), 101–115. https://doi.org/10.3917/lf.203.0101

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv :1301.3781*.

Minsky, M. (1952). *A Neural-Analogue Calculator Based Upon a Probability Model of Reinforcement*. Harvard University Psychological Laboratories.

Mladenić, D., & Grobelnik, M. (2003). Feature selection on hierarchy of web documents. *Decision Support Systems*, *35*(1), 45–87. https://doi.org/10.1016/S0167-9236(02)00097-0

Moll, R. N., Arbib, M. A., & Kfoury, A. J. (1988). Parsing, part i. In *An introduction to formal language theory* (pp. 64–79). Springer US. https://doi.org/10.1007/978-1-4613-9595-9_4

Moncla, L., & Gaio, M. (2015, November). A multi-layer markup language for geospatial semantic

annotations. *9th Workshop on Geographic Information Retrieval*. https://doi.org/10.1145/2837689.2837700

Moncla, L., & Gaio, M. (2023, April). Perdido : Python library for geoparsing and geocoding French texts. *First International Workshop on Geographic Information Extraction from Texts (GeoExT)*. https://hal.science/hal-04049794

Moncla, L., Vigier, D., & McDonough, K. (2024a). *Fr_spacy_custom_spancat_edda (revision 636c533)*. Hugging Face. https ://doi.org/ 10.57967/hf/1764

Moncla, L., Vigier, D., & McDonough, K. (2024b). *GeoEDdA : A Gold Standard Dataset for Named Entity Recognition and Span Categorization Annotations of Diderot & d'Alembert's Encyclopédie* (Version 1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10530178

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, *86*(1), 82–85. https://doi.org/10.1109/JPROC.1998.658762

Morin, R. (1989). Diderot, l'encyclopédie et le dictionnaire de trévoux. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *7*(1), 71–122. https://doi.org/10.3406/rde.1989.1034

Mouhouche, A., & El-Hajjami, A. (2014). Étude diachronique et évolution du concept de résonance en physique. *Meta*, *58*(2), 430–448. https://doi.org/10.7202/1024182ar

Moureau, F. (2001). *Le Roman vrai de l'Encyclopédie* (pp. 124–129). Gallimard. https://catalogue.bnf.fr/ark:/12148/cb36204966x

Muller, C., & Brunet, É. (1988). La statistique résout-elle les problèmes d'attribution ? *Strumenti Critici*, *III*(3), 367–387. https://hal.science/hal-04682159

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification [Journal Article]. *Lingvisticæ Investigationes*, *30*(1), 3–26. https ://doi.org/https://doi.org/10.1075/li.30.1.03nad

Nazarenko, A., Habert, B., & Salem, A. (1997). *Les linguistiques de corpus* (p. 240). Armand Colin. https://hal.science/hal-00619268

Nivre, J., & Scholz, M. (2004). Deterministic dependency parsing of English text. *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*, 64–70. https://aclanthology.org/C04-1010

Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal Dependencies. In A. Klementiev & L. Specia (Eds.), *Proceedings of the 15th conference of the European chapter of the association for computational linguistics : Tutorial abstracts*. Association for Computational Linguistics. https://aclanthology.org/E17-5001

Nogueira, A. F. R., Oliveira, H. S., Machado, J. J. M., & Tavares, J. M. R. S. (2022). Transformers for urban sound classification–a comprehensive performance evaluation. *Sensors*, *22*(22), 8874–8874.

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, *47*(2), 617–644. https://doi.org/10.1109/T-AIEE.1928.5055024

Pamuksaç, A., & Sanvido, L. (2023). Distinguer des genres textuels en français langue étrangère : Étude textométrique de marqueurs subjectifs. *Synergies Espagne*, *16*(16), 129–242.

Passeron, I. (2006). Quelle(s) édition(s) de la Cyclopœdia les encyclopédistes ont-ils utilisée(s) ? *Recherches Sur Diderot Et Sur l'Encyclopédie*, *40-41*, 287–292. https://doi.org/10.4000/rde.4342

Patel, C., Patel, A., & Patel, D. (2012). Optical character recognition by open source OCR tool tesseract : A case study. *International Journal of Computer Applications*, *55*, 50–56. https://doi.org/10.5120/8794-2784

Peterson, A., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest : Measuring polarization in westminster systems. *Political Analysis*, *26*(1), 120–128. https://doi.org/10.1017/pan.2017.39

Pierrel, J.-M., & Parisse, C. (2016, October). ORTOLANG : a French infrastructure for Open Resources and TOols for LANGuage. *5th CLARIN Annual Conference*. https://hal.science/hal-01372257

Pincemin, B. (2012). Hétérogénéité des corpus et textométrie : *Langages*, *n° 187*(3), 13–26. https://doi.org/10.3917/lang.187.0013

Pincemin, B. (2022). Sémantique textométrique. In A. Biglari & D. Ducard (Eds.), *La sémantique au pluriel. Théories et méthodes* (pp. 373–396). Presses universitaires de Rennes. https://shs.hal.science/halshs-03763801

Plesser, H. E. (2017). Reproducibility vs. Replicability : A brief history of a confused terminology. *Frontiers in Neuroinformatics*, *11*, 76. https://doi.org/10.3389/fninf.2017.00076

Ploux, S., & Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Revue TAL : Traitement Automatique Des Langues*, *39*, 161–182. https://shs.hal.science/halshs-00009433

Pruvost, J. (2022). Le regard encyclopédique de l'académie française. *Langue Française*, *214*(2), 111–120. https://doi.org/10.3917/lf.214.0111

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Quemada, B. (1968). *Les dictionnaires du français moderne, 1539-1863 : Étude sur leur histoire, leurs types et leurs méthodes*. Didier.

Quemener, E., & Corvellec, M. (2013). SIDUS–the solution for extreme deduplication of an operating system. *Linux Journal*, *2013*(235), 3 :3.

Rastier, F. (1996). *Textes & sens / institut national de la langue française-CNRS ; centre de linguistique française-paris IV ; études publiées par françois rastier*. Institut national de la langue française 1981-2001 ; Didier érudition Institut national de la langue française.

Rastier, F. (2001). Chapitre II. La sémiotique : Du signe au texte. *Formes Sémiotiques*, 51–72. https://shs.cairn.info/article/PUF_RAST_2001_01_0051?lang=fr&tab=texte-integral

Rastier, F., & Pincemin, B. (1999). Des genres à l'intertexte. *Cahiers de Praxématique*, *33*, 83–111. https://doi.org/10.4000/praxematique.1974

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT : Sentence embeddings using siamese BERT-networks*. https://arxiv.org/abs/1908.10084

Rey, A. (2006). *Antoine furetière : Un précurseur des lumières sous louis XIV / alain rey*. Fayard.

Rey, C. (2022). La professionnalisation des encyclopédistes : Un facteur d'évolution des connaissances : *Langue Française*, *N° 214*(2), 81–94. https://doi.org/10.3917/lf.214.0081

Roe, G., Gladstone, C., & Morrissey, R. (2016). Discourses and Disciplines in the Enlightenment : Topic Modeling the French Encyclopédie. *Frontiers in Digital Humanities*, *2*. https://doi.org/10.3389/fdigh.2015.00008

Romary, L., & Tasovac, T. (2018, September). TEI Lex-0 : A Target Format for TEI-Encoded Dictionaries and Lexical Resources. *TEI Conference and Members' Meeting*. https://inria.hal.science/hal-02265312

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rosset, S., Grouin, C., & Zweigenbaum, P. (2011). *Entités nommées structurées : Guide d'annotation quaero* (Notes et Documents LIMSI N° : 2011-04). LIMSI-CNRS.

Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L. A., Benureau, F. C. Y., Brown, C. T., Buyl, P. de, Caglayan, O., Davison, A. P., Delsuc, M.-A., Detorakis, G., Diem, A. K., Drix, D., Enel, P., Girard, B., Guest, O., Hall, M. G., Henriques, R. N., …Zito, T. (2017). Sustainable computational science : The ReScience initiative. *PeerJ Computer Science*, *3*, e142. https://doi.org/10.7717/peerj-cs.142

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Russell, B. (1940). *An inquiry into meaning and truth.* (p. 438). W W Norton & Co.

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC' 10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/701_Paper.pdf

Saint Pair, G. de. (2022). *Chronique rimée du Mont Saint Michel*. https://doi.org/10.34847/nkl.ec7185r8

Salgado, A., Romary, L., Costa, R., Tasovac, T., Khan, A. F., Ramos, M., Almeida, B., Carvalho, S., Khemakhem, M., Silva, R., & Lehečka, B. (2024). The morais dictionary : Following best practices in a retro-digitized dictionary project. *International Journal of Humanities and Arts Computing*, *18*(1), 125–147. https://doi.org/10.3366/ijhac.2024.0325

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Salvat, C. (2006). Les articles "œ\économie" et leurs désignants. *Recherches Sur Diderot Et Sur l' Encyclopédie*, *40-41*, 107–126. https://doi.org/10.4000/rde.352

Sayre, K. M. (1973). Machine recognition of handwritten words : A project report. *Pattern Recognition*, *5*(3), 213–228. https ://doi.org/https://doi.org/10.1016/0031-3203(73)90044-7

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio*, *9*(3), 10.1128/mbio.00525–18. https://doi.org/10.1128/mbio.00525-18

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. *Proceedings of the ACL SIGDAT-Workshop*. https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Schreibman, S. (2011). Editorial Introduction to the First Issue. *Journal of the Text Encoding Initiative*, *Issue 1*. https://doi.org/10.4000/jtei.229

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierar-

chy. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the fourth international conference on language resources and evaluation (LREC' 04)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/65.pdf

Sekine, S., Sudo, K., & Nobata, C. (2002). Extended named entity hierarchy. In M. González Rodríguez & C. P. Suarez Araujo (Eds.), *Proceedings of the third international conference on language resources and evaluation (LREC' 02)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/120.pdf

Shafranovich, Y. (2005). *Common Format and MIME Type for Comma-Separated Values (CSV) Files* (No. 4180). RFC 4180 ; RFC Editor. https://doi.org/10.17487/RFC4180

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21. https://doi.org/10.1109/JRPROC.1949.232969

Sinclair, J. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robin (Eds.), *In memory of J. R. Firth* (pp. 410–430). Longman.

Sinclair, J. (1990). *Collins COBUILD english grammar*. Harper Collins ; Distributed exclusively in Japan by Shūbun International]. https://cir.nii.ac.jp/crid/1130282269779530880

Sinclair, J. (2004). *Trust the text : Language, corpus and discourse* (1st ed.). Routledge.

Sperberg-McQueen, C. M., & Burnard, L. (1999). *TEI P3 Guidelines for the Encoding and Interchange of Machine Readable Texts* (final revision). Zenodo. https://doi.org/10.5281/zenodo.3549598

Standard generalized markup language, Pub. L. No. ISO 8879 :1986 (1986).

Surmont, J.-N. de. (2006). De la genèse à l'informatisation du trésor de la langue française et du grand robert électronique. *Revista de Filología Románica*, *23*, 55–66. https://revistas.ucm.es/index.php/RFRM/article/view/RFRM0606110055A

Tahmasebi, N., Borin, L., & Jatowt, A. (2019). *Survey of computational approaches to lexical semantic change*. https://arxiv.org/abs/1811.06278

Tan, C.-M., Wang, Y.-F., & Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Inf. Process. Manage.*, *38*(4), 529–546. https://doi.org/10.1016/S0306-4573(01)00045-0

Tarrade, L., Magué, J.-P., & Chevrot, J.-P. (2022). Detecting and categorising lexical innovations in a corpus of tweets. *Psychology of Language and Communication*, *26*(1), 313–329. https://doi.org/10.2478/plc-2022-15

Tesnière, L., & Fourquet, J. (1959). *Eléments de syntaxe structurale / lucien tesnière ; préface de jean fourquet,...* Librairie C. Klincksieck.

Teubert, W. (2005). Corpus linguistics and lexicography : The beginning of a beautiful friendship. *Lexicographica*, *20*(2004), 1–19.

Thompson, K. (1984). Reflections on trusting trust. *Commun. ACM*, *27*(8), 761–763. https://doi.org/10.1145/358198.358210

Todorov, N. (2018). Enseigner la géographie à l'école primaire et la genèse des identités territoriales en europe au tournant du XVIIIe au XIXe siècle. *Revue de Géographie Historique*, *13*.

Tutin, A., & Kraif, O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : L'apport des arbres lexico-syntaxiques récurrents. *Lidil*, *53*, 119–141. https://doi.org/10.4000/lidil.3966

Tutin, A., & Véronis, J. (1998). Electronic dictionary encoding : Customizing the TEI guidelines. In A. M. Thierry Fontenelle Philippe Hiligsmann (Ed.), *Proceedings of the 8th EURALEX international congress* (pp. 363–374). Euralex.

Underwood, T. (2016). The life cycles of genres. *Journal of Cultural Analytics*, *1*(1), 11061. https://doi.org/10.22148/16.005

Underwood, T. (2018). *The historical significance of textual distances*. arXiv. https://doi.org/10.48550/ARXIV.1807.00181

Underwood, T. (2020). Machine learning and human perspective. *PMLA/Publications of the Modern Language Association of America*, *135*(1), 92–109. https://doi.org/10.1632/pmla.2020.135.1.92

Urieli, A., & Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. *Actes de La 20e Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN'2013)*, 188–201.

Verdegem, P. (2024). Dismantling AI capitalism : The commons as an alternative to the power concentration of big tech. *AI & SOCIETY*, *39*(2), 727–737. https://doi.org/10.1007/s00146-022-01437-8

Verdier, N. (2015). Entre diffusion de la carte et affirmation des savoirs géographiques en france : Les paradoxes de la mise en place de la carte géographique au xviii e siècle. *Espace Géographique*, *Tome 44*(1), 38–56.

Viard, J. (2006). L'utilité des « désignants » dans l'analyse de la présentation par D'Alembert des propriétés physiques : L'exemple de l'élasticité et de la dureté. *Recherches Sur Diderot Et Sur l'Encyclopédie*, *40-41*, 143–154. https://doi.org/10.4000/rde.356

Vigier, D. (2017). *Autour des prépositions en, dans, dedans. Vers une approche diachronique sur corpus outillé* [Habilitation {\'a} diriger des recherches, Université Toulouse Jean Jaurès]. https://shs.hal.science/tel-03352784

Vigier, D. (2022). L'esprit encyclopédique moderne en France entre 1690 et 1902. Présentation. *Langue Francaise*, *214*(2), 9–26. https://www.cairn.info/revue-langue-francaise-2022-2-page-9.htm

Vigier, D., & Brenon, A. (2021). *La Grande Encyclopédie ou Inventaire raisonné des Sciences, des Lettres et des Arts*. 264458212, 185163346, 398759061, 1371168210. https://doi.org/10.34847/NKL.74EB1XFD

Vigier, D., Moncla, L., Lefort, I., Joliveau, T., & McDonough, K. (2022). Les articles de géographie dans le Dictionnaire Universel de Trévoux et l'Encyclopédie de Diderot et d'Alembert : *Langue Française*, *N° 214*(2), 59–80. https://doi.org/10.3917/lf.214.0059

Walsh, J. A. (2012). Comic Book Markup Language : An Introduction and Rationale. *Digital Humanities Quarterly*, *006*(1).

Wellmon, C. (2015). *Organizing enlightenment : Information overload and the invention of the modern research university*. JHU Press.

Werbos, P. (1974). *Beyond regression : New tools for prediction and analysis in the behavioral science. Thesis (ph. D.). Appl. Math. Harvard university* [PhD thesis].

Wick, C., Reul, C., & Puppe, F. (2018). Comparison of OCR accuracy on early printed books using the open source engines calamari and OCRopus. *Journal for Language Technology and Computational Linguistics*, *33*(1), 79–96. https://doi.org/10.21248/jlcl.33.2018.219

Wilks, Y. (2010). Corpus linguistics and computational linguistics. *International Journal of Corpus Linguistics - INT J CORPUS LINGUIST*, *15*, 408–411. https://doi.org/10.1075/ijcl.15.3.12wil

Williams, G. (2017). MAPPING A DICTIONARY Using Atlas ti and XML to analyse a late XVII th century dictionary. *RiCognizioni*, *7*(4), 161–180. https://doi.org/10.13135/2384-8987/2104

Willinsky, J. (2001). Chapter 9 : Wittgenstein's dictionary. *Counterpoints*, *184*, 187–219. http://www.jstor.org/stable/42976662

Yu, C., Han, R., Song, M., Liu, C., & Chang, C.-I. (2020). A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 2485–2501. https://doi.org/10.1109/JSTARS.2020.2983224

Zhu, L. (2022). Discours dictionnairique, moule phraséologique et corpus textuel. *Langages*, *225*(1), 127–146. https://www.cairn.info/revue-langages-2022-1-page-127.htm

FOLIO ADMINISTRATIF
THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : **Brenon** DATE de SOUTENANCE : **10 juin 2025**
(avec précision du nom de jeune fille, le cas échéant)

Prénoms : **Alice, Diane, Victoire**

TITRE : **Méthodes et outils pour l'étude diachronique des discours géographiques dans deux encyclopédies françaises**

NATURE : **Doctorat** Numéro d'ordre : **2025ISAL0044**

École Doctorale : **InfoMaths (ED 512)**

Spécialité : **Informatique**

RÉSUMÉ :
Dans le cadre du projet GEODE du LabEx ASLAN, cette thèse étudie en diachronie les discours géographiques dans deux encyclopédies françaises. L'*Encyclopédie, Dictionnaire Raisonné des sciences, des arts et des métiers* (*EDdA*), emblématique du siècle des Lumières, constitue la borne la plus ancienne de l'intervalle. Celui-ci s'étend jusqu'à la fin du XIX^{ème} siècle, représenté par *La Grande Encyclopédie, Inventaire raisonné des Sciences, des Lettres et des Arts* (*LGE*). Au travers de trois contributions principales, les travaux abordent des thématiques telles que la préparation des données et l'encodage en particulier, la classification automatique et les analyses textométriques.
En confrontant des méthodes simples de la théorie des graphes à des remarques structurelles sur les éléments présents dans les encyclopédies, la première étude souligne les différences fondamentales qui existent entre dictionnaires et encyclopédies en termes de structures et aboutit à la proposition d'un encodage XML-TEI pour représenter une encyclopédie, qui est appliqué sur la première édition numérique de *LGE*.
La comparaison de méthodes de classifications permet ensuite de choisir un modèle pour associer un domaine de connaissance à chaque article du corpus, ce qui rend possible la conduite d'analyses contrastives dans la dernière partie de la thèse. Un examen attentif des erreurs commises par un des classifieurs sur l'*EDdA* révèle des ressemblances existant entre les domaines et montre une Géographie très au contact des autres sciences au XVIII^{ème} siècle.
Pour finir, les deux œuvres sont comparées pour mettre en relief les changements survenus dans la manière d'écrire la Géographie. Un premier développement quantifie puis caractérise la variation dans les articles du domaine (à l'aide de statistiques sur le nombre de mots et leurs longueurs puis en étudiant les mouvements d'entrées entre domaines). Les apports de la textométrie sont enfin utilisés pour approfondir la compréhension des liens entre biographies et discours géographiques.

MOTS-CLÉS : **Humanités Numériques, Encyclopédies, Textométrie, Apprentissage Automatique, XML-TEI**

Laboratoire(s) de recherche : **LIRIS (UMR 5205) et ICAR (UMR 5191)**

Directeurs de thèse : **Laforest Frédérique**, **Vigier Denis** & **Moncla Ludovic**

Président du Jury : **(président, nommé au moment de la soutenance)**

Composition du Jury :
Favre Benoît, Professeur des Universités (Aix-Marseille Université) : Rapporteur
Galleron Ioana, Professeure des Universités (Université Sorbonne Nouvelle) : Rapporteure
Doucet Antoine, Professeur des Universités (La Rochelle Université) : Examinateur
Eglin Véronique, Professeure des Universités (INSA-LYON) : Examinatrice
Puren Marie, Maîtresse de Conférences (EPITA) : Examinatrice
Laforest Frédérique, Professeure des Universités (INSA-LYON) : Directrice de thèse
Vigier Denis, Professeur des Universités (Université Lumière Lyon 2) : Co-directeur de thèse
Moncla Ludovic, Maître de Conférences (INSA-LYON) : Invité