

MÉMOIRE DE RECHERCHE

Département des Sciences du Langage

M1 Linguistique, Informatique et Technologies du Langage

SYNTAGMES BINOMINAUX RÉCURRENTS DANS LES TITRES SCIENTIFIQUES APRÈS UN DOUBLE POINT : ESSAI D’EXPLORATION

**:**

Damien GOUTEUX  
Sous la direction de Mme Josette Rebeyrolle et M. Ludovic Tanguy

# Remerciements

J’aimerais remercier mes deux codirecteurs de recherche, Mme Josette Rebeyrolle et M. Ludovic Tanguy qui m’ont accompagné par leurs conseils et leurs encouragements tout au long de ce travail d’une année.

Je tiens également à remercier Mme Cécile Fabre et Mme Lydia-Mai Ho-Dac qui, avec M. Tanguy, ont accueilli ma démarche de reprise d’études avec intérêt et bienveillance. Cette première année de master LITL m’a permis de combiner mes deux grands intérêts que sont la linguistique et l’informatique, en me faisant arpenter de nouveaux chemins. Le rythme est soutenu, mais pour rien au monde je ne regrette ce voyage.

Je voudrais saluer mes camarades de promotion qui ont voyagé avec moi et les amies qui ont pris de mes nouvelles à chaque étape : peu importe les aléas de la route, en telle compagnie.

Et je remercie celle et ceux qui m’ont vu cheminer toute l’année au lieu de partager pleinement leurs vies. Qu’ils me pardonnent, je rentre à la maison.

« Titles consist of only a few words, but they are serious stuff. »

(Swales, 1990, p. 224)

Table des matières

[Remerciements 3](#_Toc520482046)

[Introduction 6](#_Toc520482047)

[I. Précédentes études sur les titres scientifiques 7](#_Toc520482048)

[I.1 Le titre et ses problématiques 7](#_Toc520482049)

[I.2 Traits étudiés des titres 7](#_Toc520482050)

[I.3 Caractéristiques des documents 9](#_Toc520482051)

[I.4 Corpus utilisés 10](#_Toc520482052)

[I.5 L’utilisation du double point 10](#_Toc520482053)

[II. Corpus de travail 11](#_Toc520482054)

[II.1 Présentation de HAL et extraction des données 12](#_Toc520482055)

[II.2 Traitement des données 13](#_Toc520482056)

[II.2.1 Enrichissement des données 13](#_Toc520482057)

[II.2.2 Conversions 14](#_Toc520482058)

[II.2.3 Filtrage 15](#_Toc520482059)

[II.3 Mesures du corpus 17](#_Toc520482060)

[II.3.1 Taille du corpus et types des documents 17](#_Toc520482061)

[II.3.2 Années des documents 18](#_Toc520482062)

[II.3.3 Longueurs des titres et nombre d’auteurs 19](#_Toc520482063)

[II.3.4 Domaines et nombre de domaines 21](#_Toc520482064)

[II.3.5 Marques de ponctuation et segmentation 22](#_Toc520482065)

[II.3.6 Lexique des noms communs 24](#_Toc520482066)

[III. Syntagmes et patrons 26](#_Toc520482067)

[III.1 Rappel sur les syntagmes 26](#_Toc520482068)

[III.2 Limites de notre étude 27](#_Toc520482069)

[III.3 Définition des patrons 28](#_Toc520482070)

[IV. Études des trois patrons 30](#_Toc520482071)

[IV.1 Construction itérative de trois patrons 30](#_Toc520482072)

[IV.2 Patron SN : syntagme nominal 32](#_Toc520482073)

[IV.2.A Fiche d’identité 32](#_Toc520482074)

[IV.3 Patron SP : syntagme prépositionnel 33](#_Toc520482075)

[IV.3. A Fiche d’identité 33](#_Toc520482076)

[IV.4 Patron SNC : syntagme nominal avec coordination 34](#_Toc520482077)

[IV.4. A Fiche d’identité 34](#_Toc520482078)

[IV.4.C Statistiques 34](#_Toc520482079)

[V. Étude de structures particulières 35](#_Toc520482080)

[VI. Résultats et discussions 35](#_Toc520482081)

[VI.1 Résultats 35](#_Toc520482082)

[VI.2 Discussions 35](#_Toc520482083)

[Conclusion 36](#_Toc520482084)

[Bibliographie 37](#_Toc520482085)

[Annexes 39](#_Toc520482086)

[A1. Requêtes Apache Solr sur HAL 39](#_Toc520482087)

[A1.A Requêtes 39](#_Toc520482088)

[A1.B Résultats 39](#_Toc520482089)

[A2. Codes des étiquettes de catégorie de discours de Talismane 41](#_Toc520482090)

[A3. Index des tableaux 41](#_Toc520482091)

[A4. Index des graphiques 42](#_Toc520482092)

[A5. Index des logiciels et technologies mentionnés 42](#_Toc520482093)

# Introduction

U

n titre est la porte d’entrée d’un document scientifique, que cela soit un cours, un article de recherche, un ouvrage ou une thèse. Il s’agit généralement du premier contact qu’a le lecteur avec le texte titré et parfois du seul, s’il décide ensuite de ne pas poursuivre sa lecture. Comme les titres de la presse généraliste, les titres oscillent entre deux fonctions pragmatiques, informer et attirer. La première renseigne sur le contenu du document, son champ de recherche, son sujet et parfois même ses conclusions. Elle a pour but d’aider rapidement le lecteur à décider si le document a un intérêt pour lui. La seconde fonction crée de l’intérêt : elle vise à séduire le lecteur en l’amusant ou en l’intriguant pour l’amener à vouloir continuer sa lecture.

Les mécanismes de la seconde fonction, qui peuvent aller jusqu’à masquer l’information ou l’éclairer sous un jour très particulier, sont hors de portée de notre travail. De plus, notre intuition nous porte à penser que la première fonction est bien plus présente dans les titres de textes scientifiques, la seconde n’étant présente que de façon anecdotique et même considérée comme contraire aux règles de bonnes écritures d’un titre donnés par les travaux prescriptifs.

Pour la première fonction, on peut se demander comment cette volonté d’information se traduit-elle dans la construction d’un titre en français. C’est un texte très court mais qui doit dans cet espace limité définir le sujet du document de façon complète et concise. Ceux-ci ont souvent des constructions syntaxiques spécifiques : beaucoup ne comportent pas de verbes conjugués et ne forment pas une phrase minimale. Pourtant, les titres sont parfois segmentés : par l’utilisation de points mais aussi par l’utilisation du double point, appelé aussi le ou les deux points. Ce dernier, pour de nombreuses grammaires, précède une énumération, une citation, un exemple, une cause, une conséquence, une synthèse, une description, une définition ou une explication.

Nous nous intéressons plus particulièrement à ces 4 derniers emplois : en donnant plus d’information sur ce qui le précède, le double point agit comme un marqueur où commencer à chercher cette information supplémentaire. Mais que chercher ? Les noms sont souvent considérés comme ayant le plus de contenu sémantique. Nous nous intéresserons donc aux syntagmes binominaux qui suivent immédiatement un double point dans un grand corpus de titre pour trouver les éventuelles récurrences syntaxiques et lexicales et tenter d’y apporter une explication.

La limite de deux noms est arbitraire. Nous avons l’intuition qu’elle permet déjà l’observation de phénomènes et elle vient cadrer ce travail dans les limites de l’exercice du projet de recherche du Master 1 LITL. Notre source de données sera une archive ouverte française pour avoir des titres dans cette langue, nos outils le langage de programmation Python[[1]](#footnote-1) et le logiciel de tableur Excel2.

Nous commençons par explorer des études antérieures sur les titres scientifiques pour en tirer des enseignements sur notre matière de travail. Nous constituons ensuite un important corpus de titres pour essayer de faire émerger de des récurrences. Dans un premier temps, nous les cherchons au niveau syntaxique, en nous aidant de patrons. Nous exposons en détails 3 d’entre eux. Puis, nous passons à la recherche de récurrences lexicales en essayant de fournir une explication pour chacune d’entre elles. Enfin, nous résumons et discutons nos résultats avant de conclure.

# I. Précédentes études sur les titres scientifiques

## I.1 Le titre et ses problématiques

(Hartley, 2003) et (Goodman, Thacker, & Siegel, 2001) rappellent les deux buts principaux du titre : informer et attirer. Nous nous limitons dans ce travail à la première fonction. (Mabe & Amin, 2002), interrogeant 5000 lecteurs de textes scientifiques, trouvent que 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 à son titre (Goodman, Thacker, & Siegel, 2001).

Dernier facteur d’importance du titre et de la fonction informative : il est généralement l’objet sur lequel s’effectue la recherche dans une base bibliographiques. D’où l’intérêt de privilégier les termes clés pour faciliter son indexation[[2]](#footnote-2) et de privilégier le but informatif selon (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) et (Haggan, 2004), même si la recherche en plein texte permis par les avancées technologiques rend cette affirmation moins pertinente (Goodman, Thacker, & Siegel, 2001).

Toujours selon ces trois sources, le titre est également critique bien avant sa publication, car c’est le premier élément que rencontre l’éditeur et les pairs qui décideront de l’accepter ou non.

Les titres d’articles scientifiques ont déjà été étudiés sous l’angle de différentes problématiques dans de nombreux articles. La plupart ont une articulation semblable composée de trois points successifs :

1. Obtention d’un ensemble de titres à partir de journaux scientifiques d’une ou plusieurs disciplines
2. L’analyse des titres, en en proposant éventuellement une typologie. L’analyse porte sur un ou plusieurs **traits** du titre.
3. A) Soit une étude en synchronie pour mettre en rapport cette analyse avec une ou plusieurs **caractéristiques** du document titré.

B) Soit une étude en diachronie des point analysés pour déterminer de potentielles évolutions.

A et B peuvent se combiner pour étudier une ou plusieurs caractéristiques du document titré en rapport avec les traits de son titre sur une période donnée.

Nous avons recensé, dans les articles lus, les différents traits des titres étudiés et les différentes caractéristiques des documents mises en rapport avec ceux-ci.

## I.2 Traits étudiés des titres

Un titre peut être analysé selon différents traits, nous en énumérons 6 :

* Le premier trait est la longueur d’un titre en mots, c’est le plus étudié, notamment par (Haggan, 2004), (Lewison & Hartley, 2005) et (Whissell, 2004). (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) considèrent un titre faisant plus de 20 mots comme trop longs. (Jamali & Nikzad, 2011) ne comptent que les substantifs. (Nagano R. L., 2015) comptent les mots mais ajoute un taux de substantifs. Pour le citer, « ce taux est souvent considéré comme un indicateur pour déterminer combien ce titre est informatif ».
* Le deuxième trait est le nombre de partitions ou segments, séparés par une marque de ponctuation, dans le titre. (Haggan, 2004) nomme cela des titres composés. Certains comme (Nagano R. L., 2015) décide de traiter les partitions séparément et ramènent celles-ci à une dualité titre / sous-titre. On peut compter la longueur de chacune pour les comparer ensuite.
* Le troisième trait est l’étude des marques de ponctuation qui segmentent ou terminent un titre. Dans le premier cas, la plus étudiée dans la littérature est le double point, notamment par (Dillon, 1981). (Haggan, 2004) y rajoute le point et le tiret. (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) se penchent eux sur les marques qui terminent, et plus particulièrement les points d’interrogation et d’exclamation, ainsi que les points de suspensions. S’ils admettent, comme (Jamali & Nikzad, 2011), que le point d’interrogation renforce le pouvoir d’attraction, ils mettent en garde sur le fait que l’objet principal puisse ne pas être dans le titre à la faveur d’une telle construction. Ce dernier article montre que la présence d’un point d’interrogation entraîne un nombre de téléchargement plus important mais qu’ils sont moins cités : une amélioration de l’attractivité d’un article ne garantie néanmoins pas son utilisation.
* Le quatrième trait est la présence d’acronyme. (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) mettent en garde contre leurs utilisations qui obscurcissent la compréhension du titre. Cet avis peut être remis en cause : dans un champ scientifique donné, les principaux acronymes sont connus et convoient énormément d’information en très peu de place, ce qui met en avant les autres informations.
* Le cinquième trait est la structure syntaxique du titre. (Haggan, 2004) constate que 90% des titres étudiés ne sont des unités syntactiques 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*» ici. Leech avait déjà pointé que, quoique rare à l’écrit, là où on les trouve néanmoins fréquemment est dans les titres.
* Le sixième trait est la présence de citation, détectée par la présence de guillemets. (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) classe leurs présences comme un défaut.

Certains traits sont corrélés : ainsi (Dillon J. , 1981) note que les titres incluant un double point sont plus longs, 17 mots en moyenne, que les titres n’en ayant pas, 8 mots en moyenne. Ce même compte a été fait par (Lewison & Hartley, 2005) qui trouve respectivement 14 et 11. (Jamali & Nikzad, 2011) montre également que les titres avec double point sont plus longs.

Ces traits touchent aussi bien à la syntaxe qu’à les présences de certains éléments comme les marquent de ponctuation ou les acronymes. (Jamali & Nikzad, 2011) y ajoute une dimension sémantique en classant les titres selon qu’ils indiquent seulement le sujet, pour les titres descriptifs, ou le sujet et sa conclusion, pour les titres nommés déclaratifs. Ces derniers comptent pour 46% de leur corpus. (Goodman, Thacker, & Siegel, 2001) avaient déjà classé les titres selon qu’ils contenaient différents éléments comme le sujet, la méthode employée, le jeu de données, les résultats et la conclusion. Ils trouvent que 2% des titres présentent le jeu de données, 19% présentent un résultat ou la conclusion de l’article, 33% la méthode et 40% seulement le sujet.

(Rebeyrolle, Jacques, & Péry-Woodley, 2009) apporte une dimension discursive sur les titres et intertitres des articles de la presse généraliste. Elles regardent comment ceux-ci contribue « à la construction d’un discours cohérent », les liens entre eux et le texte qu’ils chapeautent, et les divisent en deux : ceux qui gèrent les référents et ceux qui ouvrent un espace thématique. Les traits sont ensuite mis en rapport avec les caractéristiques des documents titrés.

## I.3 Caractéristiques des documents

Titre et document sont les deux faces d’une même pièce. Les caractéristiques du document apportent un éclairage supplémentaire sur son titre et permettent de les mettre en rapport avec ses traits. Nous en énumérons 6 :

* La première caractéristique est la discipline scientifique à laquelle se rattache le document. (Haggan, 2004) montre que l’utilisation de phrases complètes est un trait majeur des titres se rapportant à la biologie. On notera que les disciplines sont hiérarchisées en arbre et qu’un même document peut se rattacher à plusieurs d’entre elles. Les disciplines biologiques et médicales sont surreprésentées dans les études (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014), (Jamali & Nikzad, 2011), (Lewison & Hartley, 2005), (Goodman, Thacker, & Siegel, 2001) et (Whissell, 2004).
* La deuxième caractéristique est l’année du document. Elle peut correspondre à sa date de publication dans un journal scientifique ou de prépublication sur une plate-forme en ligne. Avec un espace temporel suffisamment grand, on peut faire une étude en diachronie sur l’évolution de certains traits des titres. C’est que fait (Dillon J. T., 1982) sur l’augmentation de l’utilisation du double point de 1880 à 1980 comme indicateur de « l’explosion des connaissances ». (Lewison & Hartley, 2005) étudient sur une période de 20 ans en prenant 5 années comme échantillon la longueur, l’utilisation du double point et le nombre d’auteurs et en ramenant ces données aux disciplines des documents.
* La troisième caractéristique est le nombre d’auteurs du document. (Lewison & Hartley, 2005) ont montré que plus il y a d’auteurs, plus le titre aura tendance à être long jusqu’à un plateau de 11 mots à partir de 4 auteurs. Ils remarquent également que certains laboratoires ont une politique très extensive des signatures comme le CERN[[3]](#footnote-3) dont les articles sont signés par plusieurs centaines de personnes.
* La quatrième caractéristique est la nationalité des auteurs, celle de la revue ou de la plate-forme où il a été publié ou prépublié.
* La cinquième caractéristique est le nombre d’accès et de téléchargements du document. Certaines plates-formes électroniques comptabilisent chaque visualisation de la notice de l’article, ouverture et téléchargement. (Jamali & Nikzad, 2011) mettent en relation le nombre de téléchargements avec la longueur du titre et la présence dans celui-ci d’un double point ou d’un point d’interrogation.
* La sixième caractéristique est le nombre de citations. Certaines plates-formes électroniques comptabilisent combien de fois l’article a été cité. (Jamali & Nikzad, 2011) mettent également en relation cette caractéristique avec les traits cités.

Tous les articles décrivent les corpus utilisés pour répondre à leurs problématiques.

## I.4 Corpus utilisés

Les articles étudiés, datés de 2004 à 2014, utilisent des corpus inférieurs à 2200 titres, le plus petit de (Haggan, 2004) démarrant à 751 titres. Seule exception, l’article de (Lewison & Hartley, 2005) culmine à 349 700 titres. La taille du corpus est importante car plus elle est grande, plus l’on trouvera un nombre de phénomènes importants, avec pour chacun un nombre d’occurrence suffisant pour que celui ne soit pas un cas trop marginal pour être intéressant.

Les titres sont piochés directement dans des journaux scientifiques renommés, entre 1 seul pour (Whissell, 2004) et 44 pour (Haggan, 2004). L’article de (Rebeyrolle, Jacques, & Péry-Woodley, 2009) se détache en piochant ses titres dans 6 journaux non scientifiques d’information, nationaux ou régionaux. L’article de (Lewison & Hartley, 2005) interroge de son côté une base de données de titres, le *Science Citation Index* qui contient de nombreuses revues, comme celui de (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) qui interroge la base MEDLINE regroupant plus de 500 revues.

Certains des travaux précédents font le choix de piocher ces titres dans des disciplines proches, comme la biologie et la médecine pour (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) pour augmenter le volume de leur corpus. D’autres choisissent des disciplines qu’ils jugent très éloignés, comme littérature contre sciences dures (Haggan, 2004) pour comparer les traits de leurs titres.

Il est à noter que si le contenu des articles est parfois inaccessible, surtout quand le paysage était dominé par les grands éditeurs de publications scientifiques, les titres des articles sont eux toujours accessibles gratuitement et donc facile à acquérir.

Nous retenons qu’il est essentiel d’utilisé un grand corpus et pour cela le recours à des traitements automatiques sera nécessaire, et que nous devons les piocher dans plusieurs disciplines, afin de pouvoir les comparer sur les traits étudiés dans les titres.

## I.5 L’utilisation du double point

Nous l’avons vu, la présence d’un double point tend à être corrélé à une longueur plus grande des titres. Les grammaires, dont (Grevisse & Goosse, 2011) et (Doppagne, 1998), montrent que le double point introduit une énumération, une citation, un exemple, une cause, une conséquence, une synthèse, une description, une définition ou une explication. Les 4 derniers points nous intéressent plus particulièrement car il ajoute des informations sémantiques très importantes à ce qui vient d’être dit.

(Haggan, 2004) remarque dans les titres scientifiques, il y a une haute fréquence d’utilisation des titres a deux segments autour d’un double point. Elle retrace cet usage à une stratégie d’écriture des titres scientifiques définie dans (Lester, 1993). Celle-ci utilise un premier syntagme nominal indiquant le domaine de recherche, le double point, puis un second syntagme nominal permettant de situer l’article dans ce domaine, soit en mentionnant son point de départ, soit son point d’arrivée, c’est-à-dire sa conclusion. La juxtaposition de ces deux informations par le double point rend plus facile leur interprétation. Haggan appelle cette construction un resserrement (*« narrowing »*). (Goodman, Thacker, & Siegel, 2001), qui ont recueilli les consignes données par les éditeurs de 4 journaux médicaux pointe qu’un d’entre eux encourageait l’utilisation du double point.

(Haggan, 2004) remarque plus globalement que les titres composés correspondent le plus souvent à deux segments séparés par un double point. Dans les trois catégories de disciplines qu’elle étudie, elle note qu’ils représentent 61% des titres en littérature, 30% en linguistique et 21% en sciences, mais sans différencier particulièrement ceux utilisant le double point pour la segmentation d’autres marques de ponctuation.

Une particularité qu’elle relève, surtout en littérature, est que la citation peut être *avant* le double point et non *après* comme dans cet exemple : *« I Fought the Law (and I cold won) » : Hip-hop in the mainstream*. Elle constate alors deux possibilités pour les auteurs. Celle de faire preuve de créativité en juxtaposant une seconde partie pertinente qui éclaire la citation, en citant l’auteur ou l’œuvre, où finalement le « véritable titre » est cette seconde partie. L’autre est d’utiliser une seconde partie plus obscure, visant à soumettre un « puzzle élégant » au lecteur l’incitant ainsi à le résoudre en lisant l’article, mais cela se rapproche de l’attractivité plutôt que de l’information.



Notre travail s’intéresse à l’aspect informatif des titres et nous n’explorons pas la fonction attractive. (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014), l’article le plus prescriptif que nous ayons étudié, indique combien la dimension informative doit primer sur celle de l’attractivité. Néanmoins, notre travail ne se place pas dans ce cadre de prescription mais de description de l’usage. Nous avons choisi également de pas nous intéresser aux sous-titres pour nous concentrer sur le titre. Nous avons choisi de les étudier globalement, sans éclater leurs segments.

Il est clair, à la lecture des travaux précédents que notre corpus doit être important, propre à contenir une grande variété de productions langagières et donc à l’observation et à la quantification d’un phénomène particulier. Toutes les études présentées traitaient des titres d’articles scientifiques en anglais, sauf (Rebeyrolle, Jacques, & Péry-Woodley, 2009) qui traitait elle des titres d’articles en français mais de la presse généraliste. S’intéresser aux titres de documents scientifiques en français est donc un apport de notre travail. Nous allons maintenant présenter à la construction de notre corpus de travail.

# II. Corpus de travail

Dans cette partie nous présentons notre corpus de travail et la méthode suivie pour l’obtenir. Nous commençons par présenter son origine et le travail d’extraction qui a été fait. Nous abordons après les traitements effectués sur les données brutes pour aboutir à notre corpus ainsi que quelques exemples de titres. Nous présentons ensuite l’outillage utilisé et les premières constatations effectuées sur notre corpus.

## II.1 Présentation de HAL et extraction des données

Nos titres sont issus de l’archive ouverte Hyper Article en Ligne[[4]](#footnote-4) (HAL) (Nivard, 2010). Elle compte, au 14 juillet 2018, 524 452 documents scientifiques et 1 563 014 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, 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[[5]](#footnote-5) (CCSD), fondé en 2000 et rattaché au Centre National pour la Recherche Scientifique (CNRS). Il existe des sous-ensembles de HAL dédiés à une discipline spécifique, 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[[6]](#footnote-6) , fondée en 1991 et limitée uniquement aux articles. Un dépôt d’un article dans HAL entraîne automatiquement la création d’une notice dans arXiv si elle 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.

Pour récupérer ces notices, il existe deux protocoles. Le premier est le protocole de moissonnage standardisé Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), version 2.0 . Ce protocole est standardisé, on peut donc accéder à d’autres archives ouvertes avec, dont arXiv, mais il est assez limité dans ses possibilités de sélection : de base, seule la possibilité de sélectionner des notices en fonction de la date est offerte. Les gestionnaires d’archives peuvent définir des sélecteurs supplémentaires mais ils sont encore très limités. Ce protocole est surtout conçu pour les robots s’assurant de la réplication des notices entre différentes archives ouvertes.

Le second protocole offert pour accéder à HAL est bien plus intéressant dans notre cas. Il permet de sélectionner finement nos données et repose sur Apache Solr (Smiley, Pugh, Parisa, & Mitchell, 2015), le moteur de recherche du projet Apache Lucene[[7]](#footnote-7). On peut sélectionner les notices à retourner en filtrant sur le contenu d’une des métadonnées, on peut définir le format des données en sortie parmi un large choix de standards (dont XML, CSV ou JSON). C’est ce second protocole et le format de sortie JSON, avec encodage des caractères en UTF-8, que nous utiliserons. JSON est un format simple de données structurées stockées dans un fichier texte dont les constituants sont les données atomiques (chaîne de caractères, nombres), les listes et les dictionnaires associant une clé à une valeur (Bray, 2017). Une requête Solr repose sur le protocole de transfert hypertexte (HTTP). Elle se présente donc sous la forme d’une adresse internet (URL) qui peut être testée de façon simple et rapide dans un navigateur internet pour visualiser son résultat.

Nous avons automatisé, à l’aide d’un script Python, la création et l’envoi de requêtes ainsi que la récupération et la sauvegarde des résultats retournés au format JSON. En une demi-heure, nous avons récupéré 304 600 titres ainsi que leurs métadonnées comme l’année de dernière modification de la notice, les domaines scientifiques associés au document, ses auteurs, son type et son identifiant unique. Nous présentons dans l’[annexe A1](#_A1._Requêtes_Apache) l’ensemble des requêtes utilisées.

## II.2 Traitement des données

Une fois les données brutes obtenues, nous avons effectués plusieurs traitements dessus qui se divisent en trois catégories : l’enrichissement des données, la conversion et le filtrage. Notons que l’avantage des titres est qu’ils résultent d’une production très travaillée par rapport à celle que d’un tweet, d’un SMS ou d’un texte sur un forums, soignée et relue. La présence de faute de grammaire ou d’orthographe est donc quasi-nulle, ce qui évite d’avoir à les considérer en ajoutant des traitements de tolérance ou de réparation dans les algorithmes.

### II.2.1 Enrichissement des données

Nous voulions enrichir nos titres en déterminant pour chaque mot son lemme et sa catégorie grammaticale à partir de la forme présente dans le titre. Dans un premier temps, nous avons utilisé le logiciel Stanford Core Natural Language Processing[[8]](#footnote-8). Celui-ci fournissait pour le français les catégories mais non les lemmes. Nous avons donc abandonné Stanford Core NLP pour passer à un logiciel développé à l’Université Jean-Jaurès, Talismane[[9]](#footnote-9) par Assaf Urieli (Urieli, 2013).

Avoir le lemme d’un mot permet de rassembler toutes ses formes fléchies sous une même entrée et de compter le nombre d’occurrences de celle-ci ensuite. La catégorie, ou étiquette POS pour *part of speech*, permet de percevoir la structure syntaxique qui est derrière les formes.

À chaque fois, nous avons conçu un script Python qui envoyait le titre brut à Stanford Core ou Talismane et récupérait le résultat du traitement. Nous n’avons pas comparé les résultats des deux en ce qui concerne les catégories pour tenter d’améliorer la fiabilité des résultats, cette question s’éloignant trop de nos priorités. Une fois les catégories et les lemmes obtenues, nous procédions à l’enregistrement de notre corpus, par un traitement rattaché à ceux que nous qualifions de conversions.

### II.2.2 Conversions

Tout au long de notre travail, il nous a fallu récupérer des données dans un format donné et les sauvegarder dans un autre. Le premier traitement de conversion a été la transformation des données au format JSON récupérées de HAL vers un format XML propre à notre travail. Le second s’occupait de la conversion des résultats obtenus auprès de Talismane puis de la sauvegarde des catégories et des lemmes en enrichissant notre format XML.

Nous n’avons pas utilisé des standards reconnus comme CoNLL-U[[10]](#footnote-10), issu de la conférence du même nom, ou TEI P5 de la communauté Text Encoding Initiative pour deux raisons. La première c’est que le contenu textuel d’un titre est très court mais nous en avons énormément. TEI P5 nous semble plus adapté pour encoder de véritables textes et CoNLL-U n’utilise pas XML mais un format texte utilisant les lignes et les tabulations pour traduire la structure des données. Nous souhaitions maîtriser notre format, sachant que celui-ci utilise XML, sa conversion vers un autre format basé sur XML ne poserait pas de problème.

Exemple de données récupérées auprès de HAL au format JSON ( { } indiquant un dictionnaire qui associe une clé à une valeur, [ ] indiquant une liste de valeurs et " " une chaîne de caractères) :

**{**

"docid": 1712921,

"domain\_s": **[** "0.shs", "1.shs.phil" **]**,

"title\_s": **[** "La logique de l'action de Michael Quante",

"Michael Quante on Logic and Action" **]**,

"authFullName\_s": **[** "Alain Patrick Olivier" **]**,

"language\_s": **[**"fr"**]**,

"docType\_s": "ART",

"modifiedDateY\_i": 2018

**}**

HAL nous donne une notice de document avec ses métadonnées. Idéalement, il y a une 1 notice pour 1 document qui possède 1 titre. Cette vision sera remise en cause plus fortement dans la partie suivante mais nous pouvons déjà voir que le champ titre correspond à une liste et que cette liste contient pour cette notice deux éléments. Le premier titre est en français et le second est sa traduction en anglais.

Malheureusement, il y a une certaine hétérogénéité dans le remplissage des différents champs mis à disposition par HAL. Pour certaines notices, le champ titre ne possède qu’un seul élément, mais il s’agit d’une concaténation du titre français avec le titre anglais, avec entre un marqueur qui n’est pas standardisé, certains utilisant « / » ou « [ » et d’autres « Titres en anglais : ». Cela nous amène à la nécessité de filtrage des données qui sera exposée dans le chapitre suivant.

Une fois transformé en XML et enrichi des catégories et lemmes, ce même titre se présente ainsi, les balises ouvrantes < > et fermantes </ > structurant les données :

<notice>

<id>1712921</id>

<type>ART</type>

<date>2018</date>

<text>La logique de l'action de Michael Quante</text>

<words>

<word><form>La</form><lemma>la</lemma><pos>DET</pos></word>

<word><form>logique</form><lemma>logique</lemma><pos>NC</pos></word>

<word><form>de</form><lemma>de</lemma><pos>P</pos></word>

<word><form>l'</form><lemma>le</lemma><pos>DET</pos></word>

<word><form>action</form><lemma>action</lemma><pos>NC</pos></word>

<word><form>de</form><lemma>de</lemma><pos>P</pos></word>

<word><form>Michael</form><lemma>Michael</lemma><pos>NPP</pos></word>

<word><form>Quante</form><lemma>\_</lemma><pos>NPP</pos></word>

</words>

<authors>

<author>Alain Patrick Olivier</author>

</authors>

<domains>

<domain>0.shs</domain>

<domain>1.shs.phil</domain>

</domains>

</notice>

Un titre a donc un identifiant correspondant à celui partagé par la notice et le document, un type correspondant à celui du document et une date. Cette date est issue de la dernière date de modification du dépôt du document. Nous avons pris cette date car c’était la seule systématiquement remplie sur tous les types de document de HAL de façon cohérente et qui indique la date de création du document scientifique.

Le titre en lui-même est présent sous une forme complète et une forme décomposée en mots, avec pour chaque mot la forme fléchie présente dans le texte, son lemme et son étiquette POS. Lorsque Talismane n’arrive pas à déterminer le lemme d’un mot, il indique ‘\_’ pour son lemme. La principale difficulté de cette étape est la gestion des caractères spéciaux dans les textes qui ont une signification spécifique pour XML comme « < », « > » ou « & ». Pour éviter cela, ils sont échappés, c’est-à-dire remplacés par un caractère neutre, lorsqu’ils sont sauvegardés sur le disque. Le caractère spécial est correctement restitué lorsque nous rechargeons en mémoire le corpus.

Enfin, nous avons la liste des auteurs et des disciplines scientifiques, appelées domaines dans HAL, auxquelles se rattache le document. On notera que les disciplines sont organisées en arbre, le chiffre avant son nom indiquant son niveau dans celui-ci, et qu’un même article peut être étiqueté sous plusieurs disciplines. Le titre présenté dans cette partie est néanmoins le fruit d’une autre catégorie de traitement : le filtrage effectué sur les données.

### II.2.3 Filtrage

Tout au long de notre travail, il a été nécessaire d’appliquer certains filtres à nos données. Tout d’abord, nous avons remarqué des incohérences dans nos données. Par exemple, certains documents avec plusieurs notices, nous avions donc des titres en double. D’autres, nous l’avons vu, concaténaient un titre français et un titre anglais, ou même étaient en anglais. Beaucoup avaient deux titres, un en français et sa traduction en anglais. Un autre cas, bien plus rare, était la présence d’un titre dans une autre langue.

Nous avons donc appliqué plusieurs filtres à nos données : en regardant le champ langue de la notice retournée par HAL, nous avons éliminé toute celle qui avaient plus d’un langage ou qui avait un langage qui n’était pas le français. Notre raisonnement était simple : si un document scientifique est en français, comme indiqué dans sa notice, son titre sera en français. Nous avons également gardé qu’un seul titre par notice s’il y en avait plusieurs, le premier, qui, d’après nos constations visuelles sur toutes les notices présentant cette caractéristique, était toujours celui en français.

En construisant un premier lexique des formes utilisés dans les titres nous avons pourtant remarqué une forte fréquence de « on », « and », « a », « in », « the », « und ». Les formes  « and », « in » et « the » appartiennent indiscutablement à l’anglais, tandis que « on » et « a » peuvent t’appartenir à l’anglais ou au français. Enfin « und » appartient lui à l’allemand. Pour mieux filtrer nos titres, utilisé un programme de détection automatique des langues écrits en Python appelé langdetect[[11]](#footnote-11) en gardant que les titres qu’il estimait être en français. Nous avons ainsi supprimé 12 205 titres.

Nous avons aussi supprimé certains titres car leurs notices nous semblaient incohérentes : 33 n’avaient pas d’auteurs, 6448 n’avaient pas de domaines associés, 1 n’avait de type de document et 7096 étaient des doublons. D’autres notices, au nombre de 11, avaient un titre vide et ont également été supprimées. À la fin, nous avions un corpus général de **278 806** titres, ce qui reste un nombre assez conséquent pour étudier un phénomène linguistique particulier dans celui-ci.

Enfin, nous avons appliqué la restriction découlant de notre problématique : nous voulions étudier les structures lexico-syntaxiques après un double point, il nous fallait donc extraire un sous-corpus, spécialisé pour notre problématique. Nous avons choisi de ne prendre que les titres contenant qu’un et un seul double point, considérant que les titres ayant plusieurs doubles points relevaient de cas très particuliers et complexes qui dépassent le cadre de notre travail. Un rapide décompte nous donne :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Nombre de « : »** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Nombre de titres | 190 123 | **86 095** | 2268 | 258 | 40 | 13 | 4 | 2 | 1 | 2 |

Tableau : nombre de doubles points dans les titres

Il y a donc 86095 titres avec un et un seul double point, soit 31% des titres de notre corpus général. Ce tableau à l’avantage de nous montrer que des cas très particuliers existent, comme deux titres avec 9 doubles points dedans, mais qu’ils sont également très rares. En les écartant, il s’agit de nettoyer nos données car nous ne nous intéressons pas à ces cas à la marge. C’est le même souci qui nous amène à considérer le nombre de mots[[12]](#footnote-12) après l’unique double point de ces 86 095 titres pour obtenir le graphisme suivant :

On voit que 99% ont entre 0 et 29 mots après le double point. Nous écartons donc tous les titres en ayant plus (466 titres) ainsi que ceux en ayant 0 (98 titres). Une explication possible pour ces derniers était que le double point annonçait un sous-titre mais que nous ne l’avons pas récupéré de HAL. Avec le recul, nous aurions pu le concaténer mais encore une fois, il s’agit d’un traitement supplémentaire pour récupérer seulement 98 titres. Nous obtenons finalement notre corpus de travail de **85 531** titres, soit 31% des 278 806 titres du corpus général.

## II.3 Mesures du corpus

Notons que, comme le souligne Cori et David (Cori & David, 2008), l’étape de sélection des données pour constituer un corpus peut comporter une part de subjectivité. Dans notre cas nous nous en exemptons car nous n’avons pas opéré pas de choix dans les titres que nous retournent HAL autres que ceux visant à assurer la cohérence de nos données. Pour ces choix, nous avons suivi la précaution méthodologique préconisée par ces deux auteurs, de faire *« un inventaire soigneux de toutes les décisions prises en amont »* que nous avons exposées dans ce document. Nous pouvons à présent observer notre corpus de travail sous plusieurs angles. Les exemples de titre cités pour éclairer notre propos sont accompagnés du nom de leur auteur ainsi que de l’année de publication du document mais ces références ne sont pas reprises dans les références bibliographiques.

### II.3.1 Taille du corpus et types des documents

Notre corpus général se rapproche de celui utilisé par (Lewison & Hartley, 2005) qui comportait 349 700 titres, tous les autres corpus des articles étudiés ne dépassaient pas les 2200 titres. Notre corpus de travail reste toujours bien au-dessus de ce seuil avec 84 293 titres. De plus notre corpus est constitué de titres de documents scientifiques en français, au contraire des articles étudiés, composés de titres en anglais ou de titres en français mais de la presse généraliste pour (Rebeyrolle, Jacques, & Péry-Woodley, 2009).

Nous nous intéressons tout d’abord au type des documents scientifiques titrés. En effet, notre corpus est différent car il contient d’autres types que des articles. Nous prenons comme hypothèse que la façon d’écrire un titre ne changent pas entre les différents types de documents. Pour vérifier cette hypothèse, il faudrait comparer les traits des titres de nos articles à ceux des aux autres types de documents.

Nous indiquons ici seulement les 7 types ayant le plus de titres ce qui représentent 93% du corpus de travail. Nous comparons chaque pourcentage avec celui de ce même type dans l’ensemble des notices de HAL pour essayer de mesurer la représentativité de notre corpus.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Nombre** | **% dans corpus** | **% dans HAL** |
| Article | 25 648 | 30% | 49.65% |
| Communication | 19 966 | 23% | 28.68% |
| Chapitre | 12 007 | 14% | 7.69% |
| Thèse | 10 632 | 12% | 5.02% |
| Mémoire | 6 716 | 8% | < 0.1% |
| Ouvrage | 2 281 | 3% | 1.8% |
| Autre publication | 2 147 | 3% | 2.29% |

Tableau : Répartition des titres par type

On remarque que ces 7 premiers types de document sont des documents textes. Les vidéo et les sons ne représentent que 0,32% des titres récupérées. On remarque également que notre corpus n’est pas tout à fait représentatif de HAL : il compte moins d’article, presque la moitié des documents dans HAL, au profit des chapitres d’ouvrage et des thèses, deux fois plus nombreux dans notre corpus, et des mémoires d’étudiants, surreprésentés : 8% contre moins de 0.1% dans HAL.

Martin (2002) pose comme notion fondamentale qu’« en raison de sa finitude, le corpus ne réalise donc qu'une part infime de ce qui est réalisable. (…) Et en toute rigueur, une grammaire construite à partir d'un corpus ne vaut que pour le corpus qui l'a produite. » Ainsi nos résultats ne pourront donc pas être élargis directement à l’ensemble des titres de HAL.

### II.3.2 Années des documents

Nous indiquons ici seulement les 8 années ayant le plus de titres ce qui représentent 99% du corpus de travail.

|  |  |  |
| --- | --- | --- |
| **Année** | **Nombre** | **%** |
| 2018 | 54 627 | 64% |
| 2017 | 21 658 | 25% |
| 2016 | 3 970 | 5% |
| 2015 | 1 996 | 2% |
| 2014 | 1 156 | 1% |
| 2013 | 817 | 1% |
| 2012 | 255 | < 0.1% |
| 1988 | 75 | < 0.1% |

Tableau : Répartition des titres par année

**Exemple :** *QUELQUES REMARQUES SUR L'ÉTUDE DE CH. BARTHEL: INFLUENCE DES MOISISSURES SUR LES FERMENTS LACTIQUES* (Dr Jaroslav Dvorak, 1925)

L’exemple ci-dessus est le plus vieux titre de notre corpus. On remarque déjà une utilisation du double point.

Si les documents vont de 1925 à 2018, les années les plus récentes sont les plus fournies : 2018 et 2017 représentant à elles seules 89% de corpus. Cette fenêtre trop réduire ne nous permet pas d’étudier en diachronie l’évolution des phénomènes autour du double point.

Comme explication, on peut avancer qu’autant la diffusion sur HAL de nouveaux articles sert directement les chercheurs, et la pratique tend à se généraliser dans le monde de la recherche en France, autant la mise en ligne d’anciens articles d’eux-mêmes ou d’autres auteurs est une tâche longue et moins gratifiante.

### II.3.3 Longueurs des titres et nombre d’auteurs

Nous utilisons Talismane pour séquencer notre titre en formes. Par « longueur en mots » nous entendons compter toutes les formes séquencées sauf celles ayant l’étiquette « PONCT », désignant une marque de ponctuation.

**Exemple :** *L'interprétation langue vocale* ***(****LV****)/****langue des signes* ***(****LS****)*** *et la question du* ***"*** *lexique* ***" :*** *inverser le regard* ***!*** *(Brigitte Garcia, 2018)*

Dans l’exemple ci-dessus, nous indiquons en rouge les éléments non pris en compte pour le calcul de la longueur. Le premier guillemet simple de l’élision est absorbé dans un seul élément, l’article défini élidé. Ce titre a donc une longueur de 17.

La moyenne de la longueur en mots des titres est de 15,5 mots, pour des titres qui vont de 2 à 69 mots. La médiane est 14,5 mots, le premier quartile est 10,5, le dernier est 18.5. Le graphisme suivant représente le nombre de titres par longueurs :

Nous regardons à présent le nombre d’auteurs par document scientifique. Les nombres les plus fréquents sont exprimés dans le tableau suivant :

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Nombre**  **d’auteurs** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| Nombre de  titres (%) | 59 182  69% | 12 035  14% | 6 015  7% | 3 310  4% | 1 765  2% | 1 065  1% | 689  1% | 415  < 0.5% |

Tableau : Nombres de titre par nombres d'auteurs

Si le nombre d’auteurs dans notre de corpus va de 1 à 147, 98% des titres ont néanmoins entre 1 et 7 auteurs et 69% ont un seul auteur, pour une moyenne de 1,8 auteurs par titre.

(Lewison & Hartley, 2005) ont montré que plus il y a d’auteurs, plus le titre aura tendance à être long jusqu’à un plateau de 11 mots à partir de 4 auteurs. Nous calculons la longueur moyenne des titres par nombres d’auteurs :

Nous devons mettre en garde sur cette première courbe qui semble erratique : à partir de 28 auteurs, le faible nombre de titres, entre 1 et 3 par nombres d’auteurs, ne nous permet pas d’affirmer quelque chose. Si on recalcule cette courbe sur les articles ayant de 1 à 8 auteurs on obtient :

On retrouve un plateau, mais entre 17 et 18 mots et non 11, et à partir de 7 auteurs et non 4. À partir de 2 auteurs, l’augmentation de la longueur moyenne est constante, mais les titres avec 1 auteur ont une longueur moyenne un peu plus importante de 1 que ceux avec 2 auteurs qui s’établit à 16.

(Lewison & Hartley, 2005) et (Dillon J. , 1981) affirmaient que la présence d’un double point augmentait la longueur du titre, de 8 à 17 pour Dillon. Nous avons calculé la longueur moyenne des titres, en comptant les formes et en ignorant les marques de ponctuation, et le nombre de doubles points dans le titre sur notre corpus général :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Nombre de « : »** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Nombre de titres | 190 123 | **86 095** | 2268 | 258 | 40 | 13 | 4 | 2 | 1 | 2 |
| Longueur moyenne des titres | 12 | **16** | 24 | 32 | 52 | 99 | 97 | 75 | 90 | 87 |

Tableau : Nombre de titres par nombres de doubles points

Notre corpus respecte cette affirmation : la présence d’un double point ou plus augmente la longueur moyenne des titres. On peut remarquer que plus y’a de doubles points, plus la longueur moyenne augmente jusqu’à 5. Au-delà, le faible nombre de titres ne permet pas d’avoir une idée fiable de la longueur moyenne, d’autant plus qu’il s’agit de constructions très particulières.

### II.3.4 Domaines et nombre de domaines

Nos documents se répartissent en 13 domaines de premier niveau. Certains titres sont apparentés à plusieurs domaines.

|  |  |  |  |
| --- | --- | --- | --- |
| **Domaine** | **Nombre** | **Domaine** | **Nombre** |
| Sciences de l’Homme et Société | 80 441 | Sciences cognitives | 1 354 |
| Sciences du Vivant | 15 614 | Chimie | 1 051 |
| Informatique | 5 244 | Mathématiques | 877 |
| Sciences de l’ingénieur | 4 469 | Économie et finance quantitative | 283 |
| Sciences de l’environnement | 3 414 | Statistiques | 228 |
| Planète et Univers | 2 157 | Science non linéaire | 22 |
| Physique | 2 013 |  |  |

Tableau : Répartition des titres par domaines

**Exemple**: *L'évaluation dans les environnements ouverts massivement multi-apprenants : une opportunité historique pour le développement la recherche fondamentale à visée pragmatique en pédagogie universitaire (Heutte Jean, 2018)*

Ce titre a été classé dans les trois domaines suivants :

|  |  |  |
| --- | --- | --- |
| **Niveau** | **Code** | **Intitulé** |
| 0 | shs | Sciences de l'Homme et Société |
| 1 | shs.edu | Sciences de l'Homme et Société/Education |
| 1 | shs.psy | Sciences de l'Homme et Société/Psychologie |

On voit que les domaines Éducation et Psychologie sont des domaines fils de Sciences de l’Homme et de Société.

Si on fait une dichotomie entre les titres référençant les Sciences de l’Homme et Société et ceux ne le faisant pas, on prend mieux en compte le poids très important de ceux-ci dans notre corpus : 61 252 contre 24 279, soit 72% et 28% respectivement. Si l’on regarde les autres domaines, on voit que cette dichotomie reprend celle entre sciences « dures » et « soft ».

### II.3.5 Marques de ponctuation et segmentation

Nous nous intéressons à ces marques sur deux points :

* Si elles sont en dernière position, elle transforme le titre en une interrogation ou une exclamation, ou du moins son dernier segment s’il y en a plusieurs.
* Si elles ne sont pas en dernière position, elles divisent le titre en segments pour les marques suivantes : « … », « : », « ; », « . », « ? », « ! » listées par (Haggan, 2004) auxquelles nous avons rajouté « … ».

**Exemple** : *Dynamique des structures* ***:*** *méthodes approchées, cinématiques* ***;*** *Analyse Modale* ***;*** *Recalage de Modèle (Claire Heinisch, 2018)*

Ce titre complexe est composé de 4 partitions, nous avons indiqué les trois marques de ponctuation qui le segmente en rouge. On peut estimer qu’il y a une « force de segmentation » associée à chaque marque de ponctuation. Intuitivement, le point-virgule semble ici établir la partition la plus forte, le double point crée une partition moyenne et enfin la virgule, qui ne crée pas de segmentation, mais ponctue une énumération.

Le tableau suivant compte combien de titres possèdant *au moins une fois* la marque de ponctuation indiquée. On détermine ensuite sur ce nombre, combien on cette marque en dernière position. Enfin, on calcule, pour les titres possédant une marque donnée, la moyenne des occurrences de cette marque.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Marque de ponctuation** | **Nombre de titres** | **%** | **Dernière position** | **%** | **Moyenne** |
| Double point | 85 531 | 100% | 0 | 0% | 1.00 |
| Point | 9 609 | 11% | 4 514 | 47% | 1.44 |
| Point d’interrogation | 9 033 | 11% | 7 579 | 84% | 1.02 |
| Guillemet français ouvrant « | 3 493 | 4% | 0 | 0% | 1.09 |
| Guillemet français fermant » | 3 708 | 4% | 0 | 0% | 1.09 |
| Guillemet anglais ouvrant “ | 447 | 0.5% | 0 | 0% | 1.14 |
| Guillemet anglais fermant ” | 460 | 0.5% | 0 | 0% | 1.13 |
| Guillemet droit " | 3 508 | 4% | 1690 | 48% | 1.80 |
| Point d’exclamation | 360 | 0.4% | 199 | 55% | 1.02 |
| Point-virgule | 341 | 0.4% | 7 | 2% | 1.28 |

Tableau : Titres avec un caractère segmentant dans notre corpus

Les guillemets français et anglais présentent l’avantage d’être différenciés entre l’ouvrant et le fermant. Comme ils doivent venir toujours par deux, nous remarquons qu’il y a un problème de cohérence dans les deux, il manque 15 guillemets ouvrants français et 13 guillemets ouvrants anglais.Le nombre de titres concernés est néanmoins très faible.

On remarque que le guillemet français est privilégié mais que l’influence anglo-saxonne n’est pas négligeable : sur la totalité des guillemets associés à une langue, ils représentent 12%. Étrangement, ces guillemets ne terminent jamais un titre, alors que 48% des titres ayant au moins un guillemet droit en a un en position terminale.

On remarque les moyennes sont très proches de 1, ce qui signifie que si un titre possède une marque de ponctuation donnée, celle-ci n’est présente qu’une seule fois dans presque tous les cas. On peut donc considérer que la présence d’un point d’interrogation ou d’exclamation transforme le titre en question, ou du moins son segment terminal.

Certaines de ces marques, lorsqu’elles sont à l’intérieur du titre, nous permettent de calculer le nombre de segments. Tous les titres comportent au moins 1 double point et donc deux segments. La moyenne s’établie à 2,14 segments par titre avec 91% des titres en ayant 2.

**Exemple**: *L'apprentissage sur le tas et la formation aux métiers de l'artisanat au Maroc****:*** *cas de la dinanderie, de la poterie et de l'ébénisterie-marqueterie (Sofiane Maïza, 2018)*

Ce titre présente une organisation typique en deux segments séparés par un double point en rouge. Nous notons qu’il ne met pas d’espace avant celui-ci, à la manière de l’anglais, alors que les règles typographiques françaises un impose un normalement. Dans les deux, l’utilisation d’une majuscule après est prohibée, sauf en cas de noms propres.

(Haggan, 2004) avait montré que l’utilisation d’une phrase complète pour titre était une caractéristique des titres en biologie. Nous divisions notre corpus de travail en deux : les titres ayant la biologie comme domaine et ceux ne l’ayant pas. Nous calculons ensuite le nombre de titre avec au moins un verbe conjugué, en ne comptant pas les participes passés et présents :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Domaines** | **Titres avec verbe conjugué** | **%** | **Titres sans verbe conjugué** | **%** |
| Titres en biologie | 1 004 | 8% | 11 276 | 92% |
| Titres non en biologie | 5 416 | 7% | 67 835 | 93% |

Tableau : Phrase complète dans les titres en fonction du domaine de la biologie

L’affirmation n’est pas confirmée sur notre corpus de titres français : il y a proportionnellement très légèrement plus de phrases complètes en biologie, mais cet écart n’est pas assez significatif et bien loin de la proportion d’un titre sur deux détectée par (Haggan, 2004) sur les titres d’articles scientifiques en anglais.

### II.3.6 Lexique des noms communs

Nous avons recensé 486 198 noms communs dans notre corpus. Nous avons calculé les plus fréquents mais nous avons aussi compté le nombre avant et après le double point segmentant nos titres. Le tableau 6 présente les 8 plus fréquents :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lemme** | **Occurrences** | **% du nombre total de noms** | **avant le « : »** | **%** | **après le « : »** | **%** |
| étude | 6089 | 1,25% | 1792 | 29% | 4297 | 71% |
| cas | 4631 | 0,95% | 227 | 5% | 4404 | 95% |
| approche | 3036 | 0,62% | 678 | 22% | 2358 | 78% |
| analyse | 3001 | 0,62% | 1114 | 37% | 1887 | 63% |
| application | 2982 | 0,61% | 258 | 9% | 2724 | 91% |
| siècle | 2766 | 0,57% | 1059 | 38% | 1707 | 62% |
| pratique | 2609 | 0,54% | 986 | 38% | 1623 | 62% |
| exemple | 2291 | 0,47% | 136 | 6% | 2155 | 94% |

Tableau : Comptes des noms communs avant et après le double point

On le voit, certains noms parmi les plus fréquents ne se retrouvent largement qu’après le double point : c’est le cas de « cas » à 95%, « exemple » à 94% et « application » à 91%. On peut donc observer, si on a une fréquence importante du nom, une affinité de celui-ci pour une position avant ou après le double point.

**Exemple :** *Sources d’informations pour l’adaptation des traitements médicamenteux chez les patients atteints d’une maladie rénale chronique* ***: état*** *des lieux des pratiques et difficultés des médecins généralistes savoyards (Mélanie Grué, 2018)*

Dans l’exemple précédant, « état » est en position 1 après le double point.

Si on fixe arbitrairement un taux minimum de 70% après le double point et une fréquence minimum de 500, c’est-à-dire plus de 0.1% de la totalité des noms communs, on obtient une classe de 26 noms ayant une grande affinité pour une position après le double point. On a également calculé la position moyenne où se situe ce nom par rapport au double point, 1 signifiant juste après le double point :

|  |  |  |  |
| --- | --- | --- | --- |
| **Lemme** | **Occurrences** | **% après le double point** | **Position moyenne** |
| cas | 4631 | 95% | 3 |
| exemple | 2291 | 94% | 3 |
| application | 2982 | 91% | 2 |
| résultat | 627 | 89% | 4 |
| perspective | 1317 | 86% | 4 |
| proposition | 468 | 85% | 3 |
| enjeu | 2133 | 83% | 3 |
| réflexion | 732 | 83% | 3 |
| enquête | 775 | 82% | 3 |
| comparaison | 627 | 82% | 3 |
| conséquence | 431 | 81% | 4 |
| défi | 447 | 80% | 3 |
| réalité | 469 | 79% | 4 |
| revue | 527 | 79% | 4 |
| approche | 3036 | 78% | 3 |
| apport | 1092 | 78% | 2 |
| regard | 763 | 75% | 3 |
| point | 595 | 75% | 4 |
| élément | 592 | 74% | 3 |
| état | 1104 | 73% | 3 |
| question | 1152 | 72% | 4 |
| lieu | 828 | 72% | 4 |
| étude | 6089 | 71% | 3 |
| outil | 1364 | 71% | 4 |
| expérience | 1203 | 71% | 4 |
| concept | 519 | 70% | 4 |

Cette classe peut être visualisée à l’aide d’un graphisme avec le nombre total d’occurrences en abscisse et le pourcentage d’occurrences après le double point  :

On remarque que la position après le double point reste strictement inférieure à 5, avec une moyenne de la classe entière à 3, ce qui rend ces noms très proches du double point.

On peut aussi remarquer que la classe inversée, qui prendrait un taux de 70% avant le double point et une fréquence minimum de 500 est réduite à un seul nom : « compte » qui compte 889 occurrences dont 71% avant le double point. À part cette exception, on peut dire qu’on constate une affinité de certains noms qu’avec la partie postérieure au double point.



La méthode présentée pour obtenir notre corpus dans cette partie est reproductible et permet d’obtenir d’en obtenir de nouveaux à partir de HAL. Le corpus que nous avons utilisé présente l’avantage d’être de grande taille et de présenter une grande variété de titres. Cette taille permet d’étudier un phénomène linguistique particulier, comme l’utilisation du double point, sur un nombre important de titres. On constate que la plupart des titres sont segmentés en deux par un double point.

L’étude du lexique des noms communs a permis de mettre en avant une classe de nom avec une affinité pour une position après le double point. Sémantiquement, ces mots appartiennent au vocabulaire de la recherche scientifique, sans être liés à une discipline particulière. La plupart sont en position deux. Un mot les sépare du double point sur leur gauche, on peut formuler l’hypothèse qu’il s’agit d’un déterminant. Nous aimerions connaître plus avant le contexte syntaxique immédiat de ces noms, voir s’ils inscrivent dans un syntagme utilisant un autre nom. Pour cela, nous allons enquêter sur les syntagmes auxquels ils appartiennent.

# III. Syntagmes et patrons

## III.1 Rappel sur les syntagmes

Le syntagme est selon (Maingueneau, Chiss, & Filliolet, 2007, p. 35), dans le cadre de la linguistique structurale, « un groupe d’éléments linguistiques [consécutifs] formant une unité [syntaxique] dans une organisation hiérarchisée », et pour (Neveu, 2017), « un syntagme est une suite de morphèmes qui forme un constituant de la proposition (ou de la phrase) et qui est organisé autour d’un noyau ». La catégorie de ce noyau donne le type du syntagme et « le syntagme exerce les mêmes fonctions syntaxiques que son noyau » (Neveu, 2017).

Concrètement, un syntagme est représenté dans notre travail par une séquence d’étiquettes POS comme « DET NC P DET NC »[[13]](#footnote-13) pour « la maison de la forêt ». Mais l’analyse en constituants immédiats nous a appris que malgré la linéarité de cette séquence, le syntagme s’organise en une structure hiérarchique à plusieurs niveaux représentable à l’aide de parenthèses, de boîtes ou d’arbres (Mounin, 2004, p. 81). (Maingueneau, Chiss, & Filliolet, 2007, p. 119) indique qu’un consensus existe pour utiliser cette dernière forme, les graphes arborescents communément appelés *arbres*.

Ainsi notre syntagme nominal, qui a pour noyau « maison », a un complément déterminatif prépositionnel qui est un syntagme prépositionnel, dont le noyau est « de », et qui contient lui-même un syntagme nominal, ayant pour noyau « forêt ».

Pour notre exemple on obtient avec en bleu les syntagmes, en orangé les noyaux :



Figure 1 : arbre d'analyse

Pour étudier notre corpus, nous avons inventorié toutes les séquences existantes d’étiquettes POS après le double point en comptant leurs nombres d’occurrences. Elles sont au nombre de 45 098.

## III.2 Limites de notre étude

Comme vu dans la partie II.3.6 Lexique des noms communs nous souhaitons étudier la classe de noms que nous avons distingués comme ayant une affinité pour se placer après le double point, de façon proche. Pour cela, nous voulons étudier les syntagmes dans lequel ils s’inscrivent et, en gardant à l’esprit leur proximité avec le double point, nous sommes confiants dans le fait qu’il s’agit du syntagme qui suit immédiatement le double point.

Nous savons qu’un syntagme ne s’étend jamais au-delà de la phrase dont il est un constituant. Nous pouvons donc exclure de notre observation tout ce qui se trouve après un point, un point d’interrogation, un point d’exclamation ou un point-virgule. La taille de notre inventaire des séquences d’étiquettes POS après le double point tombe alors à 42 942. On élimine 2156 séquences qui sont jugées équivalentes.

**Exemple :** *CAN : la culture à Nancy au prisme de ses habitants. Pratiques, regards, symboles (Cécile Bando, Lylette Lacote-Gabrysiak & Adeline Clerc-Florimond, 2018)*

L’exemple ci-dessus se traduit par la séquence d’étiquettes POS suivantes : « NPP PONCT DET NC P NPP P+D NC P DET NC PONCT NC PONCT NC PONCT NC ». La partie en rouge de la suite est ignorée, la suite devient donc équivalente à celle ayant seulement les étiquettes en noir.

Pour réduire la portée de notre étude aux dimensions de cet exercice, nous voulons borner notre étude du syntagme qui potentiellement pourrait être constitué de tout ce qui suit le double point jusqu’à la fin du segment. Il est néanmoins difficile de savoir où « couper » entre partie observée et partie non observée dans la séquence après le double point pour restreindre notre étude. Nous savons déjà néanmoins que nous voulons capturer le premier nom immédiatement après le double point. Il est également très difficile de résonner avec le grand nombre de séquences possibles, il nous faut les regrouper.

En parcourant visuellement notre inventaire dans Excel, mettant en valeur les étiquettes POS à l’aide d’un code couleur, nous avons eu trois intuitions. La première est que le syntagme nominal est le plus représenté juste après le double point. La deuxième est que ce syntagme nominal inclus souvent un syntagme prépositionnel qui est un complément du nom noyau et qui contient lui aussi un nom. Notre troisième intuition est que, même en se limitant au syntagme ayant cette structure complexe, il reste toujours une grande variété de séquences possibles :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DET | NC | P | DET | NC |  |
| DET | NC | P | NC |  |  |
| DET | NC | P+D | NC |  |  |
| NC | P | NC |  |  |  |
| NC | P | DET | NC |  |  |
| DET | NC | P | DET | NC | ADJ |

Tableau : exemples de suites de catégories correspondant à un syntagme nominal après le double point

Nous décidons de nous arrêter au deuxième nom après le double point, en prenant éventuellement un dernier adjectif. Le syntagme binominal ainsi capturé fourni assez de contexte à notre premier nom et l’ensemble reste simple à analyser. Néanmoins, pour capturer toutes les variantes des séquences, nous avons besoin d’un autre outil conceptuel : le patron.

## III.3 Définition des patrons

Nous définissons le patron comme un modèle définissant un ensemble de séquences. Chacune de ces séquences est capturée ou générée par le patron. On dit aussi que le patron et la séquence s’accorde. La caractéristique principale du patron est sa variabilité. Le modèle définit que certaines étiquettes POS peuvent apparaître de façon optionnelle dans la séquence, que d’autres peuvent être répétées, enfin que certaines doivent être choisies entre plusieurs alternatives. Par exemple, dans un syntagme nominal, un déterminant initial est optionnel ainsi que des adjectifs qualifiants le nom. Le seul élément qui est obligatoirement présent est le noyau du syntagme, le nom. Le patron « NC » n’est pas très utile et sa variabilité est nulle : il définit une seule séquence, comprenant un seul élément, un nom commun.

Pour représenter nos patrons, nous utilisons un langage spécifique très simple qui tient en 5 principes :

1. A B signifie une étiquette A suivi d’une étiquette B
2. [ A B ] signifie un choix : soit l’étiquette A, soit l’étiquette B
3. A? signifie l’optionalité : l’étiquette A peut apparaître une fois mais ce n’est pas obligé.
4. ( A B ) permet de grouper les étiquettes POS
   1. à l’intérieur d’un choix : [ (A B) C ] signifie soit A suivi de B, soit C
   2. ou pour signifier qu’une sous-séquence entière est optionnelle ( A B )? .
5. La répétition est représentée par le fait de répéter plusieurs fois la même suite d’éléments : A A . Notre langage ne permet pas de représenter la répétition infinie d’un élément.

Nous pouvons donc écrire nos patrons à l’aide de ce langage pour capturer les différentes écritures possibles d’un syntagme. Nos patrons capturent des séquences et non des structures, les deux correspondent à des syntagmes mais la séquence peut être ambigüe.

**Exemple :** Les syntagmes « Un tonneau plein de sable » et « Une plage étroite de sable » ont la même séquence d’étiquettes, « DET NC ADJ P NC » mais pas la même structure : le syntagme prépositionnel « de sable » est inclus dans le syntagme adjectival « plein de sable » dans le premier cas, alors que dans le second il est inclus dans le syntagme nominal, c’est la plage qui est faite de sable. On le prouve en supprimant « plein » dans la première phrase : « Un tonneau de sable » change le sens et montre que si plein est supprimé, « de sable » doit l’être aussi pour préserver le sens.

Lever cette ambiguïté reviendrait à faire une analyse syntaxique plus poussée des éléments après le double point, or, de façon générale c’est ce que nous voulons éviter. Nous resterons conscients de cette simplification lorsque nous analyserons les résultats fournis par les scripts automatiques. Nous pouvons à présent commencer à écrire nos patrons.

**Exemple** : Le patron qui correspond à la fois au syntagme illustré par la *Figure 1 : arbre d'analyse* et aux séquences dans le *Tableau 5: exemples de suites de catégories correspondant à un syntagme nominal après le double point* peut être écrit ainsi :

DET? NC [ (P DET?) P+D ] NC ADJ?

Ce patron correspond aux exemples proposés précédemment, mais il correspond également à d’autres constructions possibles comme « DET NC P+D NC ADJ ». Éventuellement, ce patron pourrait correspondre à des séquences qui ne sont pas représentées dans notre corpus. Ce n’est pas un problème car notre but n’est pas d’avoir un patron générant toutes les suites de notre corpus, mais de capturer toutes celles qui s’y trouvent et y correspondent, et de les regrouper sous un patron donné. Comme notre langage ne permet pas la répétition à l’infini d’élément, il est toujours possible de générer toutes les séquences s’accordant avec un patron donné. Leur nombre donne une mesure de sa variabilité. Le patron précédemment écrit décrit les 12 séquences suivantes :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NC | P | NC |  |  |  |
| DET | NC | P | NC |  |  |
| NC | P | DET | NC |  |  |
| DET | NC | P | DET | NC |  |
| NC | P+D | NC |  |  |  |
| DET | NC | P+D | NC |  |  |
| NC | P | NC | ADJ |  |  |
| DET | NC | P | NC | ADJ |  |
| NC | P | DET | NC | ADJ |  |
| DET | NC | P | DET | NC | ADJ |
| NC | P+D | NC | ADJ |  |  |
| DET | NC | P+D | NC | ADJ |  |



Nous avons exposé nos patrons ainsi que leur limitation de ne capturer que la séquence et non la structure des syntagmes. Nous avons décrit le langage mis au point pour les exprimer et indiqué une mesure de leur variabilité. Nous pouvons à présent passer à la construction de trois patrons qui couvriront la majorité de notre corpus, chacun capturant un syntagme comprenant deux noms, dans le but de retrouver de capturer le contexte environnant les noms de notre classe.

# IV. Études des trois patrons

## IV.1 Construction itérative de trois patrons

Nous avons dû arbitrer entre complexité et faisabilité : plutôt que de reconstruire l’ensemble de l’arbre syntaxique après le double point, nous nous contentons de regarder le premier syntagme et dans celui de se limiter arbitrairement dans son analyse.

Les trois types de syntagmes que nous avons décidé d’étudier sont :

1. un syntagme nominal, lui-même composé d’un sous syntagme prépositionnel
2. un syntagme prépositionnel, lui-même composé d’un sous syntagme prépositionnel
3. un syntagme nominal avec coordination, coordonnant deux syntagmes nominaux

À chacun de ces types correspondant un patron, noté respectivement SN, SP et SNC. Cette sélection s’est faite en regardant notre inventaire et en choisissant les séquences utilisées par le plus de titres pour avoir une couverture maximale. Par exemple, pour le patron SN, nous avons surligné sur en bleu, les séquences les plus fréquentes correspondant à celui-ci :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Nb titres** | **Séquence d'étiquettes POS** | | | | | |
| 1450 | NC | CC | NC |  |  |  |
| 1104 | DET | NC | ADJ |  |  |  |
| 746 | DET | NC | P | DET | NC |  |
| 666 | DET | NC | P | NC |  |  |
| 620 | NC | ADJ |  |  |  |  |
| 540 | DET | NC | P+D | NC |  |  |
| 520 | NC | P | NC |  |  |  |
| 504 | DET | NC | ADJ | PONCT |  |  |
| 496 | NC | P | DET | NC |  |  |
| 478 | NC | PONCT | NC | CC | NC |  |
| 444 | NC | P | DET | NC | ADJ |  |
| 433 | DET | NC | P | DET | NC | ADJ |

Tableau : Les séquences les plus fréquentes dans les titres

Un patron s’accordera avec l’entièreté de la séquence si le début de celle-ci s’accorde avec lui, c’est-à-dire correspond à une séquence générée par celui-ci. Si plusieurs séquences générées par un patron correspondent au début d’une même séquence du corpus, la séquence générée la plus longue sera retenue.

**Exemple :** le patron « NC ADJ? » génère deux séquences possibles : « NC » et « NC ADJ ». Si, lors de notre interrogation de notre corpus, nous tombons sur une séquence « NC ADJ P NC », son début correspond bien au début des deux séquences possibles. Nous retiendrons que la séquence du corpus correspond *le plus* à la séquence possible « NC ADJ » car plus d’éléments s’accordent.

Pour obtenir nos trois patrons correspondants, nous avons choisi une méthode itérative en se basant sur l’observation de l’inventaire, la connaissance des règles de syntaxe et un script Python. Celui effectuait deux opérations :

1. Le comptage automatique de la couverture du patron. Il y a deux types de couvertures :
   1. Le nombre de séquences d’étiquettes POS couvertes par le patron par rapport au nombre total de séquences inventoriées (42 942)
   2. Le nombre de titres auxquels il correspond, par rapport au nombre total et de titres dans notre corpus (85 531).

La plus importante est la couverture des titres, car certaines séquences sont très peu utilisées : 37 150 séquences, soit 86%, ne sont utilisées que par un seul titre.

1. La séparation en deux des séquences du corpus : d’un côté celles qui s’accordent avec notre patron, de l’autre, celles qui ne s’accordent pas. En regardant attentivement ces dernières, on peut décider alors d’augmenter la variabilité de notre patron pour qu’il génère plus de séquences et améliore ses taux de couverture.

Nous sommes partis à chaque fois d’un patron minimaliste, par exemple pour le patron SN de « NC P NC », avant de rajouter les différents éléments optionnels puis les choix et les répétitions possibles pour obtenir un patron ayant une couverture maximale sans dénaturer la nature du syntagme capturé par le patron. De ce fait, les trois patrons sont mutuellement exclusifs : une suite ne peut s’accorder qu’avec un seul d’entre eux ou aucun.

**Exemple :** le patron SN qui capture un syntagme nominal incluant un syntagme prépositionnel ayant un nom avait, à un moment des itérations, la forme suivante :

DET? ADJ? [NC NPP] [NC NPP]? ADJ? [(P DET?) P+D] ADJ? [NC NPP] [NC NPP]? ADJ?

Il génère alors 3 456 séquences possibles, la plus longue ayant 11 étiquettes et la plus courte 3. Il s’accorde avec 20 572 séquences de notre corpus, soit 47.91%, et ne s’accorde pas avec 22 370, soit 52.09%. En prenant en compte la fréquence de ces séquences dans les titres, il s’accorde avec 41 327 (48.32%) de ceux-ci. En observant les séquences avec lesquelles il ne s’accorde pas, nous avons déterminer 2 améliorations possibles :

1. Offrir la possibilité que le premier déterminant optionnel puisse être un déterminant interrogatif
2. Que le deuxième adjectif optionnel soit, au choix, un adjectif, un adverbe suivi d’un adjectif ou un adjectif suivi d’un adverbe

La nouvelle forme de notre patron est la suivante, avec en bleu les changements :

[DETWH DET]? ADJ? [NC NPP] [NC NPP]? [(ADV ADJ) ADJ (ADJ ADV)]? [(P DET?) P+D] ADJ? [NC NPP] [NC NPP]? ADJ?

Le patron génère à présent 10 368 séquences possibles, la plus longue ayant 12 étiquettes et la plus courte 3. Il s’accorde avec 21 192 séquence de notre corpus, soit 49.35%, une amélioration de +620 en absolu et de +1,44%. Il ne s’accorde pas avec 21 750 séquences, soit 50.65%. Il s’accorde avec 42 606 titres, soit 49.81% et une amélioration de 1279 en absolu et de 1.49%. On voit que le gain de couverture des titres est faible par rapport à l’explosion du nombre de séquences possibles. C’est le signe que, pour essayer de couvrir au maximum notre corpus de titres, il ne faut pas complexifier encore plus ce patron-là mais essayer d’en construire d’autres radicalement différents, en capturant autre chose qu’un syntagme nominal incluant un syntagme prépositionnel ayant un nom. Dans les trois parties qui suivent, nous présentons nos 3 patrons.

## IV.2 Patron SN : syntagme nominal

### IV.2.A Fiche d’identité

**Description** Il s’agit d’un syntagme nominal incluant un syntagme prépositionnel qui   
inclut un syntagme nominal.

**Structure minimale  
 **

**Patron minimal** NC P NC /

**Patron étendu**

[DETWH DET]? ADJ? [NC NPP] [NC NPP]? [(ADV ADJ) ADJ (ADJ ADV)]? [(P DET?) P+D] ADJ? [NC NPP] [NC NPP]? ADJ?

**Possibilités** 10 368 séquences, longues de 3 à 12 étiquettes

**Couverture du corpus** 49.35% des séquences de notre corpus (21 192 séquences sur 42 942)

49.81% des titres de notre corpus (42 606 titres sur 85 531)

**Exemples**

* **a**

## IV.3 Patron SP : syntagme prépositionnel

### IV.3. A Fiche d’identité

**Description** Il s’agit d’un syntagme prépositionnel incluant un syntagme   
 prépositionnel, les deux ayant un nom.

**Structure minimale **

**Patron minimal** P NC P NC /

**Version étendue**

[P+D P] DET? ADJ? [NC NPP] [NC NPP]? ADJ? [(P DET?) P+D] ADJ? [NC NPP] [NC NPP]? ADJ?

**Possibilités** 6 912 séquences, longues de 4 à 12 étiquettes

**Couverture du corpus** 4.81% des séquences de notre corpus (2 065 séquences sur 42 942)

4.70% des titres de notre corpus (4 023 titres sur 85 531)

**Exemples**

* a

## IV.4 Patron SNC : syntagme nominal avec coordination

### IV.4. A Fiche d’identité

**Description** Il s’agit d’un syntagme nominal avec une coordination incluant deux sous-syntagmes nominaux.

**Structure minimale **

**Patron minimal** NC CC NC /

**Version étendue**

[P+D P] DET? ADJ? [NC NPP] [NC NPP]? ADJ? [(P DET?) P+D] ADJ? [NC NPP] [NC NPP]? ADJ?

**Possibilités** 1 728 séquences, longues de 3 à 10 étiquettes

**Couverture du corpus** 7.35% des séquences de notre corpus (3 158 séquences sur 42 942)

10.34% des titres de notre corpus (8 845 titres sur 85 531)

**Exemples**

* *La mise en place du dispositif d'orientation active à l'Université : enjeux et perspectives* (Sylvain Obajtek, 2018)

## IV.5 Analyse globale des 3 patrons

### IV.5.A Couverture globale

Nos trois patrons couvrent de façon exclusive :

|  |  |  |
| --- | --- | --- |
| **Patrons** | **Couverture séquences** | **Couverture des titres** |
| Patron SN | 49.35% | 49.81% |
| Patron SP | 4.81% | 4.70% |
| Patron SNC | 7.35% | 10.34% |
| **Total** | 61.51% | 64.85% |

Tableau : couverture des trois patrons

Nous couvrons 64.85% de nos titres avec nos trois patrons. Il ne se dégage pas de l’inventaire des séquences restantes des régularités évidentes pour trouver d’autres patrons. Néanmoins, un peu moins des deux tiers de notre corpus de travail est couvert.

## IV.5.B Répétition de syntagme

Après avoir identifié différents syntagmes auxquels

## Répartition par domaine

PPP TROIS PATRONS



Nos trois patrons couvrent ensemble X % des séquences inventoriées et X % des titres.

# V. Étude de structures particulières

Dans la partie précédente, nous avons étudié uniquement les catégories et nos patrons ne portaient uniquement sur celles-ci. À présent, nous voulons étudier les lemmes qui se cachent derrière les catégories.

PPP STRUCTURES PARTICULIERES

# VI. Résultats et discussions

PPP RESULTATS

## VI.1 Résultats

## VI.2 Discussions

Parler du fait qu’on ne choisit pas l’émergence auto

Dire que notre langage ne gère pas l’infinité et ce que cela fait

Les travaux de Quiniou et al. (2012) nous mettent sur la piste de chercher des motifs émergents de syntagmes nominaux. Intuitivement, on peut penser qu’il se caractérise d’un article (optionnel), d’une suite d’adjectifs, éventuellement modifiés par un adverbe, du substantif noyau et d’adjectifs postposés, éventuellement complémentés, ce que ne sont jamais les adjectifs postposés.

Le premier point fait que nous ne saurons pas si le syntagme nominal est le sujet d’une phrase complète ou pas.

VI.1.1 Défauts remarqués

Talismane : ADJ pris comme NC : il cherche des phrases, spécificités des titres

Il est rare qu’un titre forme une phrase verbale. De plus, un titre est souvent très segmenté par un double point, des virgules voir même des points et ces partitions sont autant de phrases incomplètes. TreeTagger (Schmid, 2013) et la plupart des logiciels de lemmatisation et de catégorisation grammaticale sont parfois perplexes pour analyser de telles phrases car ils ont été plutôt pensés pour le contenu des textes et non leur titre. Leech (2000) propose la notion de C-units, des petites unités grammaticales indépendantes très présentes à l’oral, pour aborder ces constructions. Dans notre cas, l’outillage de notre mémoire se fera au cours du second semestre et nous espérons ne pas avoir à reprogrammer un tel outil adapté aux spécificités des titres.

Outil propre : « étude de cas » vu comme 3 suites N1 P N2 différentes,

Heureusement, 2 sont très marginales.

4.1.2 Ouverture sémantique

*Sur le plan sémantique, que peut-on dire à ce stade ?*

5.1 Perspectives

5.1.1 Couverture totale du corpus

5.1.2 Couverture totale des titres

5.1.3 Émergence automatique de patrons

# Conclusion

PPP CONCLUSION

Perspectives

# Bibliographie

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

Bray, T. (2017). *The JavaScript Object Notation (JSON) Data Interchange Format*. Retrieved from Internet Engineering Task Force (IETF) Tools: https://tools.ietf.org/html/rfc8259

Cori, M., & David, S. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages, (3)*, 111-129.

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)*.

Doppagne, A. (1998). *La bonne ponctuation : clarté, efficacité et présence de l’écrit* (éd. 3e). Duculot.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5(1)*, 1-26.

Goodman, R. A., Thacker, S. B., & Siegel, P. Z. (2001). What’s in a title? A descriptive study of article titles in peer-reviewed medical journals. *Science, 24(3)*, 75-8.

Grevisse, M., & Goosse, A. (2011). *Le bon usage : grammaire française.* Duculot.

Haggan, M. (2004). Research paper titles in literature, linguistics and science: dimensions of attraction. *Journal of Pragmatics, 2.36*, 293-317.

Hartley, J. (2003). Single authors are not alone: Colleagues often help. *Journal of Scholarly Communication, 34 (2)*, 108-113.

Jamali, H. R., & 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.

Lester, J. (1993). *Writing Research Papers. A complete Guide.* Harper Collins.

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

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

Maingueneau, D., Chiss, J.-L., & Filliolet, J. (2007). *Introduction à la linguistique française.* Hachette Éducation.

Martin, R. (2002). *Comprendre la linguistique, épistémologie élémentaire d'une discipline.* Paris: PUF.

Mounin, G. (dir.) (2004). *Dictionnaire de la linguistique.* Presses Universitaires de France.

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

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

Neveu, F. (2017). *Lexique des notions linguistiques.* Armand Colin.

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

Quiniou, S. C. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. *Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)* (pp. 821-833). Liège: JADT.

Rebeyrolle, J., Jacques, M.-P., & Péry-Woodley, M.-P. (2009). Titres et intertitres dans l'organisation du discours. *Journal of French Language Studies, 19*, 269-290.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *7th international conference on Language Resources and Evaluation (LREC 2010).* La Valette.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *New methods in language processing*, 154.

Smiley, D., Pugh, E., Parisa, K., & Mitchell, M. (2015). *Apache Solr enterprise search server.* Birmingham: Packt Publishing Ltd.

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

Swales, J. M. (1990). Genre Analysis. *English in Academic and Research Settings*.

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.

Whissell, C. (2004). Titles of articles published in the journal Psychological Reports: Changes in language, emotion, and imagery over time. *Psychological reports, 94(3)*, 807-813.

# Annexes

## A1. Requêtes Apache Solr sur HAL

### A1.A Requêtes

Nous proposons d’explorer un exemple de requête avec la plate-forme Apache Solr (Requête 1) que nous formatons afin de le rendre plus visible. Nous demandons les informations supplémentaires de la discipline (domain\_s), des auteurs (authFullName\_s), du type de document (docType\_s), la date de modification (modifiedDateY\_i) et bien sûr le titre (title\_s). Nous classons en fonction de la date de modification, de la plus récente à la plus vieille. Par défaut, 30 résultats sont retournés, le maximum autorisé étant de 10 000, nous en demandons ici 1000 pour que l’exécution de la requête ne prenne pas trop de temps.

https://api.archives-ouvertes.fr/search/ ?  
 wt = json &   
 fl = docid, domain\_s, authFullName\_s,   
 docType\_s, title\_s, modifiedDateY\_i &   
 indent = true &  
 sort = modifiedDateY\_i desc &  
 rows = 1000

*Requête 2 : Un exemple de requête avec l’API Apache Solr formaté pour plus de lisibilité*

Un système de cache permet de récupérer bien plus de résultats que la limite de 10 000 par requête. En triant sur l’identifiant numérique des notices, le champs docid , une clé unique car aucune notice ne possède la même valeur, on demande à Solr 1000 résultats et la création d’un cache. La première réponse de 1000 résultats comporte à la fin un identifiant. En relançant une requête et en fournissant l’identifiant fourni, on obtient les 1000 résultats *suivants*. Les deux requêtes suivantes (Requêtes 2a et 2b), la première de création de cache, la seconde de poursuite de la demande avec l’identifiant fourni par la première requête, AoFVmLIG , illustrent cette puissante fonctionnalité.

https://api.archives-ouvertes.fr/search/?wt=json&fl=docid,domain\_s,authFullName\_s,docType\_s,title\_s,modifiedDateY\_i&indent=true&sort=docid%20desc&rows=1000&cursorMark=\*

https://api.archives-ouvertes.fr/search/?wt=json&fl=docid,domain\_s,authFullName\_s,docType\_s,title\_s,modifiedDateY\_i&indent=true&sort=docid%20desc&rows=1000&cursorMark=AoFVmLIG

*Requêtes 3a et 2b : création et consultation d’un cache*

Nous avons conçu un script Python qui automatise la création d’un cache puis le passage des requêtes successives en sauvegardant au fur et à mesure les résultats. Nous avons exécuté ce script 304 fois pour obtenir 304 600 réponses, les 600 supplémentaires venant des étapes de mise au point du script. Nous pouvons à présent nous pencher sur les résultats de ces requêtes qui vont constituer notre corpus.

### A1.B Résultats

Notons que, comme le souligne Cori et David (2008), l’étape de sélection des données pour constituer un corpus peut comporter une part de subjectivité. Dans notre cas nous nous en exemptons car nous n’opérons pas de choix dans les titres que nous retournent HAL. Si nous devions préparer plus avant nos données, en particulier en écartant certains titres pour des raisons d’invalidité technique ou de doublon, nous respecterions la précaution méthodologique que préconisent ces deux auteurs, faire *« un inventaire soigneux de toutes les décisions prises en amont »* L’encadré *Résultat 4 : exemple d’un élément de résultats au format JSON* présente un élément de résultat d’une requête Apache Solr.

{

"docid" : 1675646,

"domain\_s" : [

"0.scco", "1.scco.ling", "0.scco", "1.scco.psyc"],

"title\_s" : [

"Récits d’enfants et d’adolescents - Développements typiques, atypiques, dysfonctionnements"],

"authFullName\_s" : [

"Christiane Préneron", "Claire Martinot"],

"docType\_s" : "DOUV",

"modifiedDateY\_i" : 2018

},

*Résultat 4 : exemple d’un élément de résultats au format JSON*

Les titres nous parviennent accompagnés des autres métadonnées de la notice. Pour simplifier, nous considérerons que les métadonnées de la notice sont également des métadonnées du titre. Elles ne sont pas incluses dans le contenu du titre lui-même, il s’agit des caractéristiques du document titré que l’on pourra mettre en relation avec les traits du titre. Elles sont : l’identifiant numérique de la notice, le type du texte titré, son année d’enregistrement sur HAL, indispensable pour des études en diachronie, le nombre d’auteurs, son champ et ses disciplines. On notera que les disciplines peuvent être hiérarchisées en arbre et qu’un même article peut être étiqueté sous plusieurs disciplines. En guise d’exemple, *Résultat 5 : une ligne de notre premier corpus de travail* présente une ligne de notre corpus avec le titre et ses métadonnées.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Titre** | **Type** | **Année** | **Nauteurs** | **Champ** | **Disciplines** |
| artxibo-01200715 | Deux dichotomies de la langue basque | other | 2014 | 1 | SHS | SHS.LANGUE.SOCIO |

*Résultat 5 : une ligne de notre premier corpus de travail*

Pour finir, nous sauvegardons ces données dans un format XML personnel dans un fichier encodé en UTF-8. La taille de ce fichier est de 450 Mo.

## A2. Codes des étiquettes de catégorie de discours de Talismane

Ce tableau est issu de la documentation officielle de Talismane (Urieli, 2013) accessible en ligne[[14]](#footnote-14).

|  |  |
| --- | --- |
| **Tag** | **Part of speech** |
| ADJ | Adjective |
| ADV | Adverb |
| ADVWH | Interrogative adverb |
| CC | Coordinating conjunction |
| CLO | Clitic (object) |
| CLR | Clitic (reflexive) |
| CLS | Clitic (subject) |
| CS | Subordinating conjunction |
| DET | Determinent |
| DETWH | Interrogative determinent |
| ET | Foreign word |
| I | Interjection |
| NC | Common noun |
| NPP | Proper noun |
| P | Preposition |
| P+D | Preposition and determinant combined (e.g. "du") |
| P+PRO | Preposition and pronoun combined (e.g. "duquel") |
| PONCT | Punctuation |
| PRO | Pronoun |
| PROREL | Relative pronoun |
| PROWH | Interrogative pronoun |
| V | Indicative verb |
| VIMP | Imperative verb |
| VINF | Infinitive verb |
| VPP | Past participle |
| VPR | Present participle |
| VS | Subjunctive verb |

## A3. Index des tableaux

[Tableau 1 : nombre de doubles points dans les titres 16](#_Toc520453138)

[Tableau 2 : Répartition des titres par type 18](#_Toc520453139)

[Tableau 3 : Répartition des titres par année 19](#_Toc520453140)

[Tableau 4 : Nombres de titre par nombres d'auteurs 20](#_Toc520453141)

[Tableau 5 : Nombre de titres par nombres de doubles points 21](#_Toc520453142)

[Tableau 6 : Répartition des titres par domaines 21](#_Toc520453143)

[Tableau 7 : Titres avec un caractère segmentant dans notre corpus 23](#_Toc520453144)

[Tableau 8 : Phrase complète dans les titres en fonction du domaine de la biologie 23](#_Toc520453145)

[Tableau 9 : Comptes des noms communs avant et après le double point 24](#_Toc520453146)

[Tableau 10: exemples de suites de catégories correspondant à un syntagme nominal après le double point 28](#_Toc520453147)

[Tableau 11: Les séquences les plus fréquentes dans les titres 30](#_Toc520453148)

## A4. Index des graphiques

[*Figure 1 : arbre d'analyse* 27](#_Toc520453211)

[*Requête 2 : Un exemple de requête avec l’API Apache Solr formaté pour plus de lisibilité* 37](#_Toc520453212)

[*Requêtes 3a et 2b : création et consultation d’un cache* 37](#_Toc520453213)

[*Résultat 4 : exemple d’un élément de résultats au format JSON* 38](#_Toc520453214)

[*Résultat 5 : une ligne de notre premier corpus de travail* 38](#_Toc520453215)

## A5. Index des logiciels et technologies mentionnés

Apache Lucene 12

Apache Solr 12

CoNLL-U 13

CSV 12

Excel 6

HTTP 12

JSON 12

langdetect 16

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) 12

protocole de transfert hypertexte *HTTP*

Python 6

Stanford Core Natural Language Processing 13

Talismane 13

XML 12

1. <https://www.python.org/> 2 <https://products.office.com/fr-fr/excel> [↑](#footnote-ref-1)
2. L’indexation en base de données est une opération consistant à construire un index sur une ou plusieurs clés. Celui-ci donne, pour une valeur de clé, rapidement accès aux données ainsi indexées. Exemple : indexer les articles sur leurs années de publication (= clé). On pourra retrouver facilement tous les articles de 2015 = (valeur de clé). [↑](#footnote-ref-2)
3. Organisation européenne pour la recherche nucléaire <http://home.cern/fr> [↑](#footnote-ref-3)
4. <https://hal.archives-ouvertes.fr/> [↑](#footnote-ref-4)
5. <https://www.ccsd.cnrs.fr/> [↑](#footnote-ref-5)
6. <http://arxiv.org/> [↑](#footnote-ref-6)
7. <https://lucene.apache.org/> [↑](#footnote-ref-7)
8. <https://stanfordnlp.github.io/CoreNLP/> [↑](#footnote-ref-8)
9. <http://redac.univ-tlse2.fr/applications/talismane/talismane.html> [↑](#footnote-ref-9)
10. <http://universaldependencies.org/docs/format.html> [↑](#footnote-ref-10)
11. <https://pypi.org/project/langdetect/> [↑](#footnote-ref-11)
12. Le comptage des mots se basent sur la segmentation opérée par Talismane moins les formes étiquetées comme marques de ponctuation. [↑](#footnote-ref-12)
13. Nous utilisons dans ce document les étiquettes de Talismane pour les catégories. Celles utilisées ici sont DET pour déterminant, NC pour nom commun, P pour préposition. La liste complète est donnée dans l’annexe A3. Codes des étiquettes de catégorie de discours de Talismane. [↑](#footnote-ref-13)
14. Plus précisément ce tableau se trouve ici : <http://joliciel-informatique.github.io/talismane/#section2.3.4> [↑](#footnote-ref-14)