Table des matières

[Introduction 1](#_Toc978896)

[Le titre comme objet d’étude 2](#_Toc978897)

[Le titre, un espace crucial et court 2](#_Toc978898)

[Buts d’un titre 3](#_Toc978899)

[Syntaxes d’un titre 3](#_Toc978900)

[Lexique des titres 4](#_Toc978901)

[Sémantique du titre 5](#_Toc978902)

[Méthode, données et outils 5](#_Toc978903)

[Présentation de HAL 5](#_Toc978904)

[Présentation du corpus et des outils 6](#_Toc978905)

[Présentation de la liste des noms pivots 7](#_Toc978906)

[Hypothèses et premières analyses 7](#_Toc978907)

[Analyse lexicale des noms pivots + Présence dans les titres (1, 2, 3…) 7](#_Toc978908)

[Analyse syntaxique en dépendances 7](#_Toc978909)

[Détection des constructions 7](#_Toc978910)

[Substituabilité des termes et constructions communes 7](#_Toc978911)

[Objectifs, pistes et perspectives 7](#_Toc978912)

[La sémantique derrière les constructions 7](#_Toc978913)

[La question d’une nouveau patron définitoire des NSS 7](#_Toc978914)

[Conclusion 7](#_Toc978915)

[Bibliographie 7](#_Toc978916)

# Introduction

L’étude des titres de documents scientifiques, en particulier de ceux des articles, est doublement motivé. La première motivation est didactique, 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 ; 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 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. 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 ? Dans quelles structures 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 premier document qui clôt le premier semestre, nous commençons par exposer le titre comme objet d’étude en reprenant la littérature sur le sujet. Nous poursuivons en exposant notre méthode, nos hypothèses et nos objectifs.

# Le titre comme objet d’étude

## 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. D’autres travaux mettent la longueur en relation avec d’autres caractéristiques. 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.

## 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 »,* que Hartley juge 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 cela pourrait être pour favoriser la recherche de document, en incluant les mots clés 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, 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.

La préséance du but informatif, l’impact négatif que peut avoir sur lui le but attractif et son importance relative dans le 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 analyser sur différents plans : au niveau des syntaxes possibles, du lexique employé et des différents contenus sémantiques des titres.

## Syntaxes d’un titre

Haggan (2004) montre que seulement une faible proportion de titres étant des phrases avec un verbe conjugué dans ces 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. Soler (2007) montrent 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.

Même si la plupart des titres sont des groupes nominaux, ils peuvent être segmentés. 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 73 % des titres de 314 articles publiés en contenait contre seulement 13 % dans 474 titres d’articles non publiés. Townsend (1983) confirme cette idée en trouvant deux fois plus d’utilisations du double point dans les titres publiés que dans ceux non publiés et remarquent, comme Diers et Downs, 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, 42 % des titres en avaient au moins un et ce taux a augmenté entre 1981 et 2001.

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 lié 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. Néanmoins, Ball (2009), sur un corpus de 20 millions de titres de 1966 à 2005 en sciences dures, montrent un accroissement très important de l’utilisation de point d’interrogation dans les titres.

Utilisation du présent = Haggan (2004) « un optimisme confiant projeté par l’auteur que ce qu’il reporte sera vrai pour l’éternité ».

Hyland 2004 sur les évidentials p.147

Dillon (1981 ; 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. Townsend (1983)et Diers et Downs notent un accroissement de l’utilisation du double point dans les titres.

Jacques et Sebire (2010) étudie 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, 18 contre 9 dans le *Lancet*, 16 contre 13 pour le *British Medical Journal*, et 12 contre 10 pour le *Journal of Clinical Pathology*.

À ce premier contact humain, s’ajoute aussi la perspective de la recherche dans les bases bibliographiques à l’aide de plusieurs mots clés, qui peut s’effectuer uniquement sur le titre.

## Lexique des titres

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, sont des noms abstraits comme « effet », « rôle », « comparaison », « efficacité », des nominalisations comme « transplantation », « mutation », « traitement » ou un nom de médicament ou de maladie.

Grant (2013), dans ses conseils d’écriture d’un bon titre, insiste sur le fait qu’un titre doit être spécifique.

## Sémantique du titre

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 donné, si c’est approprié, une indication de la méthode utilisée. Jamali et Nikzad (2011) propose 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.

# Méthode, données et outils

## Présentation de HAL

Nos titres sont issus de l’archive ouverte Hyper Article en Ligne[[1]](#footnote-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]](#footnote-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]](#footnote-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.

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.

## 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 (1997) 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. 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 %), chapitre d’ouvrage (66 788, 20 %) et poster (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[[4]](#footnote-4), développé à l’Université Jean-Jaurès par Assaf 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, 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.

## Présentation de la liste des noms pivots

Critères d’inclusions

# Hypothèses et premières analyses

## Analyse lexicale des noms pivots + Présence dans les titres (1, 2, 3…)

## Analyse syntaxique en dépendances

## Détection des constructions

## Substituabilité des termes et constructions communes

# Objectifs, pistes et perspectives

## La sémantique derrière les constructions

## La question d’une nouveau patron définitoire des NSS

# Conclusion

<TODO>

# Bibliographie

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

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. In Proceedings of Euralex (pp. 355-366).

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.

1. <https://hal.archives-ouvertes.fr/> [↑](#footnote-ref-1)
2. <https://www.ccsd.cnrs.fr/> [↑](#footnote-ref-2)
3. <http://arxiv.org/> [↑](#footnote-ref-3)
4. <http://redac.univ-tlse2.fr/applications/talismane/talismane.html> [↑](#footnote-ref-4)