Bilan du mémoire et projet professionnel

I. Bilan du mémoire

I.1 Problématique de départ et résultats

Notre mémoire portait sur les titres et les noms sous-spécifiés. Les titres étaient déjà l'objet d'étude de notre travail de première année, où nous avions étudié les récurrences lexico-syntaxiques après le double point. Nous avions découvert que des noms privilégient fortement la position de premier nom après le double point, comme *approche, problème* et *outil*. En l'absence de verbe dans les titres, ce premier nom est aussi la tête du segment qui démarre après le double point.

Nous avions remarqué une similitude lexicale entre ces noms et la liste des noms fréquemment employés de façon sous-spécifiée (**NSS**). Cet emploi se caractérise par un faible contenu sémantique et son sens est complété par un contenu spécifiant, relié au NSS par une construction spécificationnelle (**CS**). Ses trois fonctions sont, cognitivement, la création de concepts temporaires, sémantiquement, la caractérisation de ceux-ci, et, au niveau discursif, une capacité d'emploi quasi-pronominale pour faire référence au concept. On voulait donc savoir si les têtes détectées étaient des NSS inclus dans une construction spécificationnelle, en étendant notre recherche à toutes les têtes de segments d'un titre.

Nous avons réussi à sélectionner les têtes de segments, très majoritairement nominales, vérifiant ainsi que les titres sont essentiellement des syntagmes nominaux. Nous avons ensuite étudié la variation des têtes entre les domaines pour distinguer deux classes : les têtes spécifiques et les têtes transdisciplinaires. Pour identifier les emplois sous-spécifiés, nous utilisions d'abord les définitions opératoires fondées sur une complémentation des noms par une proposition conjonctive ou infinitive. Très sélectives, elles donnent très peu de résultats sur les titres, du fait de leur nature averbale.

Nous avons donc été obligé de recourir à une autre définition opératoire, la complémentation du nom sous-spécifié par un syntagme prépositionnel incluant un syntagme nominal : **NSS** de NC. La notion de définition opératoire est rejetée par Schmid (2018, p. 113) : elles sont un moyen opérationnel de trouver des candidats à la sous-spécification, et non une définition. Les capacités de sélection des opérationnalisations varient et, pour NSS de NC, elle est très faible souligne Schmid (2000, p. 26 ; 2018, p. 115) car elle retourne trop de bruits comme des couples partie de totalité.

Les résultats étant trop nombreux pour une annotation manuelle, nous avons essayé de les filtrer, en observant que les propositions ont un verbe et que l'anglais utilise des gérondifs ou des déverbaux pour NC : nous restreignons notre sélection aux noms désignant une action, en utilisant la base VerbAction. Néanmoins, il y avait encore trop de résultats et trop de bruit pour permettre une annotation manuelle et ainsi calculer une précision. Le calcul du rappel étant lui impossible.

I.2 Difficultés rencontrées

La première difficulté rencontrée dans notre travail a été organisationnelle : il s'agit autant d'un manque de temps que de la qualité du temps disponible, après une journée de travail et en soirée, ou entre midi et deux. Les suivis des problématiques professionnelles et universitaires étaient parfois durs à combiner. Les weekends étaient plus propices pour avancer sur le mémoire, mais je ne pouvais pas entièrement m'y consacrer, à cause d'autres matières, comme le projet en commun notamment, que nous avons très mal géré, ou les nécessités de la vie quotidienne. Une fois les autres matières terminées et grâce à au support familial, cette situation s'est améliorée mais très tardivement sur l'année. Sur la fin, la pression pour rédiger et trouver une solution au problème, ainsi que la fatigue accumulée, a handicapé mes capacités de réflexion et de recul sur mon travail.

La deuxième difficulté est un manque d'appropriation de l'outil statistique qui est indispensable à la linguistique de corpus. L'ampleur du corpus et la pression finale ne m'ont pas permis de me replonger dans mes cours et de poser le problème sereinement. Mon travail manque selon moi de graphiques, de tests de corrélation/covariance, de calculs d'intervalles de confiance et de χ^2.

La troisième difficulté est liée à un manque d'aisance avec le savoir savant. Ce savoir abondant, parcouru d'écoles et de courants stratifiés au fil des ans, est contradictoire, et multiplie les approches et les définitions, qui ne sont pas forcément concordantes, sur des objets d'études proches mais différents (NP de Huygue, 2018 vs NSS de Legallois, 2018). Ce foisonnement, observé notamment par Riegel (2006, p. 27-26), m'a semblé particulièrement aigu en linguistique, d'autant plus que les noms sous-spécifiés sont une notion complexe, brassant lexique, syntaxe, sémantique, discours et cognition.

La quatrième difficulté élargie la précédente : il s'agit de ma méthodologie de la recherche Ayant sauté les deux premières années de licence de linguistique et n'ayant jamais été formé à la méthodologie de la recherche, je traîne des manques que le rythme du master ne m'a pas permis de combler. Ces deux dernières difficultés sont visibles dans ma tendance à la redéfinition, qui est un moyen de me positionner face au foisonnement notionnel, et mon embarras à me positionner face aux autres travaux. L'exemple le plus frappant reste la longue liste d'articles mesurant la performance d'un titre en termes de nombres de citations. Ramener ce nombre aux seules caractéristiques du titre, sa longueur ou la présence d'un double point, sans prendre en compte le journal où il a été publié, les auteurs ou le contenu même de l'article, revient à se demander quelle voiture est la plus rapide en regardant seulement sa couleur. Rouge sera sûrement la réponse, mais je doute qu'elle importe.

I.3 Perspectives de recherche

En premier lieu, améliorer les calculs des têtes spécifiques et des distances entre les domaines.

Notre sélection automatique ne ramène malheureusement pas que des emplois sous-spécifiés : un tri manuel est nécessaire ensuite. Nous devons nous contenter que notre filtre sélectionne seulement des candidats à la sous-spécification. Nous aurions dû constituer un sous-ensemble du corpus, annoté manuellement, et mesurer les taux de rappel et de précision du filtre. Augmenter la sélectivité du filtre permet de se concentrer sur les candidatures les plus probables.

Les têtes transdisciplinaires sont de meilleures candidates à la sous-spécifications que les autres têtes. Néanmoins, nous n'avons pas calculé globalement si le fait d'être une tête est corrélé au fait d'être candidat à la sous-spécification. Chercher NC de NC dans notre corpus de titres remonterait bien trop de résultats, mais on peut déjà restreindre le premier NC aux seuls lemmes de nos têtes transdisciplinaires. On aurait ainsi quatre variables booléennes : (A) être tête, (B) être tête transdisciplinaire, (C) être candidat à la sous-spécification, (D) être un lemme de tête transdisciplinaire. Si B implique A et D, on peut tester les corrélations entre A et C, D et C, et mieux quantifier notre travail sur la relation entre B et C.

Nous avons repéré la construction le problème posé par X. Nous pensons qu'il y a des constructions spécificationnelles propres à certains NSS ou certaines classes de NSS, comme avoir pour objectif/résultat de X (Nakamura, 2017 ; Roze et al., 2014), qui demandent des investigations.

L'interprétation sémantique automatique des titres a toujours été ce qui a motivé mes travaux de première et seconde années. Repérer les NSS, c'est repérer dans le discours la constitution de concepts et leurs caractérisations, et empêcher l'interprétation des NSS comme des noms pleins. Nous n'avons pas pu aborder globalement ce problème, conscient à présent que c'est une problématique très vaste pour la recherche, mais notre filtre perfectionné pourrait être un petit pas dans ce sens.

II. Projet professionnel

II.1 Apports du master

Le master LITL s'appuie sur deux domaines scientifiques : la linguistique et l'informatique. À ces deux domaines principaux s'ajoutent la statistique, l'ergonomie et la recherche.

II.1.1 Apports de l'informatique : Python, Big Data et apprentissage automatisé

En informatique, le principal outil a été le langage de programmation Python, devenu en quelques années un des langages les plus utilisés dans le monde[1]. Bien que je le pratique depuis 2005, ces deux années m'ont permis d'en apprendre encore plus et de l'exploiter au maximum de ses capacités, en manipulant de larges volumes de données dans le cadre de la linguistique de corpus.

Au-delà du texte, ce changement d'échelle, m'a entraîné dans un nouveau domaine de l'informatique, appelé communément le « big data », avec ses problématiques, ses outils et les possibilités qu'il apporte. Notamment, la capacité de valoriser des données, d'en tirer informations, connaissances et prédictions, pour reprendre la pyramide DIKW[2] (Zins, 2007).

Un des outils du big data est un autre apport du master : l'apprentissage automatique supervisée ou non, techniques qui appartiennent au champ de l'intelligence artificielle. Les possibilités pratiques de cet outil sont considérables et les outils pour l'implémenter sont foisonnants, notamment en Python. L'explicabilité des résultats, enjeu fort de la recherche, est moins importante pour certaines applications professionnelles.

Nous avons vu le big data et l'apprentissage automatique sur des données non structurées, du texte en langue naturelle, mais on peut également y intégrer des données plus structurées, comme des fichiers XML ou JSON, des tableurs, des bases de données, des fichiers de logs ou des codes sources de programme dans un datalake[3]. Le traitement automatique de la langue naturelle devient à la fois un composant d'un ensemble logiciel plus vaste en vue de valoriser ces données, mais les techniques du TAL peuvent également être transposées à des langages moins naturels exprimés dans des fichiers journaux d'événements ou des fichiers sources. Un exemple d'application est l'autocomplétion lors de l'écriture de code. L'apprentissage automatique, sur un grand nombre de fichiers sources, permet ainsi d'améliorer la qualité des suggestions proposées pour compléter le code en train d'être écrit. Des plugins ou des logiciels fondés sur ce principe existent déjà : Codota, Kite, IntelliSense ou TabNine[4].

II.1.2 Apports de la linguistique : linguistique de corpus et connaissances générales

Le master nous a énormément appris sur la linguistique de corpus, et, par ricochet, sur le mode de formation des connaissances. Les approches corpus-based et corpus-driven (Biber, 2012) reflètent la dualité antique entre déduction et induction : partir du raisonnement pour aboutir à une hypothèse puis la confronter aux faits, ou, partir des faits pour raisonner dessus et aboutir à une hypothèse qui les transcendent. Quelque soit l'approche choisie, il est clair que la linguistique de corpus est supérieure à l'exemple inventé intuitivement, prisonnier du carcan subjectif du chercheur. La linguistique de corpus a néanmoins ses propres limites, car nul corpus ne peut refléter l'infini des possibles faits linguistiques, et la constitution du corpus est potentiellement porteuse de biais.

Plus globalement, nous avons beaucoup enrichi notre connaissance de la linguistique durant ces deux années, en explorant ses différentes branches et écoles, et notre compréhension du français.

---

[1] PYPL https://pypl.github.io/PYPL.html,TIOBE https://www.tiobe.com/tiobe-index/.Au détriment de Perl/Ruby.
[2] https://en.wikipedia.org/wiki/DIKW_pyramid
[3] https://fr.talend.com/resources/what-is-data-lake/
[4] https://www.codota.com/, https://kite.com/, https://tabnine.com/

II.1.3 Apports des autres domaines : statistique, ergonomie et recherche

La statistique est l'outil central, à la fois pour quantifier les données, et établir des corrélations entre elles. Nous avons observé, sans les utiliser par manque de temps, que de nombreuses bibliothèques Python et le langage R permettent de manipuler facilement données et statistiques. On peut ainsi établir en quelques lignes la normalité de la distribution d'une variable, par exemple avec un test de Shapiro-Wilk, pour savoir si on doit utiliser des méthodes paramétriques ou non paramétriques. Ils proposent avec la même facilité des tests de corrélation ou d'analyse de variance.

Les séances d'ergonomie en mode projet nous ont permis de rapidement voir les différents critères utilisés pour évaluer les interfaces homme-machine, comme ceux de Bastien et Scapin (1993), et de concevoir rapidement quelques prototypes. J'avais déjà suivi un enseignement de la conception et l'évaluation d'IHM dans le cadre du cours d'Interaction Design à la Danish Technical University (DTU), lors de mon Erasmus en dernière année d'école d'ingénieur. Un autre de mes cours danois, sur les méthodologies de la conception des logiciels, fut celui qui se rapprocha le plus d'une initiation à la recherche. Je n'ai par la suite jamais reçu d'autres formations, y compris lors de ma tentative de thèse. Le master LITL est à la fois la formation la plus dure que j'ai suivie, même si cela tient en partie à des contraintes externes, et celle qui m'aura le plus permis d'apprendre à faire et de faire de la recherche, avec les limitations qui étaient les miennes. En cela, je lui en sais fort gré.

II.2 Positionnement du master dans mon parcours professionnel

Ma position est particulière, car je suis déjà inséré dans le monde professionnel toulousain où je souhaite rester pour le moment. Mon profil est paradoxal pour un recruteur : je bénéficie de plus de huit d'expérience en informatique mais j'ai un profil junior en TAL. Cela peut expliquer le silence d'Airbus et d'Apsys concernant mes candidatures sur des postes de terminologue pour la première et d'ingénieur linguiste pour la seconde. Après réflexions et études des salaires, des types d'emplois offerts et des bilans des sociétés Inbenta France, Synapse Développement, Hubware, Coup de puce Expansion, Safety Data-CFH[5] et OneLight-studio[6] j'ai décidé de ne pas y candidater pour l'instant. Depuis plus de trois ans en mission à la direction des systèmes d'information de Thales Alenia Space (TAS), j'apprécie de travailler pour un grand groupe et la responsabilité de gérer deux applications.

Je compte intégrer les apports du master au sein de ma gestion des systèmes d'information à TAS. Un premier projet de détection dans les fichiers journaux des événements est prévu. Il s'agira d'y repérer des incidents remontés par les serveurs et signalés par un message ayant une structure fixe mais un contenu variable. Une association sera faite entre le type d'incident et le projet mentionné dedans pour alerter les bons intervenants au plus vite. Un autre projet, dont les spécifications sont moins avancées, est l'analyse des commentaires dans les fichiers de *commit*, émis lorsqu'un utilisateur valide ses modifications sur un travail collaboratif centralisé. Il s'agira d'associer ses modifications avec les tâches qui lui sont confiées. En plus de mes actions chez TAS, je sensibilise mes collègues de Scalian aux possibilités et enjeux, notamment éthique, du TAL. Notre centre de compétence sur le big data ne dispose pas d'un taliste et j'essaye d'organiser des synergies entre nos différentes compétences.

Il est difficile de savoir sur quoi portera ma prochaine mission et quand elle commencera. Il est sûr néanmoins que les compétences acquises au cours de ce master renforceront ma polyvalence, en s'ajoutant à la gestion des SI, la modélisation et la programmation. J'espère néanmoins conserver le niveau de responsabilité qui est le mien, voir l'accroître, et/ou intégrer un grand industriel dans les cinq années à venir. La piste de la gestion documentaire, où DSI et TAL se rejoignent, est à explorer.

[5] Placée en liquidation judiciaire le 2/5/19 : https://www.societe.com/societe/safety-data-cfh-434738944.html
[6] OneLight-Studio est issue de Prometil et a repris SEMIOS. L'autre partie de Prometil a été rachetée par LGM.

Références

Bastien, J. C., & Scapin, D. L. (1993). *Ergonomic criteria for the evaluation of human-computer interfaces*. Thèse de doctorat, INRIA.

Biber, D. E. (2012). Corpus-Based and Corpus-driven Analyses of Language Variation and Use. Dans *The Oxford Handbook of Linguistic Analysis*. Oxford : Oxford University Press.

Huyghe, R. (2018). Généralité sémantique et portage propositionnel: le cas de fait. *Langue française, 2018(2)*, 35-50.

Legallois, D. (2008). Sur quelques caractéristiques des noms sous-spécifiés. *Scolia, 23*, 109-127.

Nakamura, T. (2017). Extensions transitives de constructions spécificationnelles. *Langue française, 2017(2)*, 69-84.

Riegel, M. (2006). Grammaire des constructions attributives : avec ou sans copule. Dans *Construction, acquisition et communication : Études linguistiques de discours contemporains*, Engwall, G. (éd.). Stockholm : Université de Stockholm (Acta Universitatis Stockholmiensis Romanica Stockholmiensia 23).

Roze, C., Charnois, T., Legallois, D., Ferrari, S. et Salles, M. (2014). Identification des noms sous-spécifiés, signaux de l'organisation discursive. Dans *Proceedings of TALN 2014, 1*, 377-388

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin : Mouton de Gruyter (Topics in English Linguistics 34).

Schmid, H. J. (2018). Shell nouns in English-a personal roundup. *Caplletra. Revista Internacional de Filologia, (64)*, 109-128.

Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. *Journal of the American Society for Information Science and Technology, 58 (4)*, 479-493.