UNIVERSITÉ TOULOUSE Jean Jaurès

Récurrences dans les emplois des noms généraux sous-spécifiés dans les titres de documents scientifiques

Damien Gouteux, sous la direction de Mme Josette Rebeyrolle et M. Ludovic Tanguy

Table des matières

Introduction

L'étude des titres s'inscrit dans l'étude de l'écriture académique. L'étude des titres de documents scientifiques, en particulier de ceux des articles et des documents s'en approchant fonctionnellement, est doublement motivée. La première motivation est didactique, elle veut instruire ou prescrire l'art d'écrire un bon titre, notamment pour les jeunes chercheurs dont la langue maternelle n'est pas l'anglais (Wang et Bai, 2007 ; Soler, 2007 ; Cheng et al., 2012 ; Grant, 2013 ; Aleixandre-Benavent et al., 2014). Écrire un bon titre, c'est rendre service à la fois à celui dont le travail est publié et à la communauté qui pourra retrouver son travail plus facilement. La seconde motivation est une recherche d'efficacité. Dans une époque placée sous le signe du *« publish or perish »,* la maximisation des chances pour que l'article soit accepté, publié, téléchargé et cité est d'une vive importance (Jacques et Sebire, 2009 ; Jamali et Nikzad, 2011). Les deux motivations ne sont pas contradictoires, elles se complètent même : après l'apprentissage des règles d'écriture d'un titre du correct vient la recherche de l'optimisation.

Notre étude porte sur 339 687 titres de documents scientifiques tirés de l'archive ouverte HAL. Dans un espace si crucial et court, on retrouve pourtant des noms au contenu sémantique très faible, des noms généraux (Halliday et Hasan, 1976), qui se rapprochent des noms sous-spécifiés (Legallois, 2008). Un titre est pourtant écrit avec beaucoup de soin, alors comment se justifie la présence de ces noms qui n'apportent que peu d'information ? Quelle information apportent-ils ? Dans quelles constructions et à quelle position apparaissent-ils ? Quels rôles sémantiques jouent-ils ? Notre travail veut répondre à ces questions à l'aide des méthodes de la linguistique de corpus et computationnelle. Dans ce document qui clôt le premier semestre, nous commençons par exposer le titre comme objet d'étude en reprenant la littérature sur le sujet, ainsi que celle sur les noms généraux et les noms sous-spécifiés. Nous poursuivons en exposant notre méthode, nos données et nos outils. Ensuite, nous présentons nos hypothèses et nos premières analyses des récurrences d'emplois des noms généraux avant de terminer par nos pistes et perspectives d'objectifs.

I. Le titre comme objet d'étude

I.1 Le titre, un espace crucial et court

Un titre de document scientifique est un emplacement singulier. D'une part, il est le premier contact entre le document et ses lecteurs, d'autre part il s'agit d'un texte très court.

Mabe et Amin (2002) ont interrogé 5 000 lecteurs de textes scientifiques : ceux-ci lisent 1 142 titres par an, 204 résumés et seulement 97 articles. Le titre est donc l'objet le plus lu par les scientifiques mais aussi le plus discriminant : seulement 8 % des titres lus seront suivis par la lecture de l'article, alors que cette proportion s'élève à 48 % après la lecture du résumé. Les lecteurs jugent donc l'intérêt d'un article essentiellement sur son titre (Goodman, Thacker et Siegel, 2001) dans un cadre où le nombre d'articles publiés ne cesse d'augmenter (Jacques et Sebire, 2010). Bien avant qu'il soit publié, le titre est également le premier contact avec le document que les éditeurs et les pairs chargés de l'évaluer ont. De par son importance, un titre doit donc être écrit avec le plus grand soin (Aleixandre-Benavent et al., 2014).

La longueur du titre, ou plutôt sa brièveté, a fait l'objet de nombreuses études. Tout d'abord, dans un travail prescriptif visant à établir des règles d'écriture pour un titre, Aleixandre-Benavent, Montalt-Resurecció et Valderrama-Zurián (2014) considèrent un titre faisant plus de 20 mots comme étant trop long. Ils rejoignent le manuel de publication de l'American Psychological Association (1994), qui stipule qu'un titre ne devrait pas avoir plus de douze mots. De nombreux travaux mettent la longueur en relation avec d'autres caractéristiques du document titré. Ainsi, Haggan (2004) a étudié la longueur moyenne dans différentes disciplines sur de petit corpus : pour la littérature, ses 237 titres ont une longueur moyenne de 9,4 mots, les 207 titres de linguistique une moyenne de 8,8 mots et les 307 titres de sciences dures, 13,8. Lewison et Hartley (2005) étudient 8 disciplines des sciences dures avec un corpus 216 500 titres issus de différents journaux scientifiques britanniques publiés en 1981, 1986, 1991, 1996 et 2001. Ils constatent également un allongement au fil du temps avec un gain entre 2,0 et 8,7 mots dans cinq disciplines, mais un raccourcissement pour la biologie, -0.3 mots, la chimie, -9.5 mots, et la physique, -2.4 mots, pour des moyennes situées entre 8,7 mots par titre et 14,5. Whissell (2012) étudie 12 313 titres pour tirés de 65 volumes du journal *American Psychologist* de 1946 à 2010 et remarque que les titres globalement s'allongent avec le temps, passant de 7,62 mots par titre de 1946 à 1955, à 8,06 mots de 1979 à 1988 avant de très légèrement régresser à 8.01 mots par titre de 2001 à 2010. Un titre est donc court et Jacques et Sebire (2010) ont étudié la longueur médiane des 25 titres les plus cités et celle des 25 titres les moins cités de trois journaux médicaux. La médiane des longueurs est plus grande pour les plus cités avec 18 contre 9 dans le *Lancet*, 16 contre 13 pour le *British Medical Journal*, et 12 contre 10 pour le *Journal of Clinical Pathology*.

I.2 Buts d'un titre

Les titres sont donc à la fois très importants et très courts, spécificités résumées par Soler (2007) en *« informativity and economy »*. On peut s'interroger sur ce contenu critique et en premier lieu sur son but. De la littérature, deux buts émergent : d'un côté le but d'informer le lecteur sur le contenu de l'article et de l'autre la volonté d'attirer son attention. Lewison et Hartley (2005) jugent ces deux buts compatibles néanmoins la poursuite de l'un influence la réalisation de l'autre. Hartley (2005) constate ainsi que son titre *« Were there any sex differences? Missing data in psychology journals »*, jugé pas assez attractif, a été remplacé par son éditeur par *« More sex please, we're psychologists »*. Hartley juge ce dernier titre beaucoup moins informatif. Dans ce même article, Hartley lie l'allongement des titres à une volonté d'être plus informatif. D'autres auteurs insistent sur la préséance que doit tenir le but informatif, comme Grant (2013) : *« First and foremost, the title should be informative »*. Haggan (2004) ajoute que *« the pragmatic aims of the researcher are much better served by precision and explicitness in pinpointing the exact focus of the research »*. Certains auteurs vont jusqu'à voir une opposition directe entre ces deux buts (Aleixandre-Benavent et al., 2014) car pour être attractif, un titre sacrifie souvent du contenu informatif, que cela soit au profit de l'humour, d'un trait d'esprit ou pour intriguer, ce qui suppose que l'on ne dévoile pas tout.

Enfin, un dernier point vient encore favoriser le but informatif : certains systèmes de bases bibliographiques recherchent uniquement dans le titre les différents mots clés soumis par le chercheur. Whissel (2012) a montré une tendance à des titres plus concrets depuis le milieu des années 80. Grant (2013) émet l'hypothèse que cette tendance pourrait être pour favoriser la recherche du document, en incluant le plus de mots clés possibles dans le titre. Jacques et Sebire (2010) rejoignent cette hypothèse en indiquant qu'un titre plus long est susceptible d'avoir plus de mots clés. Néanmoins, il faut nuancer cette affirmation car de plus en plus de systèmes

bibliographiques permettent la recherche de ces mots clés à l'intérieur du résumé ou du texte même du document, recherche dite « plein texte ».

La préséance du but informatif, l'impact négatif que peut avoir sur lui le but attractif et son importance relative dans la recherche de document sont les raisons pour laquelle nous nous limitons dans ce travail à la fonction informative des titres, en mettant de côté la fonction d'attractivité.

Le contenu critique et concis qu'est un titre peut être aussi analysé sur différents plans. Nous commençons par l'étudier du point de vue des syntaxes possibles. À partir de celles-ci, nous essayons d'aborder une premier fois le contenu sémantique des titres. Mais la sémantique ne peut s'envisager sans le lexique employé dans le titre. Nous abordons donc dans la dernière sous-partie le lexique des titres, avec les noms généraux et sous-spécifiés qui s'y trouvent, et nous revenons sur les différents contenus sémantiques des titres.

I.3 Syntaxes des titres

Haggan (2004) montre que seulement une faible proportion de titres sont des phrases avec un verbe conjugué dans ses trois sous-corpus : 8,5 % pour les sciences dures, 4,2 % pour les lettres et 4,3 % pour la linguistique. Plus globalement, Haggan constate que 90 % des titres étudiés sont des unités syntaxiques incomplètes. Elle les rapproche des C-Units de l'anglais parlé définies par Leech (2000), *« petites unités indépendantes grammaticales »*, de la variété *« stand-alone non clausal »*. Leech avait déjà pointé que, quoique globalement rares à l'écrit, on les trouve néanmoins fréquemment dans les titres. Confortant Haggan, Wang et Bai (2007) montrent que 99 % des 417 titres du *New England Journal of Medicine* parus entre 2003 et 2005 sont des groupes nominaux. Les autres sont des phrases complètes, des groupes prépositionnels ou des groupes avec un verbe au gérondif, d'un usage bien moins fréquent. Soler (2007) montre également que *« la construction structurelle la plus récurrente correspond à la construction groupe nominal »*, 40 % en moyenne des titres des six disciplines étudiées sont des groupes nominaux, la biologie et la biochimie se distinguant par une utilisation de phrases complètes pour 51 % et 46 % respectivement des titres de son corpus. Cheng et al. (2012) montrent la même chose sur leur corpus de 796 titres en linguistique appliquée : 93 % de leurs titres sont des groupes nominaux, suivis éventuellement par un autre segment. Ils expliquent que cette nominalisation permet de condenser l'information, accompagnée de pré et post modificateurs qui permettent de spécifier l'objet d'étude (Soler, 2007 ; Wang et Bai, 2007 ; Rath, 2010). Pour Cheng et al. (2012), 90 % des modificateurs des noms sont des groupes prépositionnels, ce qui est une caractéristique de l'écriture académique (Biber et al., 1999 ; Biber et Gray, 2010). La majorité de ces groupes utilisent *of* ou *in*.

Même si la plupart des titres sont des groupes nominaux, ils peuvent être segmentés, c'est-à-dire composés de plusieurs parties. Une ponctuation segmentatrice étudiée depuis longtemps est le double point. Dillon (1981) prenait même la présence d'un double point comme un facteur de qualité du fait que sa présence dans 314 titres d'articles publiés s'élevait à 73 % contre seulement 13 % dans 474 titres d'articles non publiés. Townsend (1983) confirmait cette idée en trouvant deux fois plus d'utilisations du double point dans les titres publiés que dans ceux non publiés et remarquait, comme Diers et Downs (1994) dix ans plus tard, un accroissement de l'utilisation du double point dans les titres scientifiques de journaux : de 22 % en 1972 à 34 % en 1981 pour 266 titres issus des journaux *New Zealand Psychologist* et *New Zealand Journal of Educational Studies* pour Townsend (1983) et de 20 % dans les années 60 à 30 % dans les années 80 pour Diers et Downs (1994) dans cinq

journaux de soins infirmiers. Dans le corpus de Lewison et Hartley (2005), 42 % des titres avaient au moins un double point et ce taux a augmenté entre 1981 et 2001. Sans surprise, les titres segmentés sont plus longs : Dillon (1982) en comptait une proportion de 60 % dans son corpus et remarquait que les titres sans double point avaient une longueur moyenne de 8 mots contre 17 pour ceux en ayant. Les autres marques de ponctuation segmentatrices peuvent être un tiret, un point, un point d'interrogation ou une virgule (Anthony, 2001 ; Cheng et al., 2012).

Soler (2007) indique que les titres sous forme de question sont très peu fréquents, 5 % en moyenne ; elle cite Haggan (2004) qui explique que la présence d'une question peut être liée à un manque d'information, rendant le sujet de l'article plus difficile à percevoir. Jamali et Nikzad (2011) montrent également que les titres sous forme de question ne comptent que pour 2 % de leur corpus de 2 147 titres, soit 45 titres. Cela rejoint les recommandations d'auteurs (Gustavii, 2008 ; Aleixandre-Benavent et al., 2014) qui déconseillent l'utilisation du point d'interrogation dans un titre. Néanmoins, Ball (2009), sur un corpus de 20 millions de titres de 1966 à 2005 en sciences dures, montrent un accroissement très important, 200 %, de l'utilisation du point d'interrogation dans les titres.

Au niveau des temps du verbe des rares titres étant une phrase complète, Haggan (2004) déclare qu'il s'agit « un optimisme confiant projeté par l'auteur que ce qu'il reporte sera vrai pour l'éternité ».

I.4 Sémantiques des titres et syntaxe

Haggan (2004) signale que le titre doit indiquer « ce que l'article a établi ou ce dont il parle », ce que nous traduisons par son sujet, ou problématique, et son résultat, la réponse à la problématique. Soler (2007) indique que *« le sujet qui sera discuté plus loin dans l'article est présenté en miniature dans le titre »* et que cet objet d'étude est replacé dans le champ scientifique auquel il appartient. De façon positionnelle, Wang et Bai (2007) donnent au premier nom la fonction de dire ce dont l'article parle pour les articles médicaux de leur corpus.

Grant (2013), dans un article très didactique, indique que le titre doit contenir le sujet d'étude et si possibles les résultats, car cela attire un plus grand nombre de téléchargements et de citations selon Paiva, Lima et Paiva (2012). Dernier point pour Grant, le titre doit également donner, si c'est approprié, une indication de la méthode utilisée. Jamali et Nikzad (2011) proposent une typologie avec trois classes : les titres déclaratifs indiquent les résultats en plus du sujet, au contrait des titres descriptifs qui n'indiquent que le sujet. Enfin, le type interrogatif indique le sujet mais en faisant appel à la curiosité du lecteur.

La composition syntaxique n'explicite pas la sémantique associée. Swales et Feak (1994) indique que les titres segmentés par un double point obéissent à quatre combinaisons possibles qui sont problème : solution, général : spécifique, sujet : méthode, majeure : mineure. Poursuivant ce travail en élargissant aux autres marques de ponctuation segmentatrices, Anthony (2001) étudie les titres en informatique et fait ressortir cinq combinaisons qui sont nom – description, sujet – focalisation, sujet – méthode, description – nom, sujet – description. Il critique par ailleurs que la catégorie majeure : mineure de Swales et Feak recoupent les autres. Cheng et al. (2012) poursuivent également dans cette voie en proposant onze combinaisons, en reprenant sujet – focalisation et sujet – méthode d'Anthony et en y ajoutant sujet – description, sujet – source des données utilisées, métaphore – sujet, sujet – question, question – méthode, sujet – méthode + source des données

utilisées, métaphore – question, question – méthode, nom – méthode. Ils montrent que certaines combinaisons sont favorisées par certaines disciplines, ainsi sujet – source n'apparaît pas en informatique alors que nom – description a la plus haute fréquence, combinaison qui n'apparaît pas en linguistique appliquée. L'utilisation d'une métaphore, qui tisse une association entre deux objets, ou d'une question, suscite l'attraction du lecteur, et l'explicitation informative du sujet qui l'accompagne semble pour Cheng et al. une bonne combinaison. On peut résumer les neuf différents éléments sémantiques des combinaisons qui composent un titre ainsi :

Élément sémantique	Définition
Sujet	Le sujet de l'article
Focalisation	Une partie précise du sujet, il s'agit de le délimiter
Méthode	Méthode employée pour obtenir des résultats
Description	Une description
Nom	Une appellation
Source	Source des données utilisées
Métaphore	Association entre deux objets
Question	Une question
Résultat	Résultat de l'article (Jamali et Nikzad, 2011 ; Paiva et al, 2012 ; Grant, 2013)

Les titres apportent de l'information sur le contenu du document titré, le positionne dans un champ donné, mais également distingue son contenu des autres articles (Cheng et al., 2012).

I.5 Lexique des titres, noms généraux et considérations sémantiques

Notre étude du lexique des titres portera uniquement sur les noms car les groupes nominaux constituent l'essentiel des titres (Haggan, 2004 ; Soler, 2007 ; Wang et Bai, 2007 ; Cheng et al., 2012). Nagano (2015) calcule un taux de substantifs par titres car, selon lui, *« ce taux est souvent considéré comme un indicateur pour déterminer combien ce titre est informatif »*.

Wang et Bai (2007) étudient la structure des 417 titres de leur corpus de médecine. Ils montrent que le groupe nominal de tête, celui qui commence le titre, ont pour noyau des noms abstraits comme « effet », « rôle », « comparaison », « efficacité », des nominalisations comme « transplantation », « mutation », « traitement » ou un nom de médicament ou de maladie. Cheng et al. (2012) constatent que les têtes peuvent être spécifiques ou non à la discipline. Les neuf plus fréquents non spécifiques à une discipline sont *effet, analyse, rôle, développement, relation, étude, utilisation, connaissance et influence*.

Grant (2013), dans ses conseils pour écrire un bon titre, insiste sur le fait qu'un titre doit être spécifique et donc employer un vocabulaire précis. Hatier et al. (2016a) n'ont pas étudié particulièrement le vocabulaire des titres mais celui des articles scientifiques. Ils ont extrait un lexique scientifique transdisciplinaire des mots les plus fréquents (Hatier, 2016b). Dans ses 493 noms, on trouve par exemple *cas, étude, exemple, problème, question*. Ces noms sont des noms généraux que l'on retrouve dans les titres.

La catégorie des noms généraux *(general nouns)* a été identifiée brièvement par Vendler (1968), sous le nom de récipients *(container)*. C'est Halliday et Hasan (1976) qui ont forgé le terme de noms généraux. Ils déclarent que ces noms contribuent à la cohésion du discours, Huygue (2018) précise même qu'ils contribuent à sa structuration en partant de l'exemple de *fait*. Flowerdew (2003) déclare qu'il s'agit d'une sous-catégorie des noms abstraits. Ils existent de nombreuses approches théoriques et empiriques pour circonscrire cet objet d'étude, aboutissant à ce que Adler et Moline

(2018) qualifient de plurivalence terminologique, ne serait-ce que pour les qualifier : *low content nouns, unspecific nouns, referring nouns, label, anaphoric nouns, carrier nouns, broad sense nouns,* noms coquilles *(shell nouns)* (Schmid, 2000), noms signalant *(signalling nouns)* (Flowerdew, 2003). Halliday et Hasan (1976) les définissent selon trois critères : fréquence d'utilisation élevée, contenu sémantique faible, application référentielle vaste, en citant les exemples de *chose, objet, problème, question, idée*. L'absence de spécificité sémantique demande donc une saturation contextuelle (Adler et Moline, 2018), un remplissage (Legallois, 2008) sémantique, soit une réalisation lexicale (Winter, 1992) en discours.

À la manière des pronoms, ils ne référencent pas un objet du monde directement mais un segment discursif précédent ou subséquent, on parle de référence endophorique, respectivement anaphorique ou cataphorique. Soit ce segment discursif référencé référence lui-même à un objet dans le monde réel qui est alors la référence indirecte du nom général, soit ce segment discursif est référencé comme un segment du discours en tant que tel, le nom général ayant alors une fonction métalinguistique (Winter, 1992).

Une question est de savoir si les noms généraux sont une catégorie a priori des noms ou un emploi en discours. Adler et Moline (2018) cite l'exemple de *situation* en anglais qui peut à la fois désigner un segment discursif et faire référence au statut professionnel d'une personne. Pour détecter ce qu'il appelle les noms coquilles, Schmid (2000), dans un cadre cognitiviste et constructionnel, propose de détecter deux constructions précises : déterminant + (prémodificateur) + nom + that-clause ou wh-clause ou to-infinitive et l'autre construction étant déterminant + (prémodificateur) + nom + be + that-clause ou wh-clause ou to-infinitive. Les grammaires de construction (François et Legallois, 2006), qui associe à une structure syntaxique une signification, montrent donc que l'emploi de noms généraux s'inscrit dans des patrons récurrents. Legallois (2008), qui parlent de noms sous-spécifiés, proposent pour le français les structures spécificationnelles, un type de constructions, suivantes : nom + être + que-complétive et nom + être + de-infinitifs. Pour cette auteure, il y a une « interdéfinition entre lexique et grammaire » et les noms sous-spécifiés sont employés préférentiellement, mais non exclusivement, dans ces constructions. Notons que si Legallois rapproche les noms sous-spécifiés des noms généraux, les deux concepts ne se recouvrent pas. Les relations entre les deux ont fait l'objet de quatre interventions par Huyghe, Gerhard-Krait, Lammert et une de Schnedecker et Capeau lors des journées d'étude S'caladis de novembre 2018 à l'Université Jean-Jaurès de Toulouse. Dans l'attente d'en savoir plus et constatant leur forte proximité, nous ne les différencions pas dans ce travail.

Si le contenu sémantique des noms généraux est faible, il n'est néanmoins pas nul et permet de catégoriser sémantiquement le contenu référencé (Adler, 2018). Cette catégorisation sémantique n'est pas homogène (Legallois, 2008) et Legallois indique également que la non spécification n'est pas exactement sémantique mais informationnelle. Schmid (2000) propose une catégorisation des noms coquilles en six classes, un nom pouvant appartenir à plusieurs, que Legallois (2008) reprend :

- Domaine factuel : indique que le référent est un fait
- Domaine linguistique : indique que le référent est un objet linguistique, on rejoint la fonction métalinguistique de Winter (1992)
- Domaine mental : indique que le référent est un état cognitif
- Domaine modal : émet un jugement de modalité sur le référent (possibilité, certitude, capacité, permission, obligation)

- Domaine événementiel : indique que le référent est une activité, un procès ou un état
- Domaine circonstanciel : indique que le référent est une manière, une façon de faire

Si la nominalisation permet de condenser l'information, quel est le rôle de ces noms avec un contenu sémantique si faible, dans l'espace si contraint en taille que sont les titres ? C'est à cette question que nous voulons répondre par notre étude. Nous voulons étudier les récurrences d'emplois de ces noms afin de déterminer s'il existe des constructions les employant dans les titres et comment ses constructions s'articulent avec les éléments sémantiques déjà distingués.

II. Méthode, données et outils

II.1 Présentation de HAL

Nos titres sont issus de l'archive ouverte Hyper Article en Ligne[1] (HAL) (Nivard, 2010). Elle compte, au 12 février 2019, 565 282 documents scientifiques et 1 708 795 notices. Chaque chercheur, quelle que soit sa discipline, ou documentaliste d'un centre de recherche, est libre de déposer un document sur HAL, s'il a l'accord de ses auteurs et de son éventuel éditeur. Ce document peut-être un texte, comme un article, une thèse, un livre ou seulement un chapitre, une vidéo, un son, une image ou une carte. Pour les articles, contrairement à une publication dans une revue scientifique, il n'y a pas de contrôle par les pairs du contenu scientifique déposé. Seul un contrôle pour s'assurer du bon format du document et du respect des droits est effectué. En le déposant sur HAL, le document est rendu public et est partagé avec la communauté scientifique beaucoup plus rapidement que via une revue. Les deux options peuvent être complémentaires pour diffuser son travail. Un article déposé sur HAL sans être publié dans une revue à ce moment-là est appelé un preprint.

HAL est géré par le Centre pour la Communication Scientifique directe[2] (CCSD), fondé en 2000 et rattaché au Centre National pour la Recherche Scientifique (CNRS). Il existe des sous-ensembles de HAL dédiés à des disciplines spécifiques, HAL-SHS et MédiHAL, ou pour un type de texte spécifique comme Thèses en ligne. Les avantages des archives ouvertes, par rapport à un site d'une institution particulière ou le site web personnel d'un chercheur, sont la centralisation de l'accès, la diffusion des connaissances et la conservation pérenne des documents. La création des archives ouvertes s'inscrit dans le mouvement pour un accès libre et gratuit aux connaissances scientifiques. La plus ancienne des archives ouvertes est arXiv[3], fondée en 1991 et limitée uniquement aux articles. Un dépôt d'un article dans HAL entraîne automatiquement la création d'une notice dans arXiv s'il entre dans les disciplines couvertes par cette dernière.

Une notice est créée sur HAL lors du dépôt du document et éventuellement dupliquée dans d'autres archives ouvertes. Une notice est un ensemble d'informations sur le document scientifique déposé, appelé métadonnées, comme son titre, sa date de dépôt, son type. La notice contient tout ce qui est nécessaire à notre travail. Pour notre travail, nous considérons que les métadonnées du document sont également celles de son titre.

[1] https://hal.archives-ouvertes.fr/
[2] https://www.ccsd.cnrs.fr/
[3] http://arxiv.org/

Une archive ouverte A peut avoir la notice d'un texte scientifique hébergé sur une autre archive ouverte B, cette dernière aura alors à la fois la notice et l'intégralité du document. Dernier cas possible, il existe des documents qui ne sont pas hébergés par aucune archive ouverte mais simplement référencés par leurs notices. Il s'agit généralement de textes dont les droits appartiennent à des revues payantes. La création de telles notices se fait par le traitement automatisé des références bibliographiques des documents déposés.

II.2 Présentation du corpus et des outils

L'étude de Haggan (2004) portait sur 751 titres, celle de Soler (2007) sur 570, celle de Wang et Bai (2007) sur 417 et celle de Ball (2009) sur un corpus de 20 millions de titres. Tous travaillaient sur des titres en anglais sur un choix de disciplines données. Notre corpus compte 339 687 titres en français de documents scientifiques tirés de l'archive ouverte HAL, soit 4 9090 608 mots et 1 566 048 lemmes[4]. Nous n'avons pas restreint à une discipline donnée, ni n'avons fait de différences entre « research papers » et « review papers » comme Soler (2007), les derniers étant plus rares selon cette même auteure. HAL pouvant contenir plusieurs types de documents, comme des articles mais aussi des sons ou des vidéos, nous nous sommes limités aux articles (154 790, 45 %), communications (115 278, 34 %), chapitres d'ouvrage (66 788, 20 %) et posters (2 831, 1 %). Les documents titrés ont été écrit par un nombre d'auteurs compris entre 1 et 168 auteurs, avec 62 % écrits par un seul auteur et 99 % écrits par neuf auteurs ou moins.

Nous avons enrichi nos données en déterminant pour chaque forme présente dans nos titres, son lemme et sa catégorie (ou classe) grammaticale à l'aide du logiciel Talismane[5], développé à l'Université Jean-Jaurès par Urieli (2013). Avoir le lemme d'un mot permet de rassembler toutes ses formes fléchies sous une même entrée et de compter son nombre d'occurrences en additionnant celles de ses formes fléchies. La catégorie du discours, ou étiquette POS pour *part of speech*, est la base pour analyser ultérieurement la structure syntaxique dans laquelle les formes employées s'inscrivent. Nous avons également utilisé la capacité d'analyse syntaxique en dépendances de Talismane, que nous comptons utiliser plutôt que le modèle syntagmatique (pour une comparaison des deux voir Schwischay, 2001). Nous avons développé un ensemble de scripts en Python pour interroger nos données.

Au niveau de la segmentation, on ne peut que constater la domination du double point : 101 564 titres, représentant 30 % du corpus, contiennent au moins un double point, soit presque le double des titres contenant au moins un point : 56 849 titres, soit 17 % du corpus. Au niveau des questions, seuls 28 758 titres comptent un point d'interrogation, soit 8 % de notre corpus. Seuls 25 985 titres ont au moins un verbe conjugué à l'indicatif, l'impératif ou le subjonctif, soit 8 % de notre corpus, avec 25 476 titres ayant au moins un verbe à l'indicatif, soit 98 % des 25 985 titres ou 7,5 % de notre corpus. La présence d'un verbe conjugué dans ces trois modes traduit sa nature de phrase complète. Le titre le plus long de notre corpus a 284 mots. La longueur moyenne, en nombre de mots est 12,8, la médiane est entre 11 et 12 mots.

[4] Nous devons être prudent sur le nombre de lemmes : Talismane, le logiciel ayant permis de les obtenir ne sait pas toujours obtenir le lemme. Il met alors '_' à la place. Dans ce cas-là, nous avons substitué la forme au lemme mais nous ne sommes pas à l'abri qu'un mot X et un mot X+s ne soit pas connu de Talismane et donc stockés avec deux « pseudo-lemmes » qui sont égales à ces formes.

[5] http://redac.univ-tlse2.fr/applications/talismane/talismane.html

II.3 Présentation de la liste des noms pivots

Nous sommes arrêtés sur cinq noms pivots pour construire nos patrons. Ces cinq noms reviennent dans les différentes listes de la littérature :

Noms pivots	LST[6]	NSS[7]	NG[8]	SH[9]	Position dans l'ordre des fréquences
Cas	Oui	Oui		Oui	5e avec 8 385 occurrences, 0,005 %
Étude	Oui				1e avec 13 264 occurrences, 0,008 %
Exemple	Oui	Oui		Oui	15e avec 5 107 occurrences, 0,003 %
Problème	Oui	Oui	Oui	Oui	Au-delà de la 60e position
Question	Oui	Oui	Oui	Oui	45e avec 3 353 occurrences, 0,002 %

Problème et *question* sont clairement reconnus comme des noms généraux sous-spécifiés. *Exemple* et *cas* ne sont pas reconnus comme nom général par Halliday et Hasan (1976) mais la liste qu'ils donnent est très succincte. *Étude* a un statut beaucoup plus précaire n'étant ni reconnu par Legallois (2008) ou Schmid (2000) mais nous le gardons car il est très fréquent. Au contraire, *problème* ne fait pas partie des 60e noms les plus fréquents dans notre corpus. Une fois arrêté cette courte liste, nous pouvons la projeter sur notre corpus pour détecter les emplois des noms pivots.

III. Hypothèses et premières analyses

III.1 Projection des noms pivots sur le corpus

La première constatation est que l'on ne retrouve pas les patrons de Schmid (2000), ni les constructions spécificationnelles de Legallois (2008) dans les titres. La spécificité de l'écriture des titres, et notamment leur brièveté, s'accorde mal avec les patrons et les constructions mentionnés.

Nous retrouvons néanmoins les patrons `(Début) (NSS) (PONCT :)` et `(DET un) (NSS) (PONCT :)` de Roze et al. (2014) pour *exemple*, *problème* et *question*, ce qui nous laisse à penser que la position où se trouve le nom pivot dans le titre a toute son importance, par rapport au début du titre ou par rapport à une marque de ponctuation segmentatrice et au début du segment.

A) Cas

La projection des lemmes « le cas de » retourne 3 099 résultats. On ne retrouve les occurrences de ce patron que rarement en première position du titre : 42 fois, soit 1 %. Exemples :

- Les Dominicains et la confession royale à la Renaissance : le **cas** de Guillaume Petit, confesseur de Louis XII et de François Ier
- Exposer des objets sonores : le **cas** des chansons de Brassens

Il ne s'agit pas d'un emploi général mais d'un emploi du sens lexical plein du nom cas. Le patron dans lequel il s'inscrit peut s'écrire :

```
<Sujet> : le cas de <focalisation>
```

La focalisation introduite par *« le cas de »* est instance particulière du sujet.

[6] Lexique scientifique transdisciplinaire, Hatier, 2016b.
[7] Nom sous-spécifié, Legallois, 2008.
[8] Noms généraux, Halliday et Hasan, 1976.
[9] Noms coquilles, Schmid, 2000.

B) Étude

Si *étude* est le nom pivot le plus faiblement caractérisé comme général, il est néanmoins le plus fréquent dans notre corpus : ses 13 264 occurrences comptent pour 0.008 % des 1 566 048 lemmes qu'il contient. On dénombre 7 198 occurrences dont 3 151 en première position, soit 44 %. Exemples :

- **Etude** de l'émission acoustique lors de la cristallisation d'acide citrique avec transition de phase
- **Etude** de paramètres acoustiques des voix de patients traités pour un cancer ORL dans le cadre du projet C2SI

Le patron sémantique correspondant s'écrit :

```
Étude de <sujet>
```

C) Exemple

La projection des lemmes « un exemple de » retourne 361 résultats. On retrouve les occurrences de ce patron souvent en première position du titre : 110 fois, soit 30 %. Exemples :

- Un **exemple** de valorisation : la création d'entreprise par un chercheur au statut de fonctionnaire. Approches des spécificités juridiques françaises.
- Un **exemple** de gouvernance rénovée : la lettre de mission des inspecteurs de l'enseignement primaire français.
- Les romans de Juan Benet : un **exemple** de géopoésie

Deux emplois syntaxiques émergent : soit en première position, soit en première position après un double point. Du point de vue sémantique, deux emplois émergent également : soit *« un exemple de »* introduit une focalisation qui est une instance particulière du sujet, soit *« un exemple de »* introduit le sujet. On peut écrire les patrons ainsi :

```
Un exemple de <Sujet> : <focalisation>
```

```
<Focalisation> : un exemple de <sujet>
```

La question de l'articulation entre les deux emplois syntaxiques et les deux emplois sémantiques reste à approfondir.

Nous avons également projet les lemmes « le exemple de » qui retourne 1 776 résultats. On ne retrouve que 11 fois les occurrences de ce patron en première position. Exemples :

- Saisir la culture d'avocat général au Parlement de Paris : l'**exemple** de Pierre VI Gilbert de Voisins (1684-1769)
- Du paradoxe au style paradoxal : l'**exemple** des Caractères de La Bruyère
- Géothermie et planification énergétique territoriale : l'**exemple** du schéma régional de l'Ile-de-France

Au niveau sémantique, l'utilisation de l'article défini inverse le second patron que nous avons identifié avec l'article indéfini :

```
<Sujet> : l'exemple de <focalisation>
```

D) Problème

La projection des lemmes « le problème de » retourne 306 résultats dont 106 occurrences au début du titre, soit 35 %. Exemples :

- Science et histoire chez Hobbes : le **problème** de la méthode
- Le **problème** de la métaphysique comme problème des Lumières
- Le **problème** de la minimisation des arrêts
- La Princesse de Clèves : le **problème** de l'originalité dans la construction de l'identité

On constate deux emplois syntaxiques : l'un en première position, l'autre en première position après un double point. Les patrons sémantiques s'écrivent :

```
Le problème de <sujet>

<Sujet> : le problème de <focalisation>
```

E) Question

La projection des lemmes « la question de » retourne 727 occurrences dont 289 en première position du titre, soit 37 %.

- La **question** de l'Orient hellénisé
- La **question** des contenus scolaires
- Parler au peuple, parler au roi : la **question** des harangues (XVIIe-XVIIIe siècles)
- Conditionnel contrefactuel ou supposition : la **question** du possible chez Musil et Wittgenstein

Là encore, deux emplois émergent : soit en première position, soit juste après un double point. Nous pouvons écrire les patrons sémantiques suivant :

```
La question de <sujet>

<Sujet> : la question de <focalisation>
```

Nous avons étudié dans cette partie cinq premières projections qui soulèvent déjà des points d'intérêt. À partir de ceux-ci, nous dressons une liste de pistes d'amélioration et de perspectives dans la partie qui suit.

IV. Pistes et perspectives

Dans cette partie, nous détaillons plusieurs pistes pour atteindre ce qui pourrait être nos objectifs pour le travail du second semestre.

IV.1 Piste d'améliorations techniques

La projection des noms pivots est pour l'instant « naïve ». Il faut permettre plus de flexibilité, comme un adjectif qualifiant le nom pivot et analyser ce qui se trouve en première et deuxième positions lorsque le nom pivot vient juste après. De plus, nous devons reprendre nos outils de l'année dernière pour recompter le nombre d'apparition avant et après une marque de ponctuation segmentatrice, la position par rapport à celle-ci, sachant que la plus représentée est le double point.

Roze et al. (2014) évoque également les noms *rôle, mission, solution* que nous pouvons projeter également. Plus globalement, tous les 670 noms coquilles de Schmid (2000) sont projetables sur notre corpus mais il faut les croiser avec le Lexique scientifique transdisciplinaire de Hatier (2016b) ou notre propre lexique, en ne gardant que les termes les plus fréquents. Nous pensons à *droit* et *analyse* par exemple.

Toujours en partant de notre lexique et de ses lemmes les plus fréquents, nous pouvons également essayer de projeter d'autres lemmes comme *étude* qui ne sont pas référencés dans les travaux sur les noms généraux ou sous-spécifiés. Nous pensons également à *approche* et *système*. Ils ont déjà deux des traits des noms généraux selon Halliday et Hasan (1976) : une fréquence d'utilisation élevée et une application référentielle vaste.

Ces projections à partir d'un mot pivot ne peuvent faire l'économie d'être projeté à la fois avec un article défini et un article indéfini. L'utilisation de l'article défini, ou d'un déterminant possessif mais qui reste très rare dans les titres, est une caractéristique des noms sous-spécifiés (Legallois, 2008).

Lorsque Talismane ne sait pas obtenir le lemme d'un mot, il attribue '_' à la valeur de son lemme. Dans ce cas-là, nous avons substitué la forme au lemme mais nous ne sommes pas à l'abri qu'un mot X et un mot X+s ne soit pas connu de Talismane et donc stockés avec deux « pseudo-lemmes » égalent à ces formes au lieu d'être rassemblés sous le même lemme. Nous voulons fournir une méthode plus robuste de prise en charge des lemmes inconnus de Talismane, en les stockant et en les vérifiant soit pas par des dérivations morphologiques simples, si X existe, ne pas stocker X+s, soit par une validation manuelle des lemmes inconnus si leur nombre n'est pas trop important.

À la suite d'Anthony (2001) et Cheng et al. (2012), nous pourrions également étudier la répartition de ces constructions et des mots pivots parmi les différentes disciplines. Nous n'avons pour l'instant mis en jeu dans nos constructions que deux éléments sémantiques d'un titre : son sujet et une focalisation. Il serait intéressant d'essayer de détecter par exemple des éléments de méthode et voir l'articulation qu'ils ont avec nos structures, notamment via l'analyse en dépendance.

En effet, nous n'avons pas, pour l'instant, utilisé l'analyse syntaxique en dépendance offerte par Talismane et que nous stockons en mémoire. Elle offre une vision des liens entre les éléments du titre et les relations auxquelles le mot pivot appartient sont les plus importantes à étudier.

IV.2 Perspectives théoriques

Plusieurs perspectives théoriques s'ouvrent devant nous. La première est la spécification des constructions que nous voyons émerger autour des noms pivots détectés et notamment le fait d'établir clairement le lien entre syntaxe et sémantique dans ces constructions.

Si de nouvelles publications jettent une lumière plus affirmée sur la distinction entre noms sous-spécifiés et noms généraux, cela permettrait de positionner le travail du second semestre sur cette question également.

La question des noms non classés comme généraux ou sous-spécifiés mais très fréquents dans notre corpus se pose également. Les deux phénomènes sont-ils véritablement liés comme nous en avons fait l'hypothèse ? Est-ce que les titres offrent un espace où certains noms peuvent trouver un emploi de noms généraux de façon bien plus fréquente ? Ou bien s'agit-il d'une catégorisation des

noms différente, un ensemble où on retrouverait des noms généraux, des noms sous-spécifiés et ces autres noms. Dans ce cas-là, il faudrait en établir les critères d'appartenance et essayer d'en obtenir une liste à partir de notre corpus.

Cette question rejoint celle de la découverte de nouvelles constructions, potentiellement spécificationnelles comme celles de Legallois (2008). Nous devons pouvoir projeter des constructions « à trou » pour détecter d'éventuels mots pivots qui partageraient les mêmes constructions.

Cela amène à la question des permutations possibles entre les différents noms pivots qui est également intéressante. Si la permutation syntaxique ne fait pas de doute, c'est au niveau sémantique qu'il faut l'étudier. Nous avons mentionné que les noms généraux ont un contenu sémantique très faible mais non nul. Le remplacement de l'un par l'autre change donc la qualification apportée au segment discursif référencé. Il faut étudier les possibilités et les impacts de ces permutations.

Enfin, bien qu'un peu trivial, il faut néanmoins se demander dans une perspective didactique, si la présence de ces noms au faible contenu sémantique n'est pas un ornement inutile du titre. Il s'agit de comprendre ce qu'ils apportent au titre et s'ils sont supprimables sans conséquence et, dans ce cas, si leur présence n'est pas un défaut comme ceux que soulèvent Aleixandre-Benavent et al. (2014).

Conclusion

Notre travail s'interroge sur la présence de noms au contenu sémantique faible dans un espace où chaque mot compte, les titres. La densité informationnelle de ceux-ci semble récuser la présence de noms généraux ou sous-spécifiés comme une « perte d'espace » et pourtant ils sont bien présents. Nous remarquons également des noms très fréquents dans nos titres, non identifiés comme généraux ou sous-spécifiés.

Nos premières projections révèlent l'importance de la position de ces noms : au début du titre ou autour d'une marque de ponctuation segmentatrice, généralement un double point. Nous avons esquissé des constructions autour de ces noms qui lient syntaxe et sémantique. Ces constructions permettraient de comprendre computationnellement le titre, ce qui ouvrirait la voie à une véritable recherche sémantique dans les bases bibliographiques des titres de documents scientifiques.

Le second semestre sera l'occasion d'approfondir les pistes techniques pour améliorer nos résultats qui nous aideront ensuite à atteindre les objectifs théoriques de notre projet de recherche.

Bibliographie

Adler, S. et Moline, E. (2018). Les noms généraux: présentation. *Langue française, 2018(2)*, 5-18.

Aleixandre-Benavent, R., Montalt-Resurecció, V. et Valderrama-Zurián, J. (2014). A descriptive study of inaccuracy in article titles on bibliometrics published in biomedical journals. *Scientometrics, 101(1)*, 781-791.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (6ème édition). Washington, DC: American Psychological Association.

Anthony, L. (2001). Characteristic features of research article titles in computer science. *IEEE Transactions on Professional Communication, 44(3)*, 187-194.

Ball, R. (2009). Scholarly communication in transition: The use of question marks in the titles of scientific articles in medicine, life sciences and physics 1966–2005. *Scientometrics, 79(3)*, 667-679.

Cheng, S. W., Kuo, C. W. et Kuo, C. H. (2012). Research article titles in applied linguistics. *Journal of Academic Language and Learning, 6(1)*, A1-A14.

Diers, D. et Downs, F. S. (1994). Colonizing: a measurement of the development of a profession. *Nursing research, 43(5)*, 316.

Dillon, J. (1981). The emergence of the colon: an empirical correlate of scholarship. *American Psychologist, 36*, 879-884.

Dillon, J. T. (1982). In Pursuit of the Colon, A Century of Scholarly Progress: 1880–1980. *The Journal of Higher Education, 53(1)*.

Flowerdew, J. (2003). Signalling nouns in discourse. *English for specific purposes, 22(4)*, 329-346.

François, J. et Legallois, D. (2006). Autour des grammaires de constructions et de patterns. *Cahiers du CRISCO.* Université de Caen.

Goodman, R. A., Thacker, S. B. et Siegel, P. Z. (2001). What's in a title? A descriptive study of article titles in peer-reviewed medical journals. *Science, 24(3)*, 75-78.

Grant, M. J. (2013). What makes a good title? *Health Information & Libraries Journal, 30(4)*, 259-260.

Gustavii, B. (2017). *How to write and illustrate a scientific paper*. Cambridge University Press.

Haggan, M. (2004). Research paper titles in literature, linguistics and science: dimensions of attraction. *Journal of Pragmatics, 36(2)*, 293-317.

Hallliday, M. A. K. et Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hartley, J. (2005). To attract or to inform: What are titles for? *Journal of technical writing and communication, 35(2)*, 203-213.

Hatier, S. (2016). Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche fouillée sur corpus d'article de recherche en SHS, Thèse de doctorat, Université Grenoble Alpes, 2016.

Hatier, S., Augustyn, M., Tran, T. T. H., Yan, R., Tutin, A., & Jacques, M. P. (2016). French cross-disciplinary scientific lexicon: extraction and linguistic analysis. Dans *Proceedings of Euralex*, 355-366.

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

Jacques, T. S. et Sebire, N. J. (2010). The impact of article titles on citation hits: an analysis of general and specialist medical journals. *Journal of the Royal Society of Medicine Short Reports, 1(1)*, 1-5.

Jamali, H. R. et Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics, 88(2)*, 653-661.

Leech, G. N. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50(4)*, 675-724.

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

Lewison, G. et Hartley, J. (2005). What's in a title? Numbers of words and the presence of colons. *Scientometrics, 63(2)*, 341-356.

Mabe, M. A. et Amin, M. (2002). Dr. Jekyll and Dr. Hyde: Author-reader asymmetries in scholarly publishing. *Aslib Proceedings, 54(3)*, 149-157.

Nagano, R. L. (2015). Research article titles and disciplinary conventions: A corpus study of eight disciplines. Journal of Academic Writing, 5(1), 133-144.

Nivard, J. (2010). *Les Archives ouvertes de l'EHESS*. Récupéré sur La Lettre de l'École des hautes études en sciences sociales n°34: http://lettre.ehess.fr/index.php?5883

Paiva, C. E., Lima, J. P. D. S. N. et Paiva, B. S. R. (2012). Articles with short titles describing the results are cited more often. *Clinics, 67(5)*, 509-513.

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. (2012). English abstract nouns as conceptual shells: From corpus to cognition (Vol. 34). Walter de Gruyter.

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

Schwischay, B. (2001). *Deux modèles de description syntaxique*. Manuscript non publié, Université de Osnabrück.

Soler, V. (2007). Writing titles in science: An exploratory study. *English for Specific Purposes, 26*, 90–102.

Swales, J. M. et Feak, C. B. (1994). *Academic Writing for Graduate Students*. Ann Arbor: University of Michigan Press.

Townsend, M. A. (1983). Titular Colonicity and Scholarship: New Zealand Research and Scholarly Impact. *New Zealand Journal of Psychology, 12*, 41-43.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Toulouse: Doctoral dissertation, Université de Toulouse II-Le Mirail.

Vendler, Z. (1968). *Adjectives and nominalizations (No. 5)*. Mouton.

Wang, Y. et Bai, Y. (2007). A corpus-based syntactic study of medical research article titles. *System, 35(3)*, 388-399.

Whissell, C. (2012). The trend toward more attractive and informative titles: American Psychologist 1946–2010. *Psychological reports, 110(2)*, 427-444.

Winter, E. O. (1992). The notion of unspecific versus specific as one way of analysing the information of a fund-raising letter. *Discourse description: Diverse linguistic analyses of a fund-raising text*, 131-170.