



## 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

2017 – 2018



## 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. Ils sont parfois ardu, mais pour rien au monde je ne regrette ce voyage.

Je veux aussi saluer mes camarades de promotion qui ont voyagé avec moi et les amis 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
Introduction.....	7
I. Précédentes études sur les titres scientifiques.....	8
I.1 Le titre et ses problématiques.....	8
I.2 Traits étudiés des titres .....	9
I.3 Caractéristiques des documents .....	10
I.4 Corpus utilisés .....	11
I.5 L'utilisation du double point .....	12
II. Corpus de travail.....	14
II.1 Présentation de HAL et extraction des données.....	14
II.2 Traitement des données.....	15
II.2.1 Enrichissement des données.....	15
II.2.2 Conversions .....	16
II.2.3 Filtrage.....	18
II.3 Mesures du corpus .....	20
II.3.1 Taille du corpus et types des documents.....	20
II.3.2 Années des documents .....	21
II.3.3 Longueurs des titres et nombre d'auteurs.....	21
II.3.4 Domaines et nombre de domaines .....	24
II.3.5 Marques de ponctuation et segmentation .....	24
II.3.6 Lexique des noms communs .....	26
III. Syntagmes et patrons.....	29
III.1 Séquences d'étiquette POS et syntagmes.....	29
III.2 Limites de notre étude .....	31
III.3 Définition des patrons .....	32
III.4 Limites de nos patrons .....	34
IV. Études des trois patrons.....	35
IV.1 Définition et construction des trois patrons .....	35
IV.1.1 Présentation des trois patrons.....	35
IV.1.2 Construction itérative du patron SN .....	35
IV.1.3 Explication des séquences [NC NPP] [NC NPP]? .....	37
IV.1.4 Exclusivité mutuelle des 3 patrons.....	37
IV.2 Patron SN : syntagme nominal.....	38

IV.2.1 Fiche d'identité.....	38
IV.3 Patron SP : syntagme prépositionnel.....	39
IV.3.1 Fiche d'identité.....	39
IV.4 Patron SNC : syntagme nominal avec coordination.....	40
IV.4.1 Fiche d'identité.....	40
IV.5. Couverture globale du corpus.....	41
V. Analyse syntaxico-lexicale des résultats .....	41
V.1 Résultats du patron SNC .....	42
V.1.1 Fréquences des coordinations .....	42
V.1.2 Fréquences des noms et emplacements.....	42
V.1.3 Fréquences des couples de noms ordonnés .....	43
V.1.4 Fréquence des triplets.....	43
V.1.5 Fréquences des couples de noms non ordonnés.....	44
V.2 Résultats du patron SN.....	46
V.2.1 Fréquences des prépositions .....	46
V.2.2 Fréquences des noms en première position.....	46
V.2.3 Fréquences des noms en première position avec la préposition .....	47
V.2.4 Fréquences du nom en deuxième position.....	48
V.2.5 Fréquences des syntagmes complets.....	49
V.2.6 Exprimer la notion d'« état des lieux » .....	49
V.3 Résultats du patron SP .....	53
V.3.1 Fréquences de la première préposition .....	53
V.3.2 Fréquences du premier nom.....	53
V.3.3 Fréquences de la seconde préposition .....	54
V.3.4 Fréquences du second nom .....	54
V.3.5 Fréquences des couples (préposition 1, préposition 2).....	54
V.3.6 Fréquences des triplets (préposition 1, nom 1, préposition 2).....	55
V.4 Analyse globale des 3 patrons.....	56
V.4.1 Le champ lexical de la recherche scientifique.....	56
V.4.2 L'approche phraséologie.....	56
VI. Discussion et perspectives .....	58
VI.1 Recherche d'information par l'analyse sémantique des titres : le cas de « application ».....	58
VI.2 Limitations de l'outillage et des patrons.....	60
VI.2.1 Erreurs dans la lemmatisation et l'étiquetage POS .....	60

VI.2.2 Développement des patrons.....	62
VI.3 Zone non couverte et création de sous-corpus .....	62
VI.3.1 Zone non couverte du corpus .....	62
VI.3.2 Créations de sous-corpus .....	63
VI.4 Le cas des noms propres .....	63
VI.5 Autres structures.....	64
Conclusion .....	67
Bibliographie.....	69
Annexes .....	72
A1. Requêtes Apache Solr sur HAL .....	72
A1.A Requêtes .....	72
A1.B Résultats.....	72
A2. Codes des étiquettes de catégorie de discours de Talismane .....	74
A3. Index des tableaux .....	74
A4. Index des graphiques .....	76
A5. Index des logiciels, technologies et notions mentionnés .....	76
A6. Définition du schéma utilisé pour les corpus.....	76

## Introduction

---

UN 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 document titré et parfois du seul, s'il décide ensuite de ne pas poursuivre sa lecture. Comme les titres de la presse généraliste, les titres scientifiques 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'éclaircir 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 scientifiques et que la seconde n'est présente que de façon anecdotique. Elle peut même être considérée par des travaux prescriptifs comme contraire aux règles de bonne écriture d'un titre données.

Pour la première fonction, on peut se demander comment cette volonté d'information se traduit dans la construction d'un titre en français. C'est un texte très court mais qui doit dans un espace très 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 la catégorie de partie du discours 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. Nous essayerons d'expliquer leur apport à l'interprétation sémantique du titre.

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 est l'archive ouverte française HAL pour avoir des titres dans cette langue, nos outils sont le langage de programmation Python<sup>1</sup> et le logiciel de tableur Excel<sup>2</sup>.

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 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 dans nos résultats en essayant de fournir une explication pour chacune d'entre elles. Enfin, nous discutons des limites et perspectives de notre travail avant de conclure.

---

<sup>1</sup> <https://www.python.org/> <sup>2</sup> <https://products.office.com/fr-fr/excel>

# 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 sur son titre (Goodman, Thacker, & Siegel, 2001) dans un cadre où le nombre d'articles publiés ne cessent d'augmenter (Jacques & Sebire, 2010) et ou, pour certaines disciplines, plus de la moitié de ceux-ci ne sont jamais cités (Hamilton, 1991).

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<sup>2</sup> et de privilégier le but informatif selon (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) et (Haggan, 2004). (Jacques & Sebire, 2010) précisent que plus le titre sera long, plus il sera susceptible d'en contenir et donc d'être associé à des requêtes les contenant pour être ensuite trouvé. Néanmoins la recherche en plein texte permise par les avancées technologiques rend cette affirmation moins pertinente selon (Jamali & Nikzad, 2011) et (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 similaire 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 points 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.

---

<sup>2</sup> L'indexation en base de données est une opération consistant à construire un index sur une ou plusieurs clés. Pour une valeur de clé donnée, celui-ci donne rapidement accès aux données correspondantes. Exemple : on peut indexer les articles sur leurs années de publication (= clé). On pourra alors retrouver rapidement tous les articles de 2015 = (valeur de clé) car la base de données aura déjà trié ses données selon les valeurs de la clé.



## 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 long. (Jacques & Sebire, 2010) montrent que les 25 titres les plus cités dans 3 journaux médicaux « *ont plus de deux fois plus de mots dans le titre que les articles les moins cités* ». (Jamali & Nikzad, 2011) ne comptent que les substantifs du titre pour calculer leur longueur. (Nagano R. L., 2015) comptent tous les mots mais calcule également 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 segments** ou partitions, 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 deux titres indépendants, l'une étant le titre, l'autre le sous-titre. On peut compter la longueur de chacune pour les comparer ensuite entre elles.
- Le troisième trait est constitué 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 que ces articles sont moins cités : une amélioration de l'attractivité d'un article ne garantit donc pas son utilisation.
- Le quatrième trait est la **présence d'acronymes**. (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. (Jacques & Sebire, 2010) affirment ainsi que « *beaucoup de chercheurs peuvent plus fréquemment connaître ou utiliser l'acronyme plutôt que le nom complet* ». Ils constatent la présence d'acronyme dans un tiers des 25 articles les plus cités de trois journaux médicaux.
- Le cinquième trait est la **structure syntaxique** du titre. (Haggan, 2004) constate que 90% des titres étudiés 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* ». Leech avait déjà pointé que, quoique globalement rare à l'écrit, on les trouve néanmoins fréquemment dans les titres.
- Le sixième trait est la **présence d'une 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 légèrement plus longs.

Un autre trait, abordé par (Nagano, 2015), est le début du titre et en particulier l'usage de l'article défini « the ». Il montre que les titres en sciences dures ont moins tendance à l'utiliser.

Un autre trait, abordé par (Jacques & Sebire, 2010), concerne les lemmes du titre et plus particulièrement la présence d'un nom de pays. Du fait de sa présence, il précise son objet d'étude en le limitant, ce qui contribue au fait que l'article soit moins cité du fait de sa spécialisation.

Ces traits touchent aussi bien à la syntaxe qu'à la présence de certains éléments comme les marques 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 ils contribuent « à la construction d'un discours cohérent », les liens entre eux et les textes qu'ils chapeautent, et les divisent en deux : ceux qui gèrent les référents et ceux qui ouvrent un espace thématique.

Après avoir énuméré les traits des titres, nous pouvons aborder les caractéristiques des documents titrés avec lesquelles ils sont mis en rapport.

### 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 énumérons 6 caractéristiques :

- 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), (Jacques & Sebire, 2010) et (Whissell, 2004). (Nagano, 2015) remarque que sciences dures et sciences humaines forment deux blocs de disciplines qui se comportent semblablement au niveau des traits qu'il étudie : les sciences dures ont des titres plus longs, un taux de noms supérieur, et utilisent moins l'article défini « the » au début.
- 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 ce 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* » scientifiques. (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<sup>3</sup>, 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 a été publié ou prépublié le document. Ce qui est en jeu est la langue maternelle des auteurs, ou la langue de travail pour la revue et la plate-forme. Il peut s'agir de savoir seulement si c'est l'anglais ou non.
- 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.
- La sixième caractéristique est le **nombre de citations** du document. Certaines plates-formes électroniques comptabilisent combien de fois l'article a été cité. (Jamali & Nikzad, 2011) mettent en relation le nombre de téléchargements et de citations avec la longueur du titre et la présence dans celui-ci d'un double point ou d'un point d'interrogation. Les titres avec un double point sont légèrement moins téléchargés et cités que ceux sans double point. Les titres plus longs sont légèrement moins téléchargés et cités que ceux plus courts. Cette conclusion est exactement l'inverse de celle de (Jacques & Sebire, 2010) en ce qui concerne la présence d'un double point ou l'augmentation de la longueur et le nombre de citations. Cette différence dans les résultats peut venir du fait que le corpus de (Jacques & Sebire, 2010) ne comptait que 300 titres contre 2 172 pour celui de (Jamali & Nikzad, 2011). Un corpus trop petit risque d'introduire un biais dans les résultats par un hasard de sélection.

De même que pour les traits des titres, certaines caractéristiques des documents sont corrélées entre elles : ainsi le nombre de téléchargements est positivement corrélé au nombre de citations (Jamali & Nikzad, 2011).

(Jacques & Sebire, 2010) citent sans les reprendre d'autres caractéristiques d'un document comme le genre des auteurs qui influe sur sa probabilité d'acceptation dans une revue ou son nombre futur de citations selon (Ayres, 2008).

Tous les articles étudiés décrivent le corpus utilisé pour répondre à leurs problématiques.

## I.4 Corpus utilisés

Les articles étudiés, datés de 1981 à 2015, utilisent des corpus inférieurs à 2200 titres. Le plus petit de (Jacques & Sebire, 2010) démarre à 300 titres, suivi de celui de (Haggan, 2004) qui compte 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

---

<sup>3</sup> Organisation européenne pour la recherche nucléaire <http://home.cern/fr>

Importants, avec pour chacun un nombre d'occurrences suffisant pour que ce phénomène ne soit pas un cas marginal ou aléatoire, un accident, mais présente bien quelque chose, un fait linguistique.

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*<sup>4</sup> qui contient de nombreuses revues, comme celui de (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) qui interroge la base *MEDLINE*<sup>5</sup> regroupant plus de 500 revues et celui de (Jacques & Sebire, 2010) qui interroge la base *Web of Science*<sup>6</sup> sur 3 journaux médicaux différents.

Certains des travaux précédents font le choix de piocher ces titres dans des disciplines proches, comme la biologie et la médecine (Aleixandre-Benavent, Montalt-Resurecció, & Valderrama-Zurián, 2014) pour augmenter le volume de leurs corpus. D'autres choisissent des disciplines qu'ils jugent très éloignés, comme littérature et 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'utiliser un grand corpus et pour cela le recours à des traitements automatiques est nécessaire. Nous devons piocher nos titres dans plusieurs disciplines, afin de pouvoir les comparer sur leurs différents traits.

## I.5 L'utilisation du double point

Nous l'avons vu, la présence d'un double point tend à être corrélée 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, à chaque fois, ce qui vient après le double point ajoute des informations sémantiques très importantes à ce qui vient d'être dit avant le double point.

(Swales & Feak, 1994) dans leur manuel à destination des étudiants de second cycle universitaire, déclarent qu'un double point dans un titre sépare les idées et ils élicitent de façon non exhaustive 4 combinaisons possibles :

- problème : solution, général : spécifique, sujet : méthode, majeure : mineure.

(Goodman, Thacker, & Siegel, 2001), qui ont recueilli les consignes données par les éditeurs de 4 journaux médicaux, pointent qu'un d'entre eux encourageait même l'utilisation du double point. (Dillon J. , 1981) prenait même sa présence comme un facteur de qualité en comparant les

<sup>4</sup> <http://mjl.clarivate.com/cgi-bin/jrnlst/jloptions.cgi?PC=K>

<sup>5</sup> <https://www.ncbi.nlm.nih.gov/pubmed>

<sup>6</sup> <https://login.webofknowledge.com>

titres de 474 articles non publiés et ceux de 314 articles publiés. À sa suite, (Townsend, 1983) confirme cette idée en trouvant deux fois plus d'utilisations du double point dans les titres publiés que dans ceux non publiés. Cependant, il détermine que l'usage d'un double point est faiblement lié à l'impact de l'article, mesuré à l'aide du nombre de citations. (Jamali & Nikzad, 2011) affirment même qu'un article avec un double point reçoit moins de citations, en contraction avec (Jacques & Sebire, 2010) qui trouvent une corrélation positive entre présence d'un double point et nombre de citations.

(Haggan, 2004) remarque que dans les titres scientifiques, il y a une haute fréquence d'utilisation des titres à deux segments séparés par 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* ») ce qui correspond à la combinaison « général : spécifique » de (Swales & Feak, 1994).

(Haggan, 2004) remarque plus globalement que les titres composés correspondent le plus souvent à cette forme 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 de ceux utilisant 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-Resurrecció, & 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 un cadre de prescriptif mais de description de l'usage. Nous avons choisi de pas nous intéresser aux sous-titres pour nous concentrer sur le titre en l'étudiant comme un objet indissociable, sans l'éclater en ses éventuels 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 de phénomènes particuliers. Toutes les études présentées traitaient des titres d'articles scientifiques en anglais, sauf (Rebeyrolle, Jacques, & Péry-Woodley, 2009) qui traitaient 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 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<sup>7</sup> (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<sup>8</sup> (CCSD), fondé en 2000 et rattaché au Centre National pour la Recherche Scientifique (CNRS). Il existe des sous-ensembles de HAL dédiés à des disciplines spécifiques, HAL-SHS et MédiHAL, ou pour un type de texte spécifique comme Thèses en ligne. Les avantages des archives ouvertes, par rapport à un site d'une institution particulière ou le site web personnel d'un chercheur, sont la centralisation de l'accès, la diffusion des connaissances et la conservation pérenne des documents. La création des archives ouvertes s'inscrit dans le mouvement pour un accès libre et gratuit aux connaissances scientifiques. La plus ancienne des archives ouvertes est arXiv<sup>9</sup>, fondée en 1991 et limitée uniquement aux articles. Un dépôt d'un article dans HAL entraîne automatiquement la création d'une notice dans arXiv s'il entre dans les disciplines couvertes par cette dernière.

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

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

---

<sup>7</sup> <https://hal.archives-ouvertes.fr/>

<sup>8</sup> <https://www.ccsd.cnrs.fr/>

<sup>9</sup> <http://arxiv.org/>

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 réplique des notices entre les 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<sup>10</sup>. 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 utilisons. 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 immédiatement 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](#) l'ensemble des requêtes utilisées.

## II.2 Traitement des données

Une fois les données brutes obtenues, nous avons effectué 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, soignée et relue, par rapport à celle d'un tweet, d'un SMS ou d'un texte sur un forum. La présence de fautes 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 nos algorithmes.

### II.2.1 Enrichissement des données

Nous voulions enrichir nos données en déterminant pour chaque forme présente dans nos titres, son lemme et sa catégorie grammaticale. Dans un premier temps, nous avons utilisé le logiciel Stanford Core Natural Language Processing<sup>11</sup>. 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<sup>12</sup> par Assaf Urieli (Urieli, 2013).

---

<sup>10</sup> <https://lucene.apache.org/>

<sup>11</sup> <https://stanfordnlp.github.io/CoreNLP/>

<sup>12</sup> <http://redac.univ-tlse2.fr/applications/talismane/talismane.html>



Avoir le lemme d'un mot permet de rassembler toutes ses formes fléchies sous une même entrée et de compter son nombre d'occurrences en additionnant celles de ses formes fléchies. La catégorie, ou étiquette POS pour *part of speech*, permet de percevoir la structure syntaxique qui est derrière les formes employées.

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 obtenus, nous procédons 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 est la transformation des données au format JSON récupérées de HAL vers un format XML propre à notre travail. Le second s'occupe 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<sup>13</sup>, issu de la conférence du même nom, ou TEI P5<sup>14</sup> 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 souhaitons maîtriser notre format, sachant que celui-ci utilise XML, sa conversion vers un autre format basé sur XML ne poserait pas de problème. La définition de son schéma est donnée dans l'annexe A6. Définition du schéma utilisé pour les corpus.

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

<sup>13</sup> <http://universaldependencies.org/docs/format.html>

<sup>14</sup> <http://www.tei-c.org/guidelines/p5/>



liste contient pour cette notice deux éléments. Le premier titre est en français et le second est sa traduction en anglais. La demande des données à HAL au format JSON est un choix historique, pour des raisons de taille, il est plus léger que XML et car nous pensions travailler uniquement dans ce format au départ. En rétrospective, demander directement à HAL ses données au format XML aurait évité une étape de conversion.

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 des 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. Les caractères spéciaux sont correctement restitués 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 relié à plusieurs disciplines. Un niveau zéro indique une discipline racine générale comme « Sciences de l'Homme et Société » ou « Sciences du Vivant ».

Le titre présenté dans cette partie est 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 étaient référencés par plusieurs notices, nous avons 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 que le français ou l'anglais.

Nous avons donc appliqué plusieurs filtres à nos données : en regardant le champ langue des notices retournées par HAL, nous avons éliminé toutes celles 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 n'avons gardé également 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, nous avons utilisé un programme de détection automatique des langues écrits en Python appelé langdetect<sup>15</sup> en ne 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, 6 448 n'avaient pas de domaines associés, 1 n'avait de type de document et 7 096 étaient des doublons. D'autres notices, au nombre de 11, avaient un titre vide et ont également été supprimées. À la fin, nous avons un corpus général de **278 806** titres, ce qui reste un nombre assez conséquent pour étudier une classe de phénomènes linguistiques particuliers dans celui-ci.

---

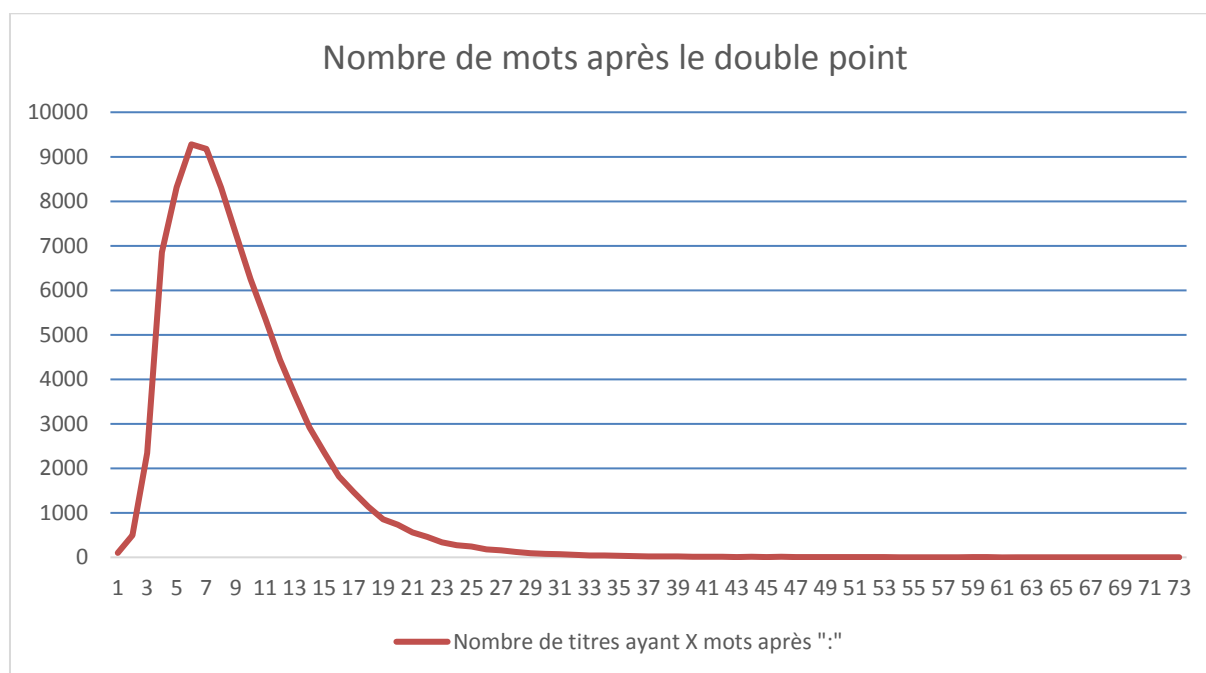
<sup>15</sup> <https://pypi.org/project/langdetect/>

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	<b>86 095</b>	2268	258	40	13	4	2	1	2

Tableau 1 : nombre de doubles points dans les titres

Il y a donc 86 095 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<sup>16</sup> 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 est 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.

<sup>16</sup> Le comptage des mots se base sur la segmentation du titre en formes opérée par Talismane, moins les formes étiquetées comme marques de ponctuation.

## II.3 Mesures du corpus

Notons que, comme le souligne (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 retournons 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(s) auteur(s), de l'année de publication du document ainsi que de ses disciplines scientifiques les plus spécialisées<sup>17</sup>, mais ces références d'exemples 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 2 200 titres. Notre corpus de travail reste toujours bien au-dessus de ce seuil avec 85 531 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 change pas entre les différents types de documents. Pour vérifier cette hypothèse, il faudrait comparer les traits des titres de nos articles à ceux des autres types de documents mais le temps nous a manqué pour cette étude.

Nous indiquons ici seulement les 7 types ayant le plus de titres, ils 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%	1,80%
Autre publication	2 147	3%	2,29%

Tableau 2 : 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.

<sup>17</sup> Si un titre est lié par son document aux disciplines de Sciences du vivant et Ingénierie des aliments, une sous-discipline de la première, nous ne mentionnons que la discipline la plus spécialisée, l'Ingénierie des aliments.

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. Elles 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 3 : Répartition des titres par année

**Exemple :** QUELQUES REMARQUES SUR L'ÉTUDE DE CH. BARTHEL: INFLUENCE DES MOISSURES SUR LES FERMENTS LACTIQUES (Dr Jaroslav Dvorak, 1925, Alimentation et Nutrition - Ingénierie des aliments)

L'exemple ci-dessus est le plus vieux titre de notre corpus. On remarque déjà une utilisation du double point et l'utilisation inhabituelle des majuscules.

Si les documents vont de 1925 à 2018, les années les plus récentes sont les plus fournies : 2018 et 2017 représentent à elles seules 89% de corpus. Cette fenêtre trop réduite 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.

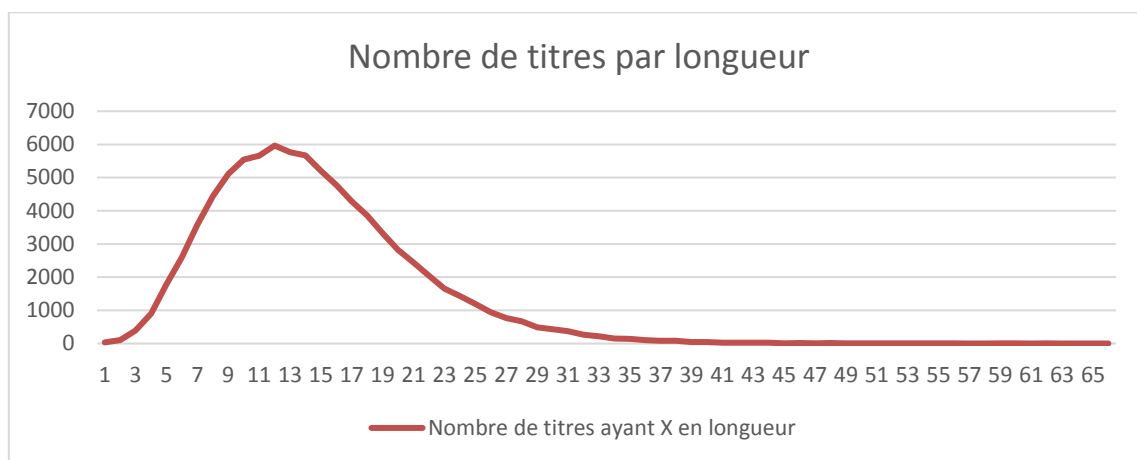
### II.3.3 Longueurs des titres et nombre d'auteurs

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

**Exemple :** L'<sub>1</sub> interprétation<sub>2</sub> langue<sub>3</sub> vocale<sub>4</sub> (LV5)//langue<sub>6</sub> des<sub>7</sub> signes<sub>8</sub> (LS<sub>9</sub>) et<sub>10</sub> la<sub>11</sub> question<sub>12</sub> du<sub>13</sub> "lexique<sub>14</sub> " : inverser<sub>15</sub> le<sub>16</sub> regard<sub>17</sub> ! (Brigitte Garcia, 2018, Linguistique)

Dans l'exemple ci-dessus, nous indiquons en rouge et gras 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 :

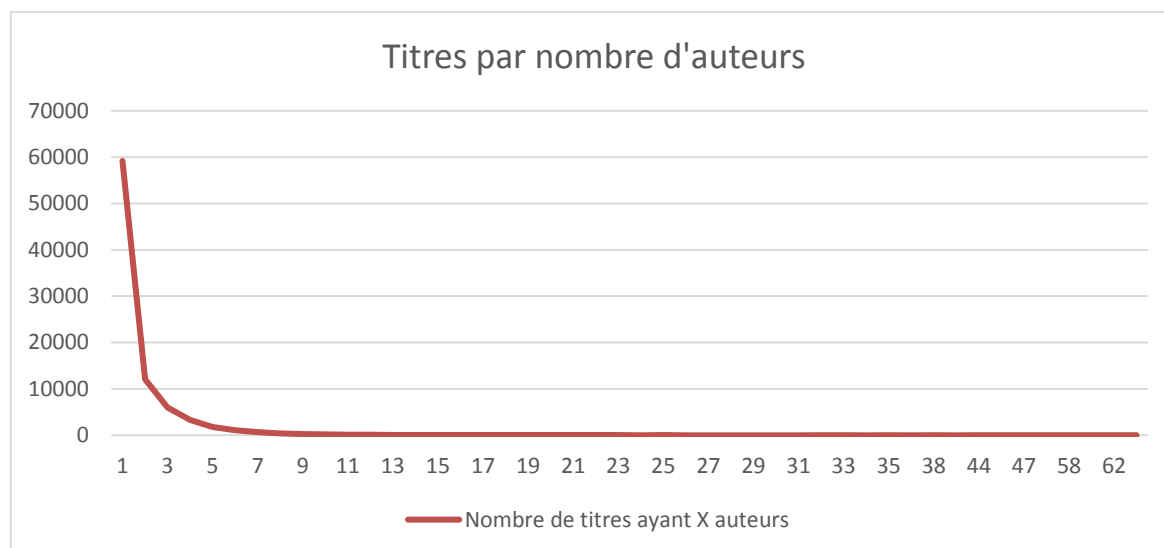


Nous regardons à présent le nombre d'auteurs par document scientifique. Les 8 nombres les plus fréquents sont exprimés dans le tableau suivant :

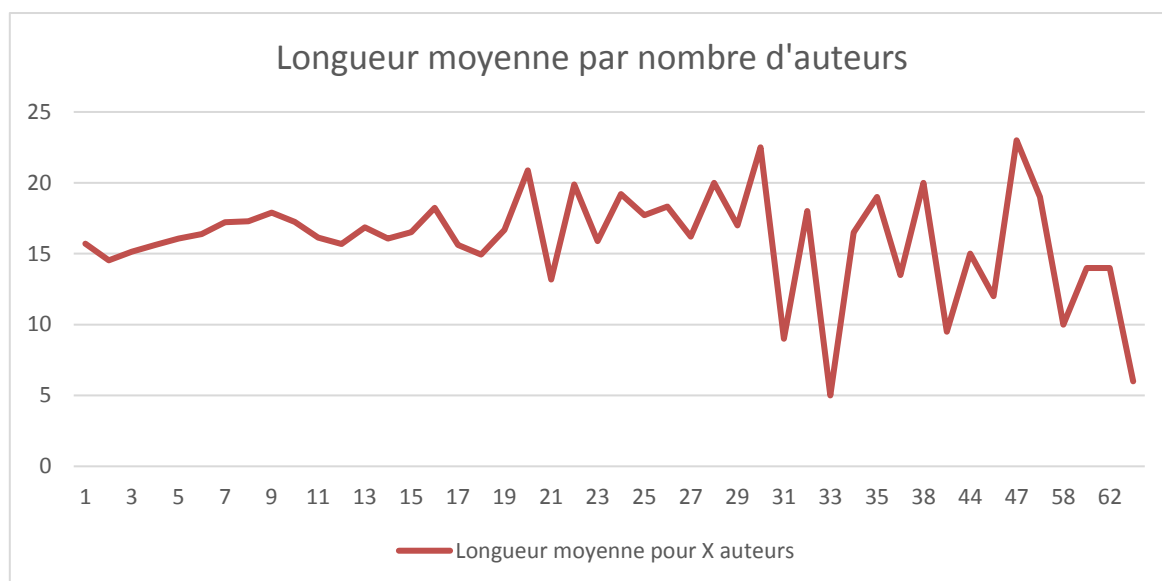
Nombre d'auteurs	1	2	3	4	5	6	7	8
Nombre de titres (%)	59 182 69%	12 035 14%	6 015 7%	3 310 4%	1 765 2%	1 065 1%	689 1%	415 < 0.5%

Tableau 4 : Nombres de titre par nombres d'auteurs

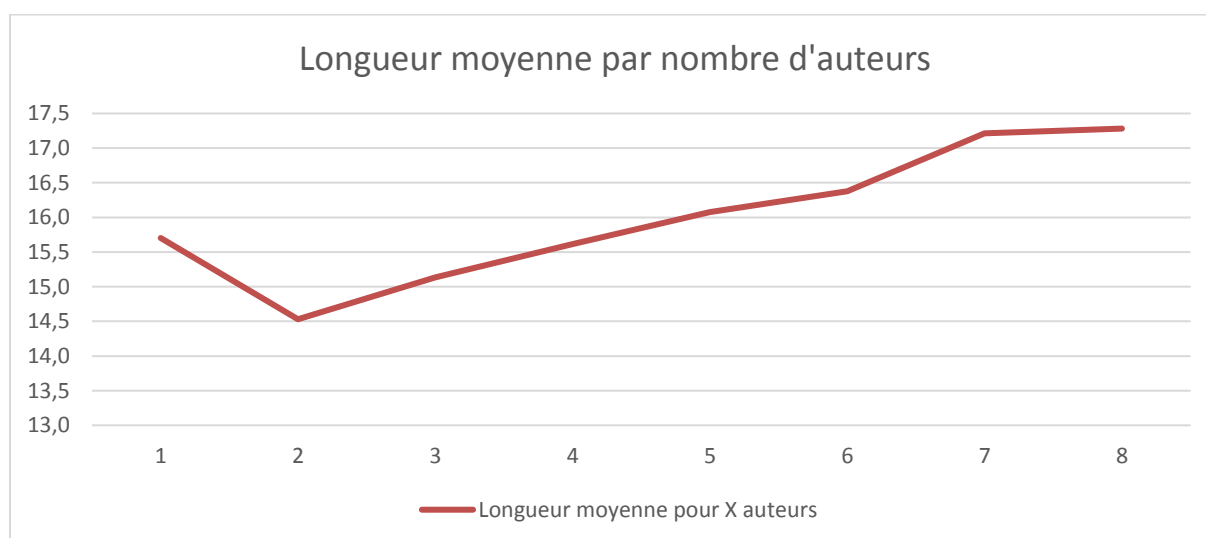
Si le nombre d'auteurs dans notre corpus va de 1 à 147, 98% des titres ont néanmoins entre 1 et 7 auteurs et 69% ont un seul auteur, pour une moyenne de 1,8 auteurs par titre.



(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 nombre 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 bien une amorce de plateau, mais situé à 17 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 jusqu’à 6, où elle s’accélère. Les titres avec 1 auteur ont une longueur moyenne de 15,5 alors que ceux avec 2 auteurs ont une longueur moyenne qui s’établit à 14,5.

(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	<b>86 095</b>	2268	258	40	13	4	2	1	2
Longueur moyenne des titres	12	<b>16</b>	24	32	52	99	97	75	90	87

Tableau 5 : Nombre de titres par nombres de doubles points et longueurs moyennes

Notre corpus confirme cette affirmation : la présence d'un double point ou plus augmente la longueur moyenne des titres. On peut remarquer que plus il y'a de doubles points, plus la longueur moyenne augmente jusqu'à 5 doubles points. 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, les multiples racines de l'arbre des disciplines. Le terme de « *domain* » en anglais est privilégié par HAL, nous utiliserons indifféremment domaine et discipline pour les désigner. Certains titres sont apparentés à plusieurs domaines :

Domaine	Code	Nombre	Domaine	Code	Nombre
Sciences de l'Homme et Société	shs	80 441	Sciences cognitives	scco	1 354
Sciences du Vivant	sdv	15 614	Chimie	chim	1 051
Informatique	info	5 244	Mathématiques	math	877
Sciences de l'ingénieur	Spi	4 469	Économie et finance quantitative	qfin	283
Sciences de l'environnement	sde	3 414	Statistiques	stat	228
Planète et Univers	sdu	2 157	Science non linéaire	nlin	22
Physique	phys	2 013			

Tableau 6 : 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, Éducation - Psychologie)

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/Éducation
1	shs.psy	Sciences de l'Homme et Société/Psychologie

Tableau 7 : Analyse des domaines de l'exemple

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 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<sub>1</sub> ; méthodes approchées, cinématiques<sub>2</sub> ; Analyse Modale<sub>3</sub> ; Recalage de Modèle<sub>4</sub>* (Jean-Michel Gènevaux, 2018, Sciences de l'ingénieur - Mécanique des structures)

Ce titre complexe est composé de 4 partitions, nous avons indiqué les trois marques de ponctuation qui le segmente en rouge et gras. 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 possèdent *au moins une fois* la marque de ponctuation indiquée. On détermine ensuite sur ce nombre, combien de titres ont cette marque en dernière position. Enfin, on calcule, pour les titres possédant une marque donnée, la moyenne des occurrences de cette marque dans le titre.

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 8 : 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 que les moyennes des occurrences par titre 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 remarque que 84% des points d'interrogation sont en position terminale, transformant le titre en question, ou du moins son segment terminal. Cette proportion tombe à 55% pour le point d'exclamation.

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'établit 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*<sub>1</sub>: cas de la dinanderie, de la poterie et de l'ébénisterie-marqueterie<sub>2</sub> (Améziane Ferguene et Abderrahmane Bellali, 2018, Économies et finances - Sociologie)

Ce titre présente une organisation typique en deux segments séparés par un double point en rouge. Nous notons qu'il n'y a pas d'espace avant celui-ci, à la manière de l'anglais, alors que les règles typographiques françaises en imposent 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 divisons 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 titres 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 9 : 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 est 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 d'occurrences 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 10 : Comptes des noms communs les plus fréquents 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 (Laure Pajean, 2018, Médecine humaine et pathologie)

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

Tableau 11 : Tableau des noms communs les plus fréquents avec pourcentage d'occurrences après le double point

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 en ordonnée :

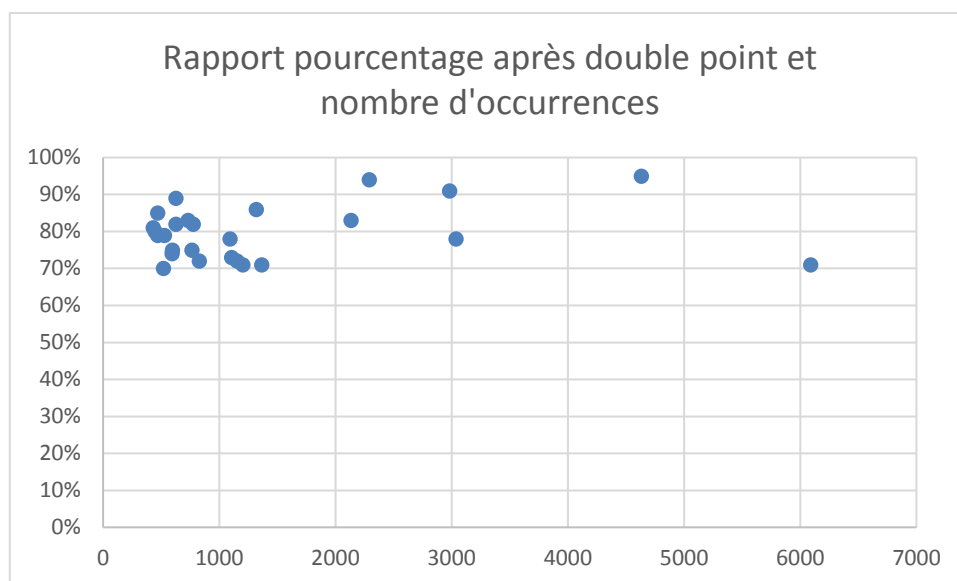


Figure 1 : Pourcentage d'occurrences après le double point et nombre total d'occurrences des noms communs les plus fréquents

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. Dans les tableaux des parties suivantes, l'appartenance d'un nom à cette classe sera rappelée en le suffixant d'un astérisque comme « état\* ».

Dans le cadre d'une typologie reposant sur la fonction référentielle (Huyghe, 2015), on peut associer tous les noms de notre classe à des noms généraux. Selon cet auteur, ces noms se distinguent par un très faible contenu sémantique, on ne sait pas vraiment à quoi fait référence une « approche » dans le monde réel. Cette « *pauvreté de leur contenu sémantique* », ce manque de « *spécifications sémantiques* », permet en retour d'avoir une « très large application référentielle ». Ces noms peuvent servir à dénoter énormément de référents, l'auteur parle de « polyvalence référentielle ».

Ces noms, appelés « *shell noun* » par (Schmid H.-J. , 2000) ont, d'après ce même auteur, pour particularité d'avoir une haute fréquence, avec la spécificité que, dans notre corpus, ils se trouvent en grande majorité après le double point. (Halliday & Hasan, 1976) en donne une liste pour l'anglais ainsi que (Hinkel, 2004, p. 274) : nous retrouvons les noms « approche » (*approach*), « élément » (*item*), « expérience » (*experience*), « résultat » (*result*) de notre classe.

La question de savoir si ces noms généraux sont sous-spécifiés et ce qu'il dénote dépasse le cadre de notre travail. Pour pouvoir néanmoins répondre à ces questions, il faut avant tout identifier leur contexte. En effet, « la fonction de noms [généraux] sous-spécifiés s'acquiert qu'en relation avec une construction spécificationnelle » selon (Adler, 2018) et c'est le contexte qui donne la clé de son interprétation référentielle. Dans ce travail, nous allons essayer d'établir quels sont ces contextes qui pourraient être récurrents avec une grande fréquence, puisque au moins un nom qui le compose à

une forte fréquence après le double point. Nous étendrons nos observations également aux contextes récurrents ne possédant pas de noms de notre classe, pour avoir une vue d'ensemble du phénomène des récurrences de syntagme après le double point.



La méthode présentée pour obtenir notre corpus dans cette partie est reproductible et permet d'obtenir de nouveaux corpus 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, les titres avec plus de segments étant relativement rares.

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. Ce sont des noms généraux. La plupart sont en position deux après le double point. Un seul mot les sépare du double point sur leur gauche, on peut formuler l'hypothèse qu'il s'agisse 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 Séquences d'étiquette POS et syntagmes

Talismane a catégorisé les différentes formes des titres. Pour chacune, nous avons son lemme et sa catégorie grammaticale, exprimée par une étiquette POS. À chaque titre correspond donc une séquence d'étiquettes POS. Nous considérons dans notre travail uniquement les étiquettes venant après le double point.

**Exemple :** « *La rue et l'écran : la négociation de l'intimité* » (Marianne Trainoir, 2018, Éducation)

Ce titre a pour séquence d'étiquettes POS après le double point : « DET NC P DET NC »<sup>18</sup>. Cette séquence est la représentation linéaire d'un syntagme.

En synthétisant les définitions de (Maingueneau, Chiss, & Filliolet, 2007, p. 35) et (Neveu, 2017), nous définissons le syntagme comme un groupe de lemmes consécutifs constituant une unité syntaxique, organisé autour d'un lemme noyau et s'inscrivant dans une organisation hiérarchisée. La catégorie de ce noyau donne le type du syntagme et « le syntagme exerce les mêmes fonctions syntaxiques que son noyau » (Neveu, 2017).

L'analyse syntagmatique<sup>19</sup>, montre que malgré la linéarité de cette séquence, le syntagme s'organise en une structure hiérarchique à plusieurs niveaux représentable « à l'aide de parenthèses,

---

<sup>18</sup> Nous utilisons dans ce document les étiquettes de Talisman pour les catégories. Celles utilisées ici sont DET pour déterminant, NC pour nom commun, P pour préposition. Il existe également ADJ pour adjectif qualificatif. La liste complète est donnée dans l'annexe A2. Codes des étiquettes de catégorie de discours de Talisman.

<sup>19</sup> Cette analyse est au cœur de l'analyse en constituant immédiat d'une phrase. Néanmoins, nous n'étudions pas des phrases entières mais des titres qui en sont rarement.

*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.

Notre exemple est, après le double point, constitué d'un syntagme nominal, qui a pour noyau le nom commun « négociation », qui 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 « intimité ».

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

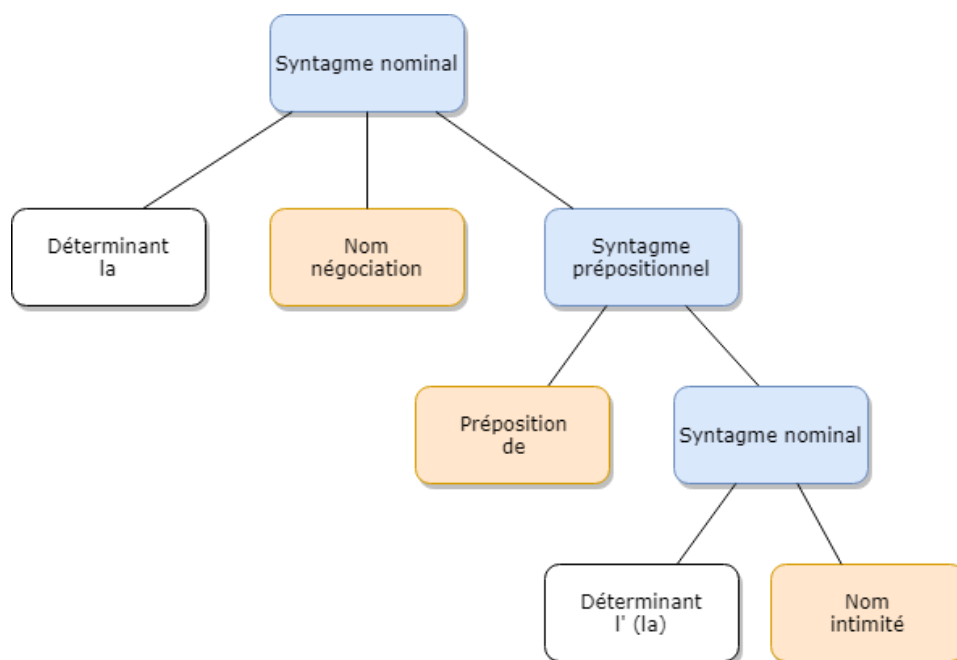


Figure 2 : arbre d'analyse syntagmatique

Le passage de la séquence linéaire d'étiquettes à la structure soulève des ambiguïtés. 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. Les structures des deux syntagmes sont les suivantes :

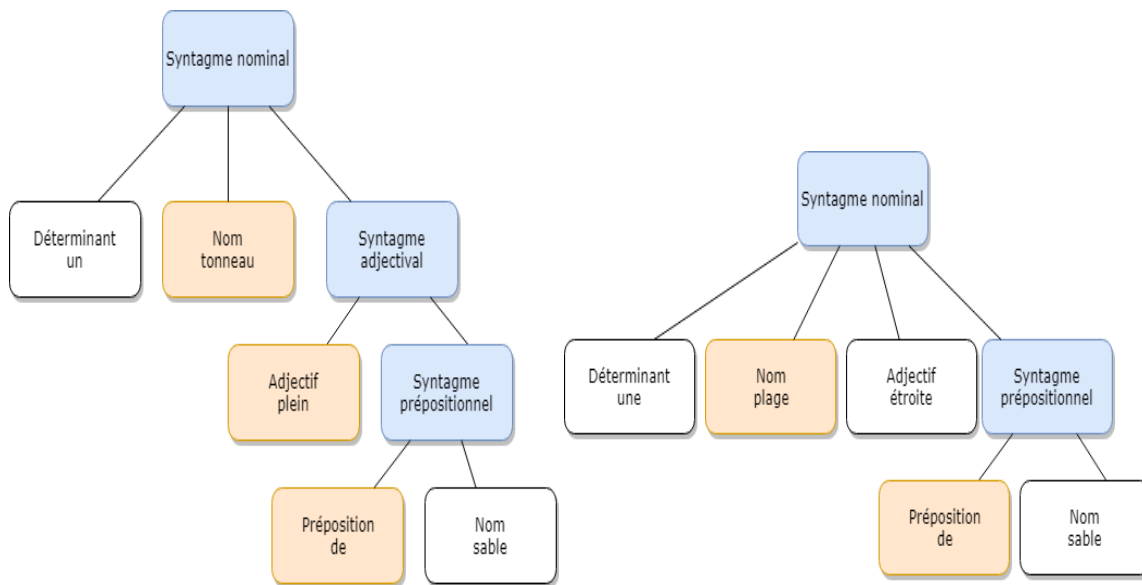


Figure 3 : Syntagmes possibles pour "DET NC ADJ P NC"

Après le double point, le titre peut comporter encore jusqu'à 29 lemmes, ce qui correspond à la longueur maximale de la séquence. Nous avons inventorié toutes les séquences existantes d'étiquettes POS après le double point en comptant leurs nombres d'occurrences. Celles-ci sont au nombre de 45 098. À certaines de ses séquences peut donc correspondre plusieurs structures possibles, potentiellement complexes avec de nombreux niveaux. Il est donc important de poser des limites à notre étude pour garantir sa faisabilité.

### 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 très proche de celui-ci. 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.

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.

Pour couper, 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 2 156 séquences qui sont jugées équivalentes.

**Exemple :** CAN : la culture à Nancy au prisme de ses habitants. *Pratiques, regards, symboles* (Cécile Bando, Lyette Lacote-Gabrysiak & Adeline Clerc-Florimond, 2018, Sciences de l'information et de la communication)

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 et gras de la suite est ignorée, la suite devient donc équivalente à celle ayant seulement les étiquettes en noir.



Pour regrouper, nous parcourons 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 12: 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 fournit 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 un patron syntaxique fini comme un modèle  nissant un ensemble de séquences d'étiquettes POS. La caractéristique principale du patron est sa variabilité. Le modèle  nit que certaines étiquettes POS sont obligatoires, d'autres répétées, que certaines peuvent apparaître de façon optionnelle et enfin que d'autres doivent être choisies entre plusieurs alternatives.

Un patron peut être utilisé de deux façons : si les séquences sont préexistantes, on peut tester si la séquence correspond au patron, on dit que le patron *capture* la séquence. On peut alors regrouper toutes les séquences capturées. C'est le cas de notre étude, les séquences sont dans nos titres. À l'inverse, on peut utiliser un patron pour *générer* des séquences. Nos patrons sont finis car l'ensemble de séquences générées est fini, on peut les dénombrer.

Pour représenter nos patrons, écrire leurs modèles, 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'optionnalité : l'étiquette A peut apparaître une fois mais ce n'est pas obligé.
4. **( A B )** permet de grouper les étiquettes POS



- a. à l'intérieur d'un choix :  $[ (A B) C ]$  signifie soit A suivi de B, soit C
  - b. 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 séquences de notre corpus.

**Exemple :** Le patron qui correspond à la fois au syntagme illustré par la *Figure 2 : arbre d'analyse* et aux séquences dans le

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 12: 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 peut 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 et seulement celles-ci. Notre but est de capturer toutes celles qui s'y trouvent et y correspondent, pour 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 précédent génère 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	

Tableau 13 : Séquences générées par notre patron

### III.4 Limites de nos patrons

Le terme de patron a déjà été utilisé par (Hunston & Francis, 2000) qui s'inscrivent dans une perspective didactique et descriptive, dirigée par les corpus. Elle remonte aux descriptions pédagogiques de l'anglais par (Hornby, 1954) et au travail sur corpus de (Sinclair, 1991). Cette école contextualiste a été étudiée par (Legallois, 2006).

En première différence, nos patrons, comme outils, portent uniquement sur le niveau syntaxique, bien que notre démarche, avec notre classe de noms, s'inscrive également au niveau lexical. (Hunston & Francis, 2000) définissent aussi bien des patrons uniquement syntaxiques comme « V n », un verbe suivi d'un syntagme nominal, que des patrons lexico-syntaxique comme « v-link ADJ about n », ou v-link désigne un verbe d'état.

Une première limitation de nos patrons est que nous n'incluons pas le niveau lexical dans, bien que cela soit techniquement possible. Nous aurions pu par exemple demander obligatoirement à avoir un ou des noms de notre de classe.

Nous n'avons pas fait cela, c'est pour voir ce qui émerge du corpus. Nous émettons l'hypothèse que nous retrouverons nos noms mais nous limiter à ceux-ci, c'est prendre le risque de manquer des phénomènes. Notre approche est *corpus driven* : ce que l'on observe dans le corpus dirige notre élaboration théorique, au lieu de servir de confirmation ou d'infirmer à une théorie construite a priori.

(Legallois, 2006) explique que pour les auteurs de l'école contextualiste « *la dimension de ce patron est indépendante de la notion de syntagme* », ce qui est le cas pour les nôtres également et constitue une seconde limitation : nos patrons capturent des séquences et non des structures. Pour mieux voir cette ambiguïté, si nous ajoutons un adjectif, en rouge et gras, au patron défini plus haut, nous obtenons le patron suivant :

DET? NC <b>ADJ?</b> [ (P DET?) P+D ] NC ADJ?
--

Ce patron capture toute aussi bien « Une plage étroite de sable » qu'un « Un tonneau plein de sable », alors que les structures des deux syntagmes sont différentes. 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 restons conscients de cette ambiguïté lors de notre analyse des résultats fournis par les scripts automatiques. Plus globalement nous considérons nos patrons syntaxiques finis comme des outils techniques auxquels nous ne prêtons pas un contenu sémantique supplémentaire par rapport à celui apporté par les lemmes qui le composent. Cela ne veut pas dire que, dans les séquences que nous capturons, la structure syntaxique n'apporte pas un contenu sémantique, mais nous verrons cela lors du dépouillage des résultats des patrons au chapitre V. Avant cela, nous devons à présent décrire les trois patrons que nous avons élaborés, ce sera l'objet du chapitre suivant.



Nous avons exposé nos patrons ainsi que leurs limitations : ne considérer que le niveau syntaxique et de ne capturer dans celui-ci 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 couvrent la majorité de notre corpus, chacun capturant un syntagme comprenant deux noms, dans le but de retrouver le contexte environnant les noms de notre classe et éventuellement déceler d'autres récurrences, dépassant le cadre d'un seul lemme.

## IV. Études des trois patrons

### IV.1 Définition et construction des trois patrons

#### IV.1.1 Présentation des 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-ci de se limiter arbitrairement dans son analyse.

Les trois types de syntagmes que nous avons décidé d'étudier, à partir de l'observation du corpus, 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 correspond 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.

#### IV.1.2 Construction itérative du patron SN

Par exemple, pour le patron SN, nous avons surligné 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 14: 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-ci effectuait deux opérations :

- 1) Le comptage automatique de la couverture du patron. Il y a deux types de couvertures :
  - a. Le nombre de séquences d'étiquettes POS couvertes par le patron par rapport au nombre total de séquences inventoriées (42 942)
  - b. Le nombre de titres auxquels il correspond, par rapport au nombre total 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% d'entre elles, ne sont utilisées que par un seul titre.

- 2) 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, ainsi 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.

**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 déterminé 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 :

[**DETH** 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 séquences 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.

#### IV.1.3 Explication des séquences [NC NPP] [NC NPP]?

On remarque que notre patron propose le choix d'un nom optionnel juste après le premier nom : [NC NPP]?. Il ne faut pas prendre celui-ci pour notre second nom. En français, la juxtaposition de deux noms communs ne nous semble pas correcte : \*« table vin ». Seule la juxtaposition de deux formes formant un seul nom propre est correcte « André Martinet ». Talismane n'assimile pas les formes en une seule et les étiquette toutes les deux « NPP » pour nom propre : « André<sub>NPP</sub> Martinet<sub>NPP</sub> ». Nous n'avons pas exclu les noms propres de notre travail pour deux raisons :

- 1) Ils sont beaucoup moins nombreux que les noms communs, ils ne les cacheront pas dans nos résultats si l'on regarde seulement les noms les plus fréquents.
- 2) Nous voulions repérer d'éventuelles récurrences également pour les noms propres, pour un éclairage supplémentaire sur celles-ci.

Pour en revenir à la possibilité de deux noms communs qui se suivent, il s'agit d'un contournement d'un défaut de Talismane : celui a tendance à étiqueter un adjectif comme un nom commun comme dans « revue critique » qui est étiqueté « NC NC » au lieu de « NC ADJ ». Pour ne pas perdre cette séquence, nous avons autorisé une autre étiquette NC après le premier nom. La même chose s'applique au second nom après la préposition.

#### IV.1.4 Exclusivité mutuelle des 3 patrons

Nos trois patrons sont mutuellement exclusifs par leurs constructions : une suite ne peut s'accorder qu'avec un seul d'entre eux ou aucun. Cette exclusivité est possible car notre patron SNC n'autorise pas d'expansion du nom de type syntagme prépositionnel pour son premier nom. Ainsi la suite « **N P N** CC N » s'accorde avec le patron SN et non le patron SNC. Cela dans le cadre de notre étude qui se limite aux deux premiers noms rencontrés après le double point.

Dans les trois parties qui suivent, nous présentons nos 3 patrons. À chaque fois nous présentons une fiche d'identité avec différents champs. Le premier est une courte description. Le second est la structure « idéale » du syntagme : celle à laquelle nous pensions en écrivant le patron, mais, à la suite de la limitation de ceux-ci, nous savons que nous pouvons obtenir d'autres structures syntagmatiques. Leur énumération dépasse le cadre de ce travail. Ensuite vient le patron minimal, les constituants obligatoires que l'on retrouve dans le patron étendu qui propose beaucoup plus de variabilité afin de capturer un maximum de séquences pour pouvoir observer le plus de phénomènes. Nous calculons ensuite la couverture du patron, en nombre de séquences couvertes et en nombre de titres couverts. Enfin, nous proposons trois exemples de titres qui correspondent au patron.

## IV.2 Patron SN : syntagme nominal

### IV.2.1 Fiche d'identité

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

#### Structure minimale idéale

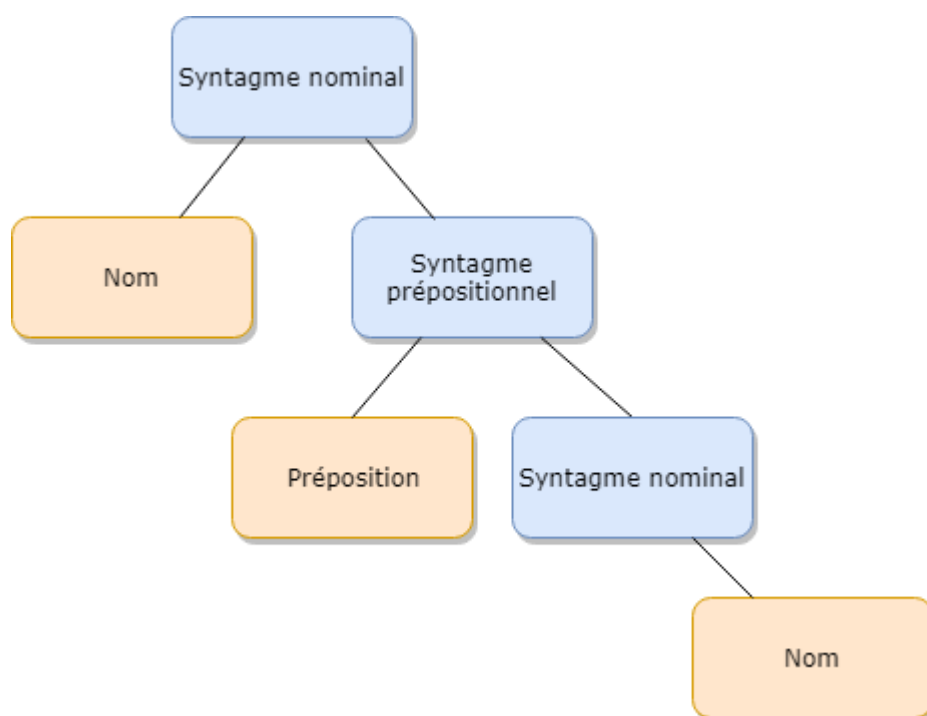


Figure 4 : Structure minimale du syntagme idéal pour le patron SN

#### Patron minimal

NC P NC

#### Patron étendu

[DETH 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

- *Représentations et images des villes de la Renaissance: l'exemple des cartes de Nancy* (Jean-Pierre Husson, 2018, Sciences de l'Homme et Société<sup>20</sup>)
- *Regard sur l'histoire de quelques prépositions de l'anglais contemporain : Apport de la diachronie* (Anne Mathieu, 2018, Sciences de l'Homme et Société<sup>14</sup>)
- *L'ethos collectif des professeurs documentalistes sur Twitter : exploration de quelques pratiques* (Florence Thiault, 2018, Sciences de l'information et de la communication)

### IV.3 Patron SP : syntagme prépositionnel

#### IV.3.1 Fiche d'identité

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

#### Structure minimale idéale

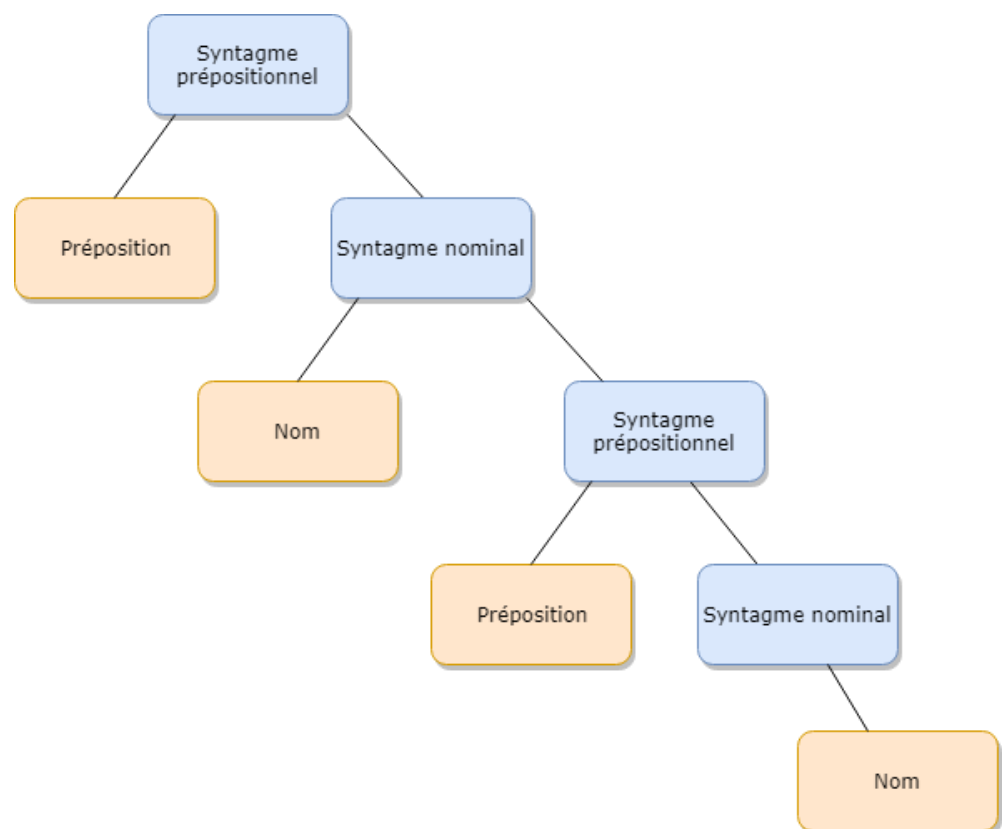


Figure 5 : structure du syntagme idéal pour le patron SP

#### 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?

<sup>20</sup> Ces titres sont sous-spécifiés au niveau de leurs domaines scientifiques. Le deuxième devrait être en Linguistique.

<b>Possibilités</b>	6 912 séquences, longues de 4 à 12 étiquettes
<b>Couverture du corpus</b>	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

- *Analogie et dynamiques discursives du figement/défigement : aux sources de la créativité lexicale et de l'économie linguistique en langue des signes française (LSF)* (Brigitte Garcia, 2018, Linguistique)
- *De la salle de cinéma à la caverne : à propos de quelques tentatives artistiques d'ensevelissement* (Thibault Honoré, 2018, Art et histoire de l'art)
- *La co-construction textuelle avec de jeunes enfants : de la phrase au texte, et vice versa* (Frédéric Torterat, 2018, Linguistique - Éducation)

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

### IV.4.1 Fiche d'identité

<b>Description</b>	Il s'agit d'un syntagme nominal avec une coordination incluant deux sous-syntagmes nominaux.
--------------------	--

#### Structure minimale idéale

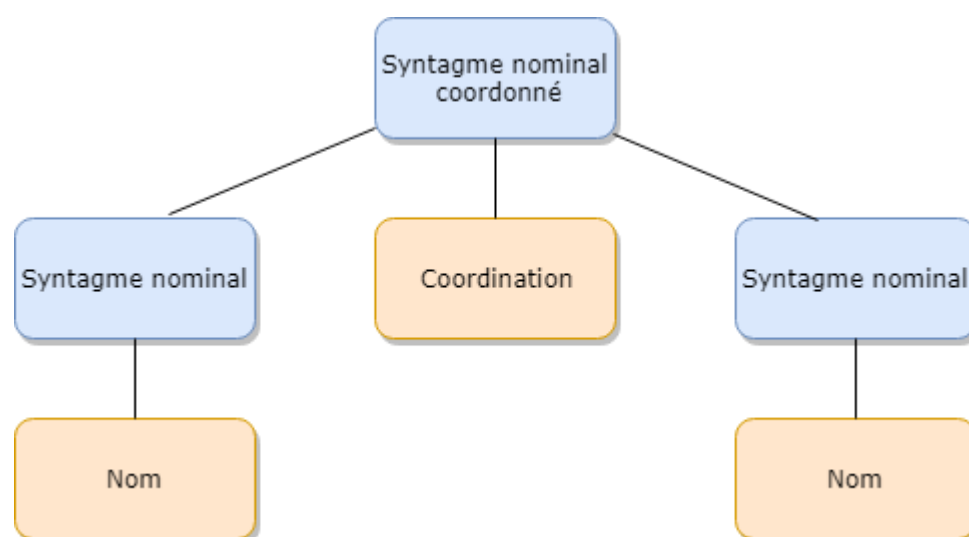


Figure 6 : structure du syntagme idéal pour le patron "NC CC NC"

<b>Patron minimal</b>	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?
---



<b>Possibilités</b>	1 728 séquences, longues de 3 à 10 étiquettes
<b>Couverture du corpus</b>	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, Éducation)
- *Les écrivains de langue russe en exil : plurilinguisme et autotraduction* (Anna Lushenkova Foscolo, 2018, Littératures)
- *L'actrice, du cinéma aux arts plastiques et au roman : genre et représentation* (Thibaut Casagrande, 2018, Littératures)

## IV.5. Couverture globale du corpus

Nos trois patrons couvrent de façon exclusive :

Patrons	Couverture des séquences	Couverture des titres
Patron SN	49,35%	49,81%
Patron SP	4,81%	4,70%
Patron SNC	7,35%	10,34%
<b>Total</b>	<b>61,51%</b>	<b>64,85%</b>

Tableau 15 : 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 ce qui nous permet de détecter certains phénomènes récurrents.



Nous avons utilisé notre langage pour définir trois patrons. Tous ont pour particularité de capturer des séquences contenant deux noms, seule la constitution de la séquence varie. L'un correspond à un syntagme nominal, le deuxième à un syntagme prépositionnel et le troisième est constitué de deux syntagmes nominaux coordonnés.

Nos trois patrons couvrent 64,85% des titres de notre corpus de travail. Assez pour remarquer un phénomène de répétitions de contextes dans lesquels s'inscrivent les noms de notre classe. Dans le chapitre suivant, nous nous proposons d'étudier ces contextes.

## V. Analyse syntactico-lexicale des résultats

Dans la partie précédente, nous avons étudié uniquement les catégories et nos patrons portaient uniquement sur celles-ci. À présent, nous voulons étudier les lemmes qui peuplent les séquences capturées par nos patrons.

## V.1 Résultats du patron SNC

### V.1.1 Fréquences des coordinations

Au niveau de la coordination, nous avons 4 coordinations qui se répartissent ainsi :

Coordination	et	ou	or	mais
Occurrences	8145	688	6	6
Pourcentage	92%	8%	< 0,1%	< 0,1%

Tableau 16 : Tableau des fréquences de la coordination du patron SNC

La conjonction de coordination « et » écrase toutes les autres. Le « or » provient de 4 titres en anglais qui ont réussi à passer nos filtres et de 2 titres français utilisant une citation anglaise dans leur construction : « *To Be or Not to Be* » et « *to join or not to join* » respectivement.

### V.1.2 Fréquences des noms et emplacements

Si on regarde le premier nom, nom 1, on remarque que les 8 plus fréquents sont :

Lemme	enjeu*	approche*	étude*	pratique	modélisation	analyse	bilan	mythe
Occurrences	278	110	105	91	84	84	75	69
Pourcentage	3%	1%	1%	1%	1%	1%	1%	1%

Tableau 17: Tableau des fréquences des lemmes en première position

Si on regarde le second nom, nom 2, on remarque que les 8 plus fréquents sont :

Lemme	perspective*	enjeu*	application*	pratique	limite	réalité*	modélisation	représentation
Occurrences	288	204	186	147	127	103	100	81
Pourcentage	3%	2%	2%	2%	1%	1%	1%	1%

Tableau 18 : Tableau des fréquences des lemmes en seconde position du patron SNC

Nous faisons suivre les noms appartenant à notre classe de noms ayant une forte affinité pour la position après le double point d'une étoile \*. Ils sont au nombre de 3 sur les 8 plus fréquents nom en première position, et de 4 sur 8 pour ceux en seconde position.

Certains noms semblent donc avoir une position interchangeable : ils peuvent être indifféremment mis à gauche ou à droite de la conjonction de coordination comme « enjeu », « pratique » et « modélisation ». D'autres semblent avoir une affinité plus grande pour un emplacement donné. Nous comparons la répartition des 8 noms les plus fréquents pour chaque position :

Lemme	Occurrences	Avant CC	Après CC
enjeu*	482	58%	42%
perspective*	308	6%	94%
pratique	238	38%	62%
application*	204	9%	91%
modélisation	184	46%	54%
étude*	149	70%	30%
limite	141	10%	90%

Tableau 19 : répartition des lemmes les plus fréquents avant et après le double point du patron SNC

On constate que certains noms ont une affinité très grande pour la seconde position : « perspective », « application » et « limite ». Pour la recherche d'affinité avec la première position, les lemmes moins fréquents « bilan », « mythe » et « principe » sont respectivement à 91%, 95% et 93% en première position, pour un nombre total d'occurrences supérieur à 50 chacun. Nous voulons à présent étudier les couples de noms, de façon ordonnée.

### V.1.3 Fréquences des couples de noms ordonnés

#### A) Quel nom 1 sachant nom 2

Pour les trois noms ayant une nette préférence pour la seconde position, nous voulons savoir quels noms sont en première position :

- Pour « perspective », deux lemmes se détachent : « enjeu » (18%) et « bilan » (17%), les autres combinaisons ne dépassant pas 3,5%.
- Pour « application », les quatre premiers lemmes associés sont : « développement » (9%), « méthode » (8%), « théorie » (7%) et « principe » (7%). Les autres associations ne dépassent pas 5%.
- Pour « limite », trois lemmes se détachent : « intérêt » (19%), « apport » (17%) et « enjeu » (10%). Les autres associations ne dépassent pas 4%.

#### B) Quel nom 2 sachant nom 1

Pour les trois noms ayant une nette préférence pour la première position, nous voulons savoir quels noms sont en deuxième position :

- Pour « bilan », un lemme est ultra-majoritaire : « perspective » à 67%. Les autres combinaisons ne dépassant pas 3%.
- Pour « mythe », un lemme est ultra-majoritaire : « réalité » à 81%. Les autres combinaisons ne dépassant pas 3%.
- Pour « principe », deux lemmes se détachent : « application » à 26% et « méthode » à 10%. Les autres ne dépassant pas 4%.

Nous voulons à présent étudier l'ensemble du triplet (nom 1, conjonction de coordination, nom2).

### V.1.4 Fréquence des triplets

#### A) Triplet avec conjonction de coordination indifférente

En regardant les fréquences des couples (nom 1, nom 2) et celles des triplets (nom1, cc, nom2) nous remarquons une tendance à utiliser les deux mêmes noms mais à les joindre alternativement par « et » ou « ou ». Cela concerne les couples de noms :

Nom 1	Nom 2	Occurrence avec « et » (%)	Occurrence avec « ou » (%)
mythe	réalité*	23 (41%)	33 (59%)
rupture	continuité	12 (50%)	12 (50%)
continuité	rupture	4 (40%)	6 (60%)

Tableau 20 : fréquence des triplets avec un choix de conjonction de coordination du patron SNC

Ces trois couples ordonnés (mythe, réalité), (rupture, continuité) et (continuité, rupture) forment l'exception : les autres couples ont une forte tendance à avoir une conjonction de coordination préférée, généralement « et ».

### B) Triplet avec conjonction de coordination préférée « ou »

Les rares couples les plus fréquents préférant « ou » sont (évolution, révolution) avec 6 occurrences, (menace, opportunité) avec 5, (rêve, réalité) avec 4 et (réalité, fiction) avec 4. On remarque que chacun des termes est en opposition sémantique forte, jusqu'au point d'être des antonymes pour (rêve, réalité) et (réalité, fiction).

### C) Triplet avec conjonction de coordination préférée « et »

Le reste des triplets utilise principalement la conjonction de coordination « et ». C'est le cas pour ceux qui correspondent aux couples de noms fréquents que nous avons identifiés, (mythe, réalité) mis à part. Nous voulons savoir si leurs fréquences sont significatives par rapport aux autres triplets :

Nom 1	CC	Nom 2	Occurrences	Fréquences
enjeu*	et	perspective*	51	0,58%
bilan	et	perspective*	50	0,57%
phénomène	et	configuration	33	0,37%
mythe	ou	réalité*	33	0,37%
intérêt	et	limite	24	0,27%
mythe	et	réalité*	23	0,26%
apport*	et	limite	21	0,24%
théorie	et	pratique	17	0,19%
pratique	et	représentation	16	0,18%
enjeu*	et	défi*	16	0,18%
développement	et	application*	16	0,18%

Tableau 21 : fréquences des triplets les plus fréquents du patron SNC

Il y a un grand étalement des triplets : aucune fréquence ne dépasse 0,6%. Cela ne vient pas d'une variabilité de la conjonction de coordination, elle est généralement « et », mais bien de la variété des séquences de noms.

Nous voulons savoir si, en ne regardant pas la position des noms mais uniquement la présence dans un couple, nous détachons des utilisations récurrentes.

#### V.1.5 Fréquences des couples de noms non ordonnés

Dans cette étude, nous regardons les couples de noms mais sans tenir compte de leurs positions (première ou seconde). Nous avons regardé les couples avec une fréquence supérieur à 10, soit plus de 0,1% de notre corpus. Nous donnons, pour chaque couple, son premier membre A, son second membre B, le nombre d'occurrences totales qui est l'addition des occurrences de (A, B) et (B, A) et le pourcentage de couples dans l'ordre (A, B) :

Premier membre A	Second membre B	Occurrences	% dans ordre
mythe	réalité*	56	100%
enjeu*	perspective*	51	98%
bilan	perspective*	50	100%
phénomène	configuration	33	94%

rupture	continuité	24	71%
Intérêt	limite	24	<b>100%</b>
apport*	limite	21	<b>100%</b>
théorie	pratique	17	89%
pratique	représentation	16	53%
enjeu*	défi*	16	73%
développement	application*	16	<b>94%</b>
étude*	modélisation	15	<b>94%</b>
place	rôle	15	83%
méthode	outil*	15	65%
approche*	modélisation	15	<b>94%</b>
représentation	pratique	14	47%
méthode	application*	14	<b>93%</b>
enjeu*	pratique	14	70%
théorie	application*	13	<b>100%</b>
principe	application*	13	<b>100%</b>
enjeu*	limite	13	81%
évolution	perspective*	10	<b>100%</b>
situation	perspective*	10	<b>100%</b>
expérimentation	modélisation	10	59%
continuité	rupture	10	29%
caractérisation	modélisation	10	<b>100%</b>
observation	modélisation	9	<b>100%</b>
modélisation	application*	9	<b>100%</b>
forme	enjeu*	9	<b>100%</b>
fait	chiffre	9	<b>100%</b>
enjeu*	politique	9	<b>90%</b>
discours	pratique	9	82%
cause	conséquence*	9	<b>90%</b>
aspect	application*	9	<b>100%</b>

Tableau 22 : Fréquences des couples de noms non ordonnés du patron SNC

On remarque que certains ordres sont figés dans notre corpus : on dira toujours « A CC B » et donc la plupart du temps « A et B ». Dans ce syntagme, l'ordre de A et de B semble obéir à une règle qui dépend des lemmes A et B et non pas du fonctionnement syntaxico-sémantique de la conjonction de coordination, pour laquelle l'ordre de A et B est indifférent. Ainsi, on ne trouve jamais dans le corpus « réalité CC mythe » mais toujours « mythe CC réalité ». Pourtant, « mythe » existe aussi en seconde position, mais il est associé avec « histoire » (deux fois), « péril » et « Ouroboros » en première position.

Il en va de même pour de nombreux couples : nous avons mis en gras ceux se trouvant entre 90 et 100% dans l'ordre défini par les colonnes du tableau. Les plus remarquables couples ne se trouvent que dans cet ordre dans notre corpus : « bilan et perspective », « intérêt et limite », « apport et limite », « théorie et application », « principe et application », « évolution et perspective », « situation et perspective », « caractérisation et modélisation », « observation et modélisation », « modélisation et application », « forme et enjeu » et « fait et chiffre ».

Il faut toujours ramener ce pourcentage à la fréquence totale : plus cette dernière est grande, plus nous pouvons avoir confiance dans notre pourcentage, c'est-à-dire qu'il ne s'agit pas d'un accident de constitution de corpus mais d'une réelle tendance. Nous pouvons avancer trois types d'explications qui ne sont pas mutuellement exclusives :

- La première est sémantico-logique : pour faciliter la compréhension, il est logique, dans un texte, d'exposer en premier la situation présente puis d'aborder le futur. C'est le cas pour « bilan et perspective » et « situation et perspective ». De même, « caractérisation et modélisation » et « observation et modélisation » relève de la logique d'étapes successives dans une démarche inductive.
- La seconde relève de la rhétorique : c'est une stratégie connue de parler des avantages d'un objet avant d'en aborder les limites. Le lecteur est d'abord « chargé » d'émotions positives devant ce qu'apporte l'objet, avant que le rédacteur n'aborde ses limites, espérant les faire mieux accepter grâce aux émotions déclenchées, c'est le *pathos* défini par Aristote qui l'oppose à la raison, *logos* et à l'aura de l'orateur, l'*ethos*. C'est le cas des syntagmes « intérêt et limite » et « apport et limite ».
- La troisième relève de la tradition dans la démarche scientifique, qui veut que l'on présente d'abord la théorie avant d'aborder la pratique. C'est le cas pour « théorie et application », « principe et application » et « modélisation et application ».

## V.2 Résultats du patron SN

### V.2.1 Fréquences des prépositions

Les prépositions présentes dans plus de 1% des titres sont les suivantes :

Lemme	Occurrences	Fréquences
de	29615	70%
à	4322	10%
sur	1881	4%
pour	1806	4%
en	1677	4%
dans	1012	2%
entre	611	1%
par	583	1%

Tableau 23 : Fréquence des prépositions du patron SN

On remarque le poids ultra-majoritaire de « de » qui est la préposition employée par 70% des 42 606 titres dont la séquence d'étiquettes POS a été capturée par notre patron SN. Contrairement aux conjonctions de coordination « et » et « ou », les prépositions ont une direction sémantique : les syntagmes « l'enfant de la maison » et « la maison de l'enfant » ne sont pas équivalents sémantiquement, au contraire de « la maison ou la cabane » et « la cabane ou la maison ». Il faut en tenir compte quand nous allons analyser plus avant les résultats de ce patron. Nous passons à présent aux noms en première position.

### V.2.2 Fréquences des noms en première position

Il y a 5 120 lemmes différents en position 1. Les lemmes présents dans plus de 1% des titres sont :

n°	Lemme	Occurrences	Fréquences
1	cas*	2895	7%
2	étude*	1983	5%
3	application*	1923	5%
4	exemple*	1657	4%
5	analyse	790	2%
6	approche*	686	2%
7	apport*	648	2%
8	état*	567	1%
9	Rôle	554	1%
10	effet	442	1%
11	enjeu*	437	1%

Tableau 24 : fréquences des noms en première position du patron SN

Nous regardons à présent comment ce premier nom se combine avec les différentes prépositions.

### V.2.3 Fréquences des noms en première position avec la préposition

Il y a 7 992 combinaisons (nom 1, préposition). Celles présentes dans plus de 1% des titres sont présentées dans le tableau suivant. La première fréquence, F1, est par rapport au nombre total de titres couverts par le patron. La seconde fréquence, F2, est par rapport au nombre d'occurrences du nom seul en première position. Elle permet de mesurer le taux d'utilisation de ce nom avec cette préposition.

n°	Lemme nom	Lemme préposition	Occurrences	F1	F2
1	cas*	de	2885	6,8%	99,7%
2	application*	à	1692	4,0%	88,0%
3	exemple*	de	1590	3,7%	96,0%
4	étude*	de	1438	3,4%	72,5%
5	analyse	de	659	1,5%	83,4%
6	apport*	de	583	1,4%	90,0%
7	état*	de	565	1,3%	99,6%
8	rôle	de	468	1,1%	84,5%

Tableau 25 : fréquences noms en première position avec la préposition du patron SN

Des 11 noms les plus fréquents en première position, seuls « approche », « effet » et « enjeu » n'apparaissent pas dans ce tableau. On voit qu'il y a une très faible dispersion entre les différentes prépositions. La préposition préférée est toujours « de », sauf pour « application » qui préfère « à ».

Notons bien que si « de » et « sur » peuvent être sémantiquement équivalents dans « une étude de X » et « une étude sur X » ce n'est généralement pas le cas et cela dépend aussi bien du nom qui suit que de la préposition. Nous avons étudié pour chaque premier nom les prépositions qui le suivent. Nous avons sélectionné pour chacun des 11 noms les plus fréquents leurs 4 prépositions associées les plus fréquentes :

Lemme nom	Prép. 1	Fréq.	Prép. 2	Fréq.	Prép. 3	Fréq.	Prép. 4	Fréq.
cas*	de	99,7%	dans	0,2%	à	0,1%	en	< 0,1%

étude*	de	73,7%	sur	8,6%	en	3,8%	à	3,2%
application*	à	88,0%	de	4,5%	en	3,8%	dans	1,7%
exemple*	de	96,1%	en	1,2%	dans	1,1%	à	0,5%
analyse	de	85,4%	par	5,9%	en	2,9%	à	2,2%
approche*	de	47,4%	par	30,8%	en	6,3%	pour	5,4%
apport*	de	90,1%	à	4,2%	pour	4,2%	dans	0,6%
état*	de	99,6%	à	0,2%	de	0,2%		
rôle	de	87,7%	dans	6,7%	pour	4,5%	en	0,5%
effet	de	75,1%	sur	22,4%	en	0,9%	à	0,7%
enjeu*	de	66,4%	pour	25,4%	à	3,2%	en	2,5%

Tableau 26 : fréquences des noms en première position et des différentes prépositions après pour le patron SN

Il se dégage, pour 9 d'entre eux, qu'une préposition est privilégiée à plus de 70% et que les autres ne dépassent pas 10%. Pour « approche » et « enjeu », la répartition est plus équilibrée.

On peut remarquer, bien que ce problème dépasse le cadre de ce travail, que l'on peut parfois substituer une préposition à une autre sans altérer le sens. Lorsqu'elle est possible, la substitution de préposition pose la question du choix de celle-ci par le locuteur. Nous pouvons remarquer que certaines prépositions ont un sens physique primaire, par exemple « dans » spécifie qu'on se trouve à l'intérieur d'un espace, « sur » qu'on se trouve dessus un objet. Mais l'usage affaiblit le contenu sémantique de la préposition en la rendant susceptible d'accepter n'importe quel lemme après, quel que soit son contenu sémantique. La préposition « dans » peut ainsi accepter, outre un espace géographique, un espace social comme une institution, ou intellectuel comme un champ disciplinaire.

L'étude du corpus, en catégorisant ontologiquement et/ou fonctionnellement (Huyghe, 2015) le nom qui suit la préposition, pourrait nous permettre de répondre à cette question en construisant un tableau de préférences, comme par exemple : la préposition « dans », dans la construction « application dans », est suivie pour X% d'un lieu géographique, pour Y% d'un domaine scientifique, pour Z% d'un organe biologique, etc. », permettant ainsi de comparer la versatilité de son emploi.

## V.2.4 Fréquences du nom en deuxième position

Il y a 11 543 lemmes différents en position 2. Aucun ne sont présents dans plus de 1% des titres : il y a une plus grande variété en deuxième position qu'en première. Le tableau suivant montre les lemmes présents dans plus de 0,5% des titres :

Lemme	Occurrences	Fréquence
cas*	410	0,96%
lieu*	365	0,86%
étude*	362	0,85%
recherche	319	0,75%
pratique	270	0,63%
analyse	263	0,62%
littérature	247	0,58%

Tableau 27 : fréquences du nom en deuxième position pour le patron SN



Encore une fois, ce sont des lemmes issus du champ lexical de la recherche lexical qui sont les plus fréquents. Nous pouvons à présent passer à l'étude des fréquences des syntagmes complets.

## V.2.5 Fréquences des syntagmes complets

Nous étudions à présent la fréquence des syntagmes complets, en considérant le nom en première position, celui en seconde position et la préposition entre. Sur les 33 127 combinaisons possibles, nous plaçons notre seuil de sélection à 39 occurrences, soit plus de 0,09% des titres de notre corpus de travail, sachant qu'aucune combinaison n'est présente dans plus de 1% des titres :

n°	Nom 1	Prép.	Nom 2	Occurrences	Fréquences
1	<b>état*</b>	<b>de</b>	<b>lieu*</b>	336	0,79%
2	étude*	de	cas*	307	0,72%
3	<b>revue*</b>	<b>de</b>	<b>littérature</b>	171	0,40%
4	point*	de	vue	158	0,37%
5	<b>état*</b>	<b>de</b>	<b>art</b>	86	0,20%
6	retour	de	expérience	80	0,19%
7	mise	en	place	67	0,16%
8	mise	en	évidence	60	0,14%
9	contribution	à	étude*	59	0,14%
10	mise	en	perspective*	45	0,11%
11	élément*	de	réflexion*	42	0,10%
12	cas*	de	étude*	42	0,10%
13	<b>état*</b>	<b>de</b>	<b>connaissance</b>	40	0,09%
14	application*	à	étude*	40	0,09%
15	<b>état*</b>	<b>de</b>	<b>question*</b>	40	0,09%
16	cas*	de	entreprise	39	0,09%
17	acte	de	colloque	39	0,09%

Tableau 28 : Fréquences des triplets (nom 1, préposition, nom 2)

On remarque sur les 17 combinaisons sélectionnées, « état » est présent en première position 4 fois, « mise », 3 fois et « cas », 2 fois.

À partir de ce tableau, nous étudions un point précis, celui de l'équivalence sémantique de plusieurs des expressions répertoriées.

## V.2.6 Exprimer la notion d'« état des lieux »

Le tableau des expressions les plus fréquentes comptent cinq expressions sémantiquement équivalentes que nous avons mises en gras et qui comptent pour 1,55% des triplets. Une recherche supplémentaire dans notre corpus ramène d'autres expressions équivalentes :

Nom 1	Prép.	Nom 2	Occurrences	Fréquences
état*	de	lieu*	336	0,79%
revue*	de	littérature	171	0,40%
état*	de	art	86	0,20%
état*	de	connaissance	40	0,09%
état*	de	question*	40	0,09%
état*	de	recherche	21	0,05%
état*	de	savoir	4	0,01%

revue*	de	connaissance	3	0,01%
état*	de	débat	2	< 0,01%
état*	de	champ	1	< 0,01%
état*	de	documentation	1	< 0,01%
état*	de	littérature	1	< 0,01%
état*	de	réflexion	1	< 0,01%
revue*	de	question*	1	< 0,01%

Tableau 29 : fréquences des différentes expressions pour la notion d'état des lieux dans le patron SN

Les expressions avec un grand nombre d'occurrence doivent être regardées en premier lieu :

### Exemples :

- *Le lexique de la météorologie en Corse : état des lieux et perspectives* (Stella Retali Medori, 2018, Linguistique)
- *L'expérience des apprenants en e-formation : revue de littérature* (Gilles Dieumegard, Marc Durand, 2018, Éducation – Psychologie - Anthropologie sociale et ethnologie)
- *Systèmes d'alerte anti-collision : état de l'art et impact du niveau de fiabilité et du moment de déclenchement* (Alexandra Fort, Mercedes Bueno, Christophe Jallais, 2018, Psychologie et comportements - Santé publique et épidémiologie)
- *Avifaune du Niger: état des connaissances en 1986* (Patrick Giraudoux, René Degauquier, P.J. Jones, Jean Weigel, Paul Isenmann, 2018, Biodiversité et Ecologie - Ecologie, Environnement)
- *Les premiers monastères d'Auvergne à la lumière de la documentation textuelle et archéologique (V e -X e siècle) : état de la question* (Damien Martinez, 2018, Archéologie et Préhistoire)

À chaque fois, le sujet est donné par le premier segment du titre, celui avant le double point, et le second délimite la portée de l'article en fixant son but : établira un bilan ces connaissances. Plus un triplet est fréquent plus on peut estimer que son figement est important.

Les expressions avec un nombre d'occurrence de 1 doivent être regardées avec suspicion car on ne peut savoir s'il s'agit d'une possibilité de variation d'une expression plus ou moins figée, qui serait par la même moins figée puisqu'elle accepterait des variations, ou de l'originalité d'un auteur qui déformerait une expression plus figée.

Pour déterminer cela, on regarde le titre dans son entièreté pour savoir si le nom en deuxième position est bien un synonyme ayant pour référent les travaux scientifiques déjà faits sur ce sujet :

1. *Le management des connaissances : un état du champ* (Pascal Lièvre, 2017, Gestion et management - Economies et finances)
2. *Le travail de Florus de Lyon sur la prédestination : un état de la documentation conservée* (Pierre Chambert-Protat, 2017, Etudes classiques – Histoire – Religions)
3. *Conséquences d'accidents majeurs de barrages : état des réflexions de l'INERIS pour l'évaluation de la gravité* (Thibault Balouin, Anabel Lahoz, 2014, Sciences de l'ingénieur)
4. *Système familial et attachement : revue de la question* (S. Pinel-Jacquemin, Chantal Zaouche-Gaudron, 2017, Psychologie)

Il est clair que pour les exemples (2) et (3) ce n'est pas le cas. Un second discriminant est de savoir le nombre d'auteurs de ces occurrences uniques, pour savoir s'il y a derrière un seul auteur, ce qui accrédi terait la thèse d'une originalité, d'un écart, ou un groupe d'auteurs, tous ayant alors validé le titre qui ne résulterait plus d'une originalité individuelle. L'exemple (1) est le résultat d'un seul auteur, nous préférons l'écarter comme les deux autres, alors que nous prenons le (4), car il a deux auteurs.

Si on reprend toutes les expressions que nous estimons sémantiquement équivalentes, celles qui ne sont pas en italique dans le tableau, nous avons 705 titres (0,82% du corpus de travail) qui les utilisent. On peut se demander si le choix d'utilisation de ces syntagmes sémantiquement équivalents est lié à une discipline donnée. Pour cela on dresse le tableau suivant, en utilisant le code des disciplines (se référer à Tableau 6 : Répartition des titres par domaines) :

Nom 1	Prép	Nom 2	shs	sdv	sdu	info	ssco	phys	spi	sde	math	chim	stat	qfin	nlin
état*	de	lieu*	<b>176</b>	<b>138</b>	3	5	4	0	3	<b>14</b>	0	0	1	2	0
revue*	de	littérature	34	<b>127</b>	0	3	5	0	5	4	0	0	0	1	1
état*	de	art	25	12	1	<b>34</b>	2	3	<b>18</b>	6	0	1	1	0	0
état*	de	connaissance	13	27	1	0	0	0	0	3	0	0	0	0	0
état*	de	question*	40	0	0	0	1	0	0	0	0	0	0	0	0
état*	de	recherche	20	0	0	0	0	0	1	0	0	0	0	0	0
état*	de	savoir	4	0	0	0	0	0	0	0	0	0	0	0	0
revue*	de	connaissance	0	<b>3</b>	0	0	0	0	0	0	0	0	0	0	0
état*	de	débat	2	0	0	0	0	0	0	0	0	0	0	0	0
état*	de	littérature	1	0	0	0	0	0	0	0	0	0	0	0	0
revue*	de	question*	1	0	0	0	0	0	0	0	0	0	0	0	0
<b>Total par discipline</b>			<b>316</b>	<b>307</b>	<b>5</b>	<b>42</b>	<b>12</b>	<b>3</b>	<b>27</b>	<b>27</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>1</b>

Tableau 30 : fréquence des expressions pour exprimer la notion d'état des lieux du patron SN par disciplines

Nous écartons les disciplines pour lesquels nous n'avons pas assez d'occurrences pour statuer, avec un total inférieur à 27, en italique dans le tableau. Sur les autres, nous constatons certaines préférences pour une expression donnée. Ainsi, les Sciences de l'Homme et Société (shs) ont une préférence à 56% pour l'expression utilisant les lemmes (état, de, lieux). Les Sciences du Vivant (sdv) ont une préférence pour deux combinaisons équitablement réparties : (état, de, lieux) à 45% et (revue, de, littérature) à 41%. L'Informatique (info) a une nette préférence pour (état, de, art) à 81% mais sur un nombre d'occurrences beaucoup plus faible (42). Avec 27 occurrences chacune, les Sciences de l'ingénieur (spi) et les Sciences de l'environnement (sde) montrent une préférence pour (état, de, art) à 67% et (état, de, lieu) à 52% respectivement. Le choix de ces expressions figées est donc une question d'habitude de la discipline.

On peut également étudier la variation du premier et du second nom.

Nom 1	shs	sdv	info	spi	sde
<b>état*</b>	<b>89%</b>	<b>58%</b>	<b>93%</b>	<b>81%</b>	<b>85%</b>
<b>revue*</b>	<b>11%</b>	<b>42%</b>	<b>7%</b>	<b>19%</b>	<b>15%</b>

Tableau 31 : Répartition des lemmes en position 1 par disciplines

Pour le premier nom, seules les Sciences du Vivant (sdv) hésitent réellement entre « état » et « revue ». Les autres affichent une nette préférence pour « état » ou « revue » et « état » est toujours privilégié.

Nom 2	shs	sdv	info	spi	sde
lieu*	56%	45%	12%	11%	52%
littérature	11%	41%	7%	19%	15%
Art	8%	4%	81%	67%	22%
connaissance	4%	10%	0%	0%	11%
question*	13%	0%	0%	0%	0%
recherche	6%	0%	0%	4%	0%
savoir	1%	0%	0%	0%	0%
débat	1%	0%	0%	0%	0%

Tableau 32 : Répartition des lemmes en position 2 par disciplines

Cette façon de voir n'est valable que si on assume que toutes les combinaisons sont valables, or certaines ne se trouvent pas dans notre corpus comme (revue, de, savoir). Il vaut mieux calculer la probabilité du nom 2 selon le nom 1 par discipline. Il nous a semblé intéressant de la faire pour les Sciences du Vivant (sdv) pour savoir si l'habitude disciplinaire porte bien sur l'expression dans son entièreté ou sur un choix spécifique pour le nom 1 et un autre choix spécifique pour le nom 2, les deux étant indépendants :

Nom 2	Avec état	Avec revue	Total	P (Nom 2   état)	P (Nom 2   revue)
lieu*	138	0	138	78%	0%
art	12	0	12	7%	0%
connaissance	27	3	30	15%	2%
question*	0	0	0	0%	0%
recherche	0	0	0	0%	0%
savoir	0	0	0	0%	0%
débat	0	0	0	0%	0%
littérature	0	127	127	0%	98%
Total	177	130		100%	100%

Tableau 33 : Probabilité du nom 2 sachant le nom 1 pour les Sciences du Vivant (sdv)

On constate qu'il n'y a pas une indépendance des deux si le nom 1 est « revue » : à 98% on aura littérature après, ce qui témoigne du figement du syntagme. Si le nom 1 est « état », il y a une légère variabilité entre lieu, connaissance et art, ce qui témoigne d'un figement moindre.

Notre limite de deux noms après le double point ne nous empêche pas de constater l'existence de syntagmes récurrents plus long lors de notre exploration du corpus de travail. Ainsi on peut considérer le syntagme à 3 noms suivant :

Nom 1	P	Nom 2	CC	Nom 3	Occurrences
état*	de	art	et	perspective*	11
état*	de	connaissance	et	perspective*	2
état*	de	lieux*	et	perspective*	41
état*	de	question*	et	perspective*	3
état*	de	recherche	et	perspective*	1

Tableau 34 : Syntagmes contenant notamment (état, de, NC, et, perspective)

Deux constructions à 3 noms utilisent pour les noms 2 et 3 des lemmes issus de la liste des noms 2 les plus fréquents que nous avons établie pour notre patron SN : « état des lieux des connaissances » (3 occurrences) ou « état des lieux des savoirs » (1 occurrence). Ces deux constructions explicitent ce qui est les autres de l'ordre de l'implicite. Il n'y a pas de lieu comme référent dans l'expression « état des lieux ». Cette absence de référent est une des caractéristiques du figement des expressions (Legallois & Tutin, 2013).

Notons que l'on peut également trouver une des expressions que nous avons remarquées avant le double point comme dans : « *La prison en Suisse, un **état des lieux** : un point de vue français* » (Annie Kensey, Jean-Lucien Sanchez 2018, Histoire – Héritage culturel et muséologie) et « ***État des lieux** de la Filière Thrombectomie en Aquitaine : analyse des délais pré, inter et intra hospitaliers* » (Ludovic Lucas, 2017, Médecine humaine et pathologie). Ce qui se trouve avant le double point sortant du cadre de cet exercice, nous n'avons pas quantifié le phénomène. Notons néanmoins que le premier exemple est l'œuvre de deux auteurs, donc il ne peut s'agir d'un phénomène dû à l'originalité d'un individu unique.

### V.3 Résultats du patron SP

Ce patron est celui qui a la plus faible couverture : seulement 4 023 titres de notre corpus.

#### V.3.1 Fréquences de la première préposition

Les prépositions en première position ayant 25 occurrences ou plus sont :

Préposition	Occurrences	Fréquence
de	2324	57,8%
vers	676	16,8%
à	472	11,7%
pour	219	5,4%
entre	193	4,8%
sur	64	1,6%
en	25	0,6%
dans	25	0,6%

Tableau 35 : fréquence de la première préposition du patron SP

On remarque encore une fois la prévalence de « de » qui a elle seule compte pour presque de la moitié des prépositions.

#### V.3.2 Fréquences du premier nom

Les noms en première position ayant 25 occurrences ou plus sont :

Nom 1	Occurrences	Fréquence
propos	255	6%
recherche	100	2%
approche*	53	1%
modèle	36	0,9%
analyse	36	0,9%
origine	34	0,8%
théorie	32	0,8%
compréhension	32	0,8%

prise	28	0,7%
construction	27	0,7%
mise	26	0,6%
histoire	25	0,6%

Tableau 36 : fréquence du nom en première position du patron SP

On retrouve les mêmes lemmes issus du champ lexical de la recherche scientifique que pour nos patrons SN et SNC.

### V.3.3 Fréquences de la seconde préposition

Les prépositions en seconde position ayant 25 occurrences ou plus sont :

Préposition	Occurrences	Fréquence
de	2207	55%
à	1551	39%
en	93	2%
dans	35	1%
pour	32	1%
entre	28	1%

Tableau 37 : fréquence de la deuxième préposition du patron SP

Même si « de » reste la première préposition, on voit que « à » a une importance non négligeable. On voit se dessiner une structure utilisant « de ... à ... ».

### V.3.4 Fréquences du second nom

Il y a 1 897 noms possibles en seconde position. Ceux ayant 25 occurrences ou plus sont :

Nom 1	Occurrences	Fréquence
pratique	59	1,5%
cas*	46	1,1%
analyse	29	0,7%
modèle	27	0,7%

Tableau 38 : noms les plus fréquents en deuxième position dans le patron SP

Avec le patron SNC, on atteignait pour le deuxième nom une fréquence de 3% mais ce patron possède la propriété d'équivalence sémantique entre les deux arguments de la conjonction de coordination. Avec le patron SN, on ne dépassait pas 0,96% avec le lemme « cas » en seconde position. Ici, on est dans un intermédiaire ou émerge « pratique » et encore une fois « cas ».

### V.3.5 Fréquences des couples (préposition 1, préposition 2)

Ces couples vont nous permettre de visualiser l'utilisation conjointe des prépositions. Nous retenons les couples ayant plus de 25 occurrences :

Préposition 1	Préposition 2	Occurrences	Fréquence
de	à	1495	37%
de	de	703	17%
vers	de	565	14%
à	de	467	12%
pour	de	185	5%

entre	de	162	4%
sur	de	58	1%
de	en	34	1%
vers	en	33	1%
vers	à	26	1%
dans	de	25	1%

Tableau 39 : fréquence des couples de prépositions dans le patron SP

On remarque la présence du couple (de, de) en deuxième position.

### V.3.6 Fréquences des triplets (préposition 1, nom 1, préposition 2)

C'est cette analyse qui est la plus intéressante car en incluant le deuxième nom, rien ne se détache véritablement comme récurrences. Dans les triplets, au contraire, on constate plusieurs récurrences :

Préposition 1	Nom 1	Préposition 2	Occurrences	Fréquence / triplets
à	propos	de	255	6,3%
à	recherche	de	76	1,9%
à	origine	de	23	0,6%
vers	modèle	de	23	0,6%
de	concept	à	22	0,5%
vers	approche*	de	21	0,5%
de	théorie	à	21	0,5%
pour	approche*	de	19	0,5%
de	analyse	de	18	0,4%
vers	compréhension	de	18	0,4%
vers	prise	en	18	0,4%
pour	histoire	de	16	0,4%
de	usage	de	15	0,4%
vers	émergence	de	15	0,3%
à	source	de	13	0,3%
de	mythe	à	13	0,3%
de	recherche	à	13	0,3%
pour	lecture	de	13	0,3%
sur	trace	de	13	0,3%
vers	construction	de	13	0,3%

Tableau 40 : fréquence des triplets (préposition 1, nom 1, préposition 2) dans le patron SP

La locution prépositive « à propos de » est la plus fréquente mais aussi la plus intéressante car elle semble indiquer une délimitation précise du sujet du document titré.

#### Exemples :

1. Reprises de prothèses articulaires septiques compliquées et prothèses tumorales avec traitement de surface à l'argent : à propos de 3 cas (Eric Denes, F. Fiorenza, V. Vacquerie, G. Cordier, B. Abraham, G. Gosheger, Pierre Weinbreck , 2018, Médecine humaine et pathologie)

2. *De la salle de cinéma à la caverne : à propos de quelques tentatives artistiques d'ensevelissement* (Thibault Honoré, 2018, Art et histoire de l'art)
3. *Analyse du circuit du médicament : à propos des effets indésirables inévitables et évitables* (Catherine Piquet Diakhate, 2018, Sciences pharmaceutiques)

On voit qu'il est difficile d'identifier sémantiquement une catégorie commune aux noms qui viennent après la locution prépositive. L'exemple (2) se détache des deux autres par sa difficulté de compréhension. Le contenu amené par la locution prépositive est tout aussi sibyllin. Les deux autres exemples indiquent bien une délimitation du sujet, dans un cas à « 3 cas », dans l'autre « aux effets indésirables » mais sans que nous détectons une récurrence sur **le plan sémantique plus fine**.

## V.4 Analyse globale des 3 patrons

### V.4.1 Le champ lexical de la recherche scientifique

La première constatation globale est que l'on retrouve, pour les deux noms dans les patrons SNC et SP et pour le premier nom dans le patron SN, des noms issus du champ lexical de la recherche scientifique. Par rapport à notre classe de noms ayant une affinité pour se trouver après le double point :

- On retrouve dans les syntagmes récurrents étudiés les noms : « cas », « exemple », « application », « résultat », « perspective », « enjeu », « réflexion », « revue », « approche », « apport », « point », « élément », « état », « question », « lieu », « étude », « expérience »
- On ne retrouve pas dans les syntagmes récurrents étudiés les noms : « proposition », « enquête », « comparaison », « conséquence », « défi », « réalité », « regard », « outil », « concept ». Cela laisse à penser que si les fréquences de ces noms considérés isolément étaient élevées, celles des syntagmes dans lesquels ils s'inscrivent sont trop basses pour nos filtres.

Lorsque l'on étend la recherche de récurrence à tout un syntagme, on voit qu'il y a une concurrence entre plusieurs syntagmes. Notre étude à portée sur celui signifiant le concept l'état de l'art car c'était le plus fréquent. Le choix d'un syntagme pour l'autre est dicté par la discipline, même si une variabilité est parfois possible, à deux niveaux : dans le choix du syntagme en entier ou dans le choix du second nom. Cette dernière variabilité est néanmoins beaucoup plus restreinte, le choix de « revue » en premier nom entraînant presque automatiquement celui de « littérature » en deuxième.

### V.4.2 L'approche phraséologie

Pour expliquer ces figements nous pouvons nous tourner vers la phraséologie, « *l'étude des séquences lexicales perçues comme préconstruites* » selon (Legallois & Tutin, 2013) sur lesquelles nous nous appuyons dans cette partie.

Il s'agit de considérer que « *le figement est affaire de continuum* », des expressions figées, appelées aussi unités phraséologiques, indécomposable, jusqu'au associations passagères et libres. Entre, un ensemble de collocations, présences communes et répétées dans l'environnement lexical, et de colligations, constructions communes et répétées dans l'environnement grammatical, plus ou moins récurrentes.



La phraséologie est une ancienne branche de la lexicologie qui a connu récemment un élargissement de son domaine d'étude, au point d'imprégner de nombreux champs de la linguistique : syntaxe, linguistique textuelle, sémantique lexicale et psycholinguistique.

Cette extension va jusqu'à supposer l'existence d'un « *principe phraséologique de la langue [...], selon lequel les locuteurs sélectionneraient des pans de la langue préconstruits, intégrant à la fois lexicale et grammaire* » selon (Sinclair, 1991) traduit par (Legallois & Tutin, 2013).

Cette vision se rapproche l'école contextualiste, dont la grammaire de patrons est l'exemple que nous avons le plus étudié (Hunston & Francis, 2000) et de l'école constructionniste, comparée à la première dans (Legallois, 2006). Le lemme ne doit plus être considéré seulement comme une unité indépendante et libre dans son emploi mais également, voir primairement, comme s'inscrivant dans un ensemble d'utilisations lexico-syntaxique, que cela soit un réseau de constructions ou une accumulation de patrons (Legallois, 2006), qui s'imposent lors de la production linguistique.

(Hoey, 2005) va plus loin en proposant la théorie du « *lexical priming* », traduite par amorçage lexicale par (Legallois & Tutin, 2013) : « *un lexème est acquis grâce à ses occurrences dans les discours et textes, il se charge cumulativement des contextes et des co-textes dans lesquels il est "rencontré"* » (Legallois, 2006). Si on excepte l'apprentissage de fiches de vocabulaire dénuées d'exemples, un lemme est effectivement appris par le locuteur toujours en contexte, donc dans une expression avec un degré de figement. Lorsque le locuteur devient producteur, la volonté d'employer ce lemme déclenche le rappel des contextes et co-textes où il a été rencontré, que le locuteur peut alors reprendre dans sa production.

Il y a donc un processus circulaire qui pousse à la répétition du contexte dans lequel le lemme existe, allant jusqu'à former des unités phraséologiques tout à fait figées. La linguistique ~~cognitiviste~~ et la psycholinguistique se sont penchées sur ce phénomène. (Wray, 2002) rejoint le principe phraséologique du langage lorsqu'elle affirme que la grammaire, comme compétence psycholinguistique, « *a pour origine un répertoire d'unités phraséologiques contextuellement situées* », selon la traduction de (Legallois & Tutin, 2013), et c'est dans la « *déformabilité des unités mémorisées pour générer des énoncés nouveaux* » que réside la flexibilité de la compétence. Sans cela le locuteur serait condamné à pouvoir uniquement répéter des unités déjà entendues. Ces unités ont une fonction principale, la promotion de l'intérêt personnel du locuteur. Il dispose de « *l'avantage cognitif que confèrent les unités phraséologiques mémorisées et donc rapidement disponibles et facilement énonçables* » et, communes aux locuteurs d'une même langue, également facilement recevables.

Cette vision phraséologique ajoute une dimension explicative pour comprendre pourquoi nous avons identifié des expressions dont l'ordre est fixé, comme « bilan et perspective » et « mythe et/ou réalité », ou dont la structure syntagmatique et lexicale est fixée comme « revue de littérature » dont les occurrences de variations sont infimes.



Nous avons dans cette partie étudié les principaux résultats de l'utilisation de nos patrons sur notre corpus de travail. Si les patrons se limitaient au niveau syntaxique, nous avons ici pleinement abordé le niveau lexical, constatant le figement ou le quasi-figement de certaines expressions fréquentes. Une explication globale est fournie par la phraséologie, qui veut que l'utilisation d'un

lemme ne soit pas complètement libre : celle-ci s'inscrit dans un ensemble de collocations et colligations que le locuteur apprend avec le lemme et qu'il restitue, avec un certain degré de liberté, lorsqu'il emploie le lemme.

Nous pouvons à présent prendre du recul sur notre travail et ses résultats pour essayer d'identifier des points d'améliorations et les perspectives possibles à sa suite.

## VI. Discussion et perspectives

### VI.1 Recherche d'information par l'analyse sémantique des titres : le cas de « application »

Si on prend le nom « application » et ses 1 923 occurrences en première position, on constate pour l'utilisation des différentes prépositions :

Nom	Préposition	Occurrences	Fréquences
application*	à	1693	88,0%
application*	de	87	4,5%
application*	en	74	3,8%
application*	dans	33	1,7%
application*	sur	15	0,8%
application*	pour	14	0,7%
application*	chez	3	0,2%
application*	par	1	0,1%
application*	pendant	1	0,1%
application*	vers	1	0,1%
application*	avec	1	0,1%

Tableau 41 : fréquences d'utilisation de différentes prépositions avec "application"

Pour mieux comprendre l'utilisation des différentes prépositions, nous tirons quelques exemples de notre corpus :

1. *Approche préventive pour une réduction des Hydrocarbures Aromatiques Polycycliques (HAP) dans les fours à pyrolyse : **application à la cémentation gazeuse à basse pression*** (Tsilla Bensabath, 2018, Génie des procédés - Génie chimique)
2. *Rapports méninges des nerfs spinaux dans leur trajet intra et extra foraminaux : **application en pratique chirurgicale*** (Thomas Wavasseur, 2017, Médecine humaine et pathologie)
3. *Mise en œuvre d'un outil SIG et d'un processus d'analyse multicritère semi-automatisé pour l'aménagement du territoire : **application dans le cadre de la révision du SCoT des Vosges Centrales*** (Matthieu Chevallier, 2017, Sciences de l'ingénieur)
4. *Comparaison de différentes méthodes d'interprétation de la prédiction de l'eau corporelle par la méthode de dilution de l'eau lourde : **application chez le chevreau mâle*** (P. Schmidely, J. Robelin, P. Bas, 1989, Biologie de la reproduction - Alimentation et Nutrition - Biologie du développement)

Dans l'exemple (1), « application à la cémentation », la cémentation est un traitement. « application en cémentation », « application dans la cémentation », « application sur la cémentation » paraît équivalent mais pas « chez », qui est réservé pour les personnes et les animaux

comme dans l'exemple (4). Certaines prépositions sont donc substituables tout en préservant le sens de l'expression.

Dans l'exemple (2), « application en pratique chirurgicale », plusieurs prépositions sont substituables : « application dans la pratique chirurgicale », « application à la pratique chirurgicale » ou « application sur la pratique chirurgicale », même si cette dernière nous semble un peu moins évidente. Les exemples (1), (2) et (4) ont en commun une structure sémantique articulée autour du double point, qui correspond à celle décrite par (Haggan, 2004) : le titre rétrécit progressivement le sujet étudié. Dans la première partie avant le double point, il annonce un sujet que l'expression « application P » va préciser en restreignant son champ d'application à un exemple donné. Ce sera un traitement (la cémentation), un domaine (la pratique chirurgicale), ou un animal d'un sexe précis et à un stade de développement précis (le chevreau mâle). On peut étendre cette notion de champ d'application, voire d'exemple d'application, à celui de délimitation du sujet.

Pour l'exemple (3), « dans le cadre de » est une expression très figée. Nulle autre préposition ne serait possible ici mais cela poursuit le même but : préciser le champ d'application de ce qui vient avant le double point. L'information additionnelle après le double point sert donc à délimiter le sujet.

On peut élaborer un nouveau type de patron, cette fois-ci lexico-syntaxico-sémantique. Lexical car il se base sur la présence du lemme APPLICATION (nom), et soit d'un lemme prépositionnel pris dans A, DANS, SUR, soit d'une expression figée « dans le cadre de », associés dans une configuration syntaxique, le syntagme, qui comprend également un nom avec plusieurs traits sémantiques possibles : /traitement/, /domaine scientifique/ et /animal/. Ce patron s'inscrit lui-même dans un patron de titres, qui peut s'écrire :

Objet scientifique : délimitation
-----------------------------------

Mais l'objet scientifique est lui-même décomposable dans nos exemples. Le premier nom donne son sujet, avec des syntagmes nominaux comme « approche préventive », « rapportes méninges des nerfs spinaux », « mise en œuvre d'un outil SIG et d'un processus [...] » ou « comparaison de différentes méthodes d'interprétation ». Un syntagme prépositionnel contribue ensuite à le délimiter avec « dans les fours à pyrolyse », « dans leur trajet intra et extra foraminal », à préciser son but « pour l'aménagement du territoire » ou en explicitant le moyen avec « par la méthode de dilution de l'eau lourde ».

Considérons à présent deux nouveaux exemples avec la préposition « de » :

5. *Echanges thermiques chez le porcelet nouveau-né* : **application de la méthode du bilan d'énergie** (P. Berbigier, J. Le Dividich, A. Kobilinsky, 1978, Zootechnie)
6. *Analyse des variations individuelles en nutrition animale* : **application de l'analyse en composantes principales** à l'étude de la sécrétion lipidique du lait de chèvre (D. Sauvart, P. Morand-Fehr, 1978, Zootechnie)

Ces exemples sont bâtis autrement. Dans les deux cas, l'utilisation de la préposition « de » indique ce qui est appliqué, ici une méthode et une analyse particulières, qui correspondent au moyen utilisé. Les exemples diffèrent après. Pour (5), le premier segment fournit le champ d'application. Pour (6), le premier segment donne à la fois le sujet scientifique et une première délimitation du sujet sur son champ d'application : « en nutrition animale ». Dans le second segment,

le syntagme prépositionnel commençant par « à » donne une seconde délimitation. Finalement, on peut donc relire nos exemples ainsi :

n°	Sujet	Délimitation pré « : »	But	Moyen	Délimitation post « : »
1	Approche préventive	dans les fours à pyrolyse	pour une réduction des Hydrocarbures Aromatiques Polycycliques (HAP)		la cémentation gazeuse à basse pression
2	Rapports méninges des nerfs spinaux	dans leur trajet intra et extra foraminaux			la pratique chirurgicale
3	Mise en œuvre d'un outil SIG et d'un processus d'analyse multicritère semi-automatisé		pour l'aménagement du territoire		la révision du SCoT des Vosges Centrales
4	Comparaison de différentes méthodes d'interprétation de la prédiction de l'eau corporelle			par la méthode de dilution de l'eau lourde	le chevreau mâle
5	Echanges thermiques	chez le porcelet nouveau-né		la méthode du bilan d'énergie	
6	Analyse des variations individuelles	en nutrition animale		l'analyse en composantes principales	la sécrétion lipidique du lait de chèvre

Tableau 42 Décomposition sémantique des titres

Il semble que le rétrécissement étudié par (Haggan, 2004), opéré par le segment après le double point sur le segment avant le double point, soit réalisé également par des syntagmes prépositionnels pouvant apparaître aussi bien avant qu'après le double point. On peut se demander si la délimitation entre le sujet et les autres composantes est forte : dans l'exemple (3) la « mise en œuvre d'un outil SIG et d'un processus » peut être aussi bien considéré comme le moyen, laissant la colonne sujet vide pour cet exemple. Le sujet serait alors ce qu'on choisit de mettre en avant, par une position particulière, la première, parmi toutes les composantes du titre.

## VI.2 Limitations de l'outillage et des patrons

Notre travail a été possible par l'utilisation d'outils puissants et efficaces comme Python, Excel, Talismane et nos patrons. Néanmoins, nous avons remarqué quelques défauts et nous les énumérons dans cette partie, tout en citant des voies d'amélioration.

### VI.2.1 Erreurs dans la lemmatisation et l'étiquetage POS

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 segments sont autant de phrases incomplètes. TreeTagger (Schmid, 1994) 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. (Haggan, 2004) remarque que les titres se rapproche plus d'un des types de C-units définies par (Leech, 2000). Ces petites unités grammaticales indépendantes sont très présentes à l'oral et peu à l'écrit, sauf dans les titres.

Nous ne pouvons que constater les défauts de Talismane à étiqueter en catégories du discours correctement certaines formes. Par exemple voici une partie postérieure à un double point d'un titre : « Approche sémiotique et poétique ». Talismane l'étiquette « NC NC CC NC » alors que la séquence correcte devrait être « NC ADJ CC ADJ ». Plus généralement, Talismane a une forte proportion a considéré un adjectif comme un nom commun, rendant plus difficile les calculs statistiques. Les syntagmes « étude observationnelle » et « revue critique » sont ainsi vues comme « NC NC » et donc non comptés comme utilisant « étude » et « revue » respectivement pour son nom en première position. Nous avons effectué des corrections supplémentaires pour prendre en compte ce problème.

De plus, Talismane n'arrive pas, pour de nombreuses formes, à retrouver leurs lemmes, donnant '\_' à la place. Il ne tolère aucune erreur : l'absence d'une lettre à « vengeance » n'est pas corrigée en vengeance, de même que l'absence d'accent, « chainon » n'est pas corrigé « chaînon ». On remarque également que la présence de majuscules le perturbe fortement : « Exemple » et « EXAMPLE » n'étaient pas associés au lemme « exemple », de même « Art » pour le lemme « art ».

Talismane se comporte différemment devant un prénom et un nom propre : le prénom aura pour lemme sa forme, par exemple « Irène », alors que le nom propre aura pour lemme « \_ » comme pour « Némirovsky ». Talismane a aussi une tendance à catégoriser des formes en noms propres comme « s » (773 occurrences), « p » (369), J (301), I (275) ou « La » (470), « Il » (241) et « Le » (267). Un nettoyage manuel est ensuite nécessaire.

De plus, il manque à Talismane un vocabulaire spécialisé propre à chaque science. À ces formes inconnues, il n'associe aucun lemme, alors que la reprise de la forme, avec éventuellement la suppression de morphèmes grammaticaux classiques, comme le « -s » du pluriel, donnerait de meilleurs résultats.

Enfin, nous avons constaté des problèmes d'encodages de caractères non uniformes, non repérable visuellement : le lemme « étude » figurait ainsi trois fois dans la table des fréquences des lemmes, comme s'il s'agissait de trois lemmes différents. Heureusement, un lemme était ultra-majoritaire, et la correction a été faite manuellement en agrégeant à celui-ci les occurrences des deux formes minoritaires qui, à l'œil nu, ne s'en différenciaient aucunement.

Face à ces trois problèmes, nous pensons qu'un étiqueteur-lemmatiseur spécialement créé pour les titres permettrait d'obtenir de meilleurs résultats. Nous ne prétendons pas qu'il serait meilleur que Talismane en général, fruit d'un long travail, mais nous pensons qu'il pourrait arriver à de meilleurs résultats spécifiquement sur les titres. S'il est possible de réinjecter une liste de triplets (forme, lemme, catégorie) issue de cet outil projeté dans Talismane, nous pourrions alors bénéficier des traitements plus avancés de ce dernier, comme l'analyse des dépendances, sans devoir tout refaire, pour peu qu'elle ne soit pas mise en difficulté face aux constructions particulières des titres.

## VI.2.2 Développement des patrons

Nous avons vu que nos patrons capturent des séquences linéaires d'étiquettes POS et non des organisations structurales. Les résultats retournés sont donc ambigus. Pour le patron SN, nous nous sommes étendus sur le problème d'un syntagme prépositionnel inclus soit directement dans un syntagme nominal, soit dans un syntagme adjectival (« plein de N »). Une solution serait d'interdire un adjectif entre le nom et la préposition, mais nous perdriions de nombreux titres, ou d'interdire certains adjectifs appelant fréquemment une expansion prépositionnelle, donnant une dimension lexicale à nos patrons. Une approche plus ambitieuse serait de créer des patrons capturant des organisations structurales syntagmatiques. Cela nécessiterait au préalable la construction et le stockage de l'analyse syntaxique du titre.

Nous pouvons également envisager la construction de patrons avec plus de deux noms pour capturer les séquences comme « état des lieux et perspectives » ou « historique et état des lieux ». Cette limitation de deux noms peut nous faire perdre d'autres séquences récurrentes, qui potentiellement étendent certaines que nous avons déjà repérées.

Sur l'impossibilité de définir un patron infini dans notre langage, nous pensons que cette limitation est bénéfique : elle écarte le risque de capturer automatiquement certains titres marginaux qui ont une suite de 4 ou 5 noms communs consécutifs par exemple. On peut toujours les capturer volontairement avec « NC NC NC NC NC? » néanmoins. Les fonctions cognitives du lecteur ne peuvent gérer une suite trop longue d'un même élément, comme les expansions prépositionnelles à la suite par exemple.

Une amélioration conséquente à la recherche de structures et non plus de séquences linéaires et de s'affranchir de l'obligation d'avoir un ordre linéaire, une « *contrainte qui constitue une limite des approches classiques par segments répétées* » (Legallois & Tutin, 2013). En surmontant celle-ci, (Longrée & Mellet, 2013) présente la notion de motif. Prolongeant cette notion de motif, (Quiniou, 2012) propose une méthode pour les faire émerger automatiquement d'un corpus afin de comparer la stylistique de différents genres littéraires.

Une autre voie d'amélioration est l'adoption d'un langage standardisé pour exprimer nos patrons, comme le *Corpus Query Language* ou CQL<sup>21</sup>. Des interpréteurs existent, comme le *Corpus Query Processor* ou CQP. Néanmoins la maîtrise technique complète de nos patrons actuels nous permet d'effectuer des statistiques fines dessus, pour chaque élément du patron par exemple, ou combinaison d'éléments, et de représenter les résultats au format Excel avec la possibilité de mettre en valeur graphiquement certains éléments.

## VI.3 Zone non couverte et création de sous-corpus

### VI.3.1 Zone non couverte du corpus

Bien que nos 3 patrons SN, SNC et SP couvrent 61.51% des possibles séquences d'étiquettes après un double point et 64.85% des titres de notre corpus de travail, il reste plus d'un tiers, à chaque fois, qui ne soit pas couvert. Nous avons stipulé qu'aucun patron ne nous apparaissait intuitivement à la visualisation des séquences. Un script peut néanmoins être écrit pour faire

---

<sup>21</sup> [http://txm.ish-lyon.cnrs.fr/bfm/files/QuickRef\\_CQL\\_BFM.pdf](http://txm.ish-lyon.cnrs.fr/bfm/files/QuickRef_CQL_BFM.pdf) pour une présentation rapide de CQL par l'équipe de TXM.

émerger de nouveaux patrons ou motifs, de façon automatique, à la manière de ceux de (Quiniou, 2012), ou de façon semi-automatique.

Intuitivement, la catégorie absente de nos patrons est le verbe, soit sous la forme d'un infinitif, soit d'une forme conjuguée. Dans ce dernier cas, le premier élément après le double point serait un premier nom ou pronom, qui serait le sujet du verbe. Il serait intéressant de savoir quelles personnes sont utilisées pour ce pronom et surtout quelles personnes n'apparaissent jamais, si tel était le cas. On peut ensuite envisager un deuxième nom qui aurait une fonction de complément d'objet du verbe, direct ou indirect.

Augmenter le taux de couverture permet de détecter de nouveaux phénomènes et de lever le voile sur les dernières zones d'ombre de notre corpus de travail.

### VI.3.2 Créations de sous-corpus

Les articles que nous avons étudiés dans notre état de l'art portaient tous seulement sur la catégorie des articles scientifiques. Nous avons pris comme hypothèse que le titrage des autres types de documents, comme les ouvrages, les chapitres d'ouvrage, les vidéos, les mémoires et les rapports, ne se différenciait pas de celui des articles. Pour appuyer notre hypothèse, il s'agit dans les deux cas de titres élaborés par des membres de la communauté scientifique.

Néanmoins cette hypothèse devrait être vérifiée par la création d'un sous-corpus de notre corpus de travail constitué uniquement de titres d'articles. Sa réalisation technique ne pose pas de soucis. Notre corpus de travail étant constitué à 30% de ceux-ci, il serait d'une taille de 25 000 titres environ, suffisamment étendu pour détecter de nombreux phénomènes linguistiques et les étayer avec une fréquence assez grande pour distinguer les phénomènes dus au hasard de ceux motivés par des raisons concrètes. On pourrait alors comparer nos résultats avec ceux obtenus sur l'ensemble du corpus de travail.

On peut également subdiviser ce premier sous-corpus ou le corpus de travail en différents sous-corpus classés par disciplines. Un document pouvant être rattaché à plusieurs disciplines, il faut faire attention à ceux dont c'est le cas. Un sous-corpus des titres en biologie, doit-il par exemple contenir les titres rattachés à la biologie et l'informatique ? Le plus simple est de considérer les titres rattachés à une seule discipline en premier et de n'opérer que sur les disciplines de niveau 0, les plus larges. Nous avons déjà créé pour notre travail deux sous-corpus de notre corpus de travail de façon dichotomique : d'un côté les titres des documents rattachés à la discipline « Sciences de l'Homme et Société » (0.shs) et de l'autre ceux n'y étant pas rattachés. Le premier compte 61 252 titres et le second 24 279.

### VI.4 Le cas des noms propres

Bien que nos patrons capturent aussi bien les noms communs que les noms propres en première et seconde position, nous n'avons pas poussé l'étude de ces derniers. D'un part car ils étaient beaucoup moins fréquents que les noms communs (voir le Tableau 10 : Comptes des noms communs les plus fréquents avant et après le double point) sauf le premier d'entre eux, France, avec 2806 occurrences. Le second, Paris, ne compte lui que 827 occurrences. Voici le tableau des plus fréquents, nettoyés des incohérences ramenées par Talismane :

n°	Lemme	Occurrences	%	% du nombre total	avant le « : »	%	après le « : »
----	-------	-------------	---	-------------------	----------------	---	----------------



				de noms			
1	<b>France</b>	2806	0,03%	1330	0,47%	1476	0,53%
2	<b>Paris</b>	827	0,01%	314	0,38%	513	0,62%
3	<b>Europe</b>	672	0,01%	365	0,54%	307	0,46%
4	Saint	541	< 0,01%	207	0,38%	334	0,62%
5	Jean	476	< 0,01%	209	0,44%	267	0,56%
6	<b>Afrique</b>	457	< 0,01%	322	0,70%	135	0,30%
7	Moyen	376	< 0,01%	201	0,53%	175	0,47%
8	<b>Nord</b>	338	< 0,01%	154	0,46%	184	0,54%
9	<b>Réunion</b>	320	< 0,01%	169	0,53%	151	0,47%
10	Pierre	264	< 0,01%	117	0,44%	147	0,56%
11	<b>Italie</b>	264	< 0,01%	118	0,45%	146	0,55%
12	<b>Sud</b>	244	< 0,01%	110	0,45%	134	0,55%
13	<b>Bretagne</b>	232	< 0,01%	106	0,46%	126	0,54%
14	<b>Lyon</b>	226	< 0,01%	76	0,34%	150	0,66%
15	<b>Espagne</b>	205	< 0,01%	100	0,49%	105	0,51%
16	<b>Allemagne</b>	199	< 0,01%	93	0,47%	106	0,53%

Tableau 43 : fréquence des noms propres dans notre corpus de travail

On remarque que la France se détache fortement. HAL étant développé dans un milieu francophone, elle reçoit surtout les résultats de la recherche universitaire francophone qui intuitivement porterait plus sur la France, cela pouvant donc expliquer cette prééminence. Un triplet (NC P NC) avec 25 occurrences étaient « cas de la France » exprimant une délimitation du sujet à ce pays. Au-delà de la France, 12 des 16 plus fréquents lemmes, en gras, concernent un emplacement géographique. (Jacques & Sebire, 2010) indiquent que la présence d'un nom géographique dans le titre précise son sujet mais aussi le délimite et délimite également l'intérêt que les chercheurs peuvent lui porter, ce qui a pour conséquence qu'il est moins cité. À notre connaissance, HAL ne permet pas de récupérer le nombre de citations et de téléchargements automatiquement, nous n'avons pas pu confirmer cette information.

On remarque la présence de Jean (476) et Pierre (264) qui pourrait correspondre à des documents scientifiques en rapport avec l'hagiographie, le lemme « Saint » étant lui présent 541 fois, insuffisamment néanmoins pour couvrir toutes les occurrences des prénoms cités (740 en tout).

La répartition des lemmes avant et après le double point ne relève rien de caractéristique.

## VI.5 Autres structures

En reprenant notre liste des noms de notre classe n'étant pas apparus dans les syntagmes les plus fréquents et en étudiant les résultats du patron SN, nous entrevoyons d'autres syntagmes récurrents, mais avec un nombre d'occurrences et/ou une fréquence plus basse que ceux définis par nos filtres. Ainsi un le couple (outil, de) et le couple (outil, pour) ont respectivement 224 et 94 occurrences, un autre déjà suffisant pour témoigner d'une utilisation récurrente. Il laisse entrevoir une structure sémantique de la forme :

Désignation de l'outil : (outil, de / pour) but de l'outil



Cette construction ajoute une information importante à l'outil désigné dans la première partie : son but. On pourrait ainsi construire une liste d'outils ayant un but semblable.

**Exemple :** *Micro-impression de BMP-2 et fibronectine sur des matériaux mous : **un outil pour recréer la niche de cellules souches in vitro*** (Vincent Fitzpatrick, 2018, Biotechnologies)

On remarque dans l'exemple précédent le syntagme prépositionnel délimiteur « sur des matériaux mous » dans la partie de désignation de l'outil.

Autre exemple, les couples (regard, sur), 124 occurrences, et (regard, de), 58 occurrences. Le couple (regard, sur) semble donner la problématique, alors que le premier segment avant le double point donne le sujet de façon moins précise que dans d'autres titres, comme dans les exemples suivants :

#### Exemples :

- *De l'esquisse à l'œuvre enregistrée : **regard sur une poïétique du rock*** (Philippe Gonin, 2018, Musique, musicologie et arts de la scène)
- *Les nouvelles prisons françaises : **Regard sur l'acceptabilité sociétale des établissements pénitentiaires*** (Gerald Billard, 2018, Géographie)

Il serait alors possible d'envisager un catalogue de syntagmes récurrents associés à des sémantiques, de façon à faire « comprendre » automatiquement le titre au niveau sémantique à un algorithme, ou du moins à récupérer ses principaux composants : sujet, problématique, but, moyen, délimitations, même si ces notions doivent être auparavant clarifiées, définies et reliées.

Cette décomposition sémantique du titre ne pourra être complète qu'en résolvant l'interprétation référentielle des noms généraux de notre classe, dont « outil » et « regard » font parties. Pour « outil », intuitivement, la réponse semble se trouver dans le premier segment. Pour « regard », il n'y a pas d'intuition qui émerge à la lecture des exemples du corpus.



Cette partie nous a permis d'énumérer des perspectives et des améliorations possibles de notre travail. Que cela soit au niveau de la lemmatisation et de la catégorisation, des patrons, ou de certaines corrections, les outils peuvent toujours être améliorés. Ces améliorations doivent être mises en rapport avec leurs coûts qui peuvent être très élevés, d'où l'acceptation de certains défauts et limitations. Nos corpus, celui général et celui de travail, étiquetés et lemmatisés, constituent déjà des ressources utilisables par d'autres, et la création de sous-corpus spécialisé pour de nouvelles problématique est facilement faisable. Le cas des noms propres n'a pas été traité en profondeur et mériterait également qu'on s'y attarde.

La compréhension sémantique des titres ouvre la perspective de l'élaboration d'outils de recherche sémantique dans les archives ouvertes. Ils amélioreraient grandement leurs capacités actuelles face au foisonnement de documents scientifiques produits. À partir de titres comme « *FLEMM: un analyseur flexionnel du français à base de règles* » (Namer, 2018, Linguistique<sup>22</sup>) une base de connaissance pourrait être construite. Lorsqu'un utilisateur chercherait « FLEMM », le

---

<sup>22</sup> Cet article ne se trouve pas dans HAL.

logiciel de recherche pourrait lui proposer d'autres analyseurs flexionnels du français en plus des articles portant directement sur ce logiciel.

## Conclusion

---

Si on met de côté la fonction d'attraction, le titre d'un document scientifique doit relever le défi d'informer au maximum le lecteur du contenu du document dans un espace très contraint, plus petit qu'un paragraphe : 15,5 lemmes en moyenne. Les auteurs, en plus du sujet, vont parfois jusqu'à mettre leurs conclusions dans le titre, rendant la contrainte de place encore plus forte.

Dans ce contexte, l'utilisation du double point présente l'avantage de d'être économe en nombre de mots outils et d'être facile à interpréter par le lecteur. D'où une utilisation assez forte, dans 30% des 278 806 titres de documents scientifiques que nous avons récupérés de l'archive ouverte HAL. Les 85 531 titres de notre corpus de travail, comportant un et un seul double point, privilégient une organisation en deux segments séparés par le double point. Si le premier segment présente généralement le sujet général de l'article, le second segment présente une information supplémentaire sur celui-ci, avec une classe particulière de lemmes, des noms généraux, ayant une forte affinité pour se situer après le double point.

Dans le but d'analyser la mise en œuvre de cette information et sa nature sémantique, nous avons établi trois patrons syntaxiques linéaires portant sur les séquences d'étiquettes POS juste après le double point. Nous avons ensuite étudié les syntagmes auxquels elles correspondaient. Pour délimiter notre travail, nous nous sommes limités arbitrairement aux syntagmes binominaux, avec deux noms communs.

Nos trois patrons sont : « NC P NC », « P NC P NC » et « NC CC NC » où NC représente un nom commun, P représente une préposition et CC une conjonction de coordination. Le premier patron couvre près de 50% des titres de notre corpus, le second 5% et le dernier 10%. Soit un total de couverture de 65% du corpus de travail, assez grand pour nous permettre d'observer des phénomènes assez fréquents pour ne pas être de simples accidents.

Nous avons ensuite étudié les lemmes qui peuplent les séquences les plus fréquentes capturées par nos patrons. Dans les trois, nous avons constaté l'utilisation récurrente de noms issus du vocabulaire général du domaine scientifique comme « étude », « cas », « approche », « analyse », « application », « pratique », « exemple », « enjeu », « perspective », « modélisation », « limite ». Certains de ces noms avaient déjà été repérés comme faisant partie de notre classe ayant une affinité très forte pour se situer après le double point, comme « enjeu », « résultat », « approche », « élément », « expérience ». On peut remarquer que l'ensemble de ces noms sont tous des noms généraux, un type fonctionnel de nom dont la particularité est d'avoir un faible contenu sémantique mais une très large application référentielle, comme le montre leur transdisciplinarité. Notre classe peut gagner à être redéfinie non pas sur le simple critère de la répartition des lemmes avant ou après le double point, mais par leurs présences dans des syntagmes récurrents après le double point.

Nous avons remarqué dans nos résultats certaines expressions plus ou moins figées. Comme le figement absolu de l'ordre des 4 couples formés par les lemmes « mythe » et « réalité », « bilan » et « perspective », « intérêt » et « limite », « théorie » et « application » au sein d'un syntagme du type « NC CC NC ». Nous avons remarqué l'existence de syntagmes « NC P NC » récurrents et sémantiquement équivalents comme « état des lieux / de l'art / des connaissances / de la question /

de la recherche » et « revue de littérature / des connaissances ». Il en va de même pour le patron « P NC P NC » avec des syntagmes comme « à propos de N ».

Ces figements laissent supposer que l'écriture des titres fait appel à des lemmes en réseau, et l'utilisation d'un provoque chez le locuteur un rappel des cooccurrences et colligations dans lesquels il l'a précédemment vu. Le locuteur peut alors réutiliser celles-ci, tout en ayant la capacité de les faire varier pour produire de nouvelles combinaisons.

Nous pensons que ce travail est une première exploration de notre corpus, apportant une première étude sur un phénomène particulier, la récurrence de syntagmes binominaux apportant une information additionnelle, qui peut être une délimitation, à la partie du titre avant le double point. Des délimitations peuvent également se trouver avant le double point mais il se pose alors la question de savoir où finit le sujet et où commence ses délimitations, mais cela était hors de nos limites de travail. De plus, la décomposition sémantique du titre et la résolution de l'interprétation référentielle des noms généraux ouvrent des perspectives très intéressantes.

Dans notre travail, nous avons également écarté le reste du titre après et le segment avant le double point. Nos premières constatations nous laissent à penser que des phénomènes de récurrence peuvent également survenir avant le double point, ainsi des délimitations. Autre point, nos limitations influent directement sur notre conception des patrons et les séquences capturées. Ainsi, la séquence « NC P NC CC NC » est capturée par le patron « NC P NC » mais comporte pourtant « NC CC NC » plus loin. Savoir si la coordination est entre le premier et le dernier noms ou le second et le dernier noms requiert une analyse syntaxique plus complète. À la suite de celle-ci, une redéfinition de nos patrons est nécessaire pour éventuellement capturer des structures de syntagmes et non des séquences linéaires, et éventuellement capturer des syntagmes plus complexes comme « état des lieux et perspectives » qui indique à la fois un état du présent et propose des éléments futurs. Nous avons également écarté l'étude des noms propres pour nous concentrer sur les noms communs, mais une étude plus approfondie de ces derniers est également envisageable.

Toutes ces limitations peuvent être dépassées, et l'outillage peut être amélioré. La création de sous-corpus pour étudier des points particuliers peut également être envisagé. Les perspectives sont nombreuses et vastes : une caractérisation plus poussée des disciplines par rapport à certaines caractéristiques des titres, la recherche d'information dans les titres pour concevoir une recherche sémantique dans les archives ouvertes, un analyse détaillée des noms généraux dans les titres, ou encore des travaux à portée didactique à l'intention des étudiants et jeunes chercheurs pour l'écriture de titres. Nous espérons que nos prochains travaux nous permettrons d'explorer certains de ces points.

## Bibliographie

---

- Adler, S. (2018). Sémantique des noms généraux sous-spécifiés et construction du sens. *Langages*, 71-86.
- 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.
- Ayres, I. (2008). *Super crunchers: How anything can be predicted*. Hachette UK.
- 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.
- Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Hamilton, D. P. (1991). Research Papers: Who's Uncited Now? *Science*, 25.
- Hartley, J. (2003). Single authors are not alone: Colleagues often help. *Journal of Scholarly Communication*, 34 (2), 108-113.
- Hinkel, E. (2004). *Teaching Academic English as a Second Language Writing: Practical techniques in vocabulary and grammar*. New Jersey: Lawrence Erlbaum Associates.
- Hoey, M. (2005). *Lexical Priming: A new theory of language*. New York / Abingdon: Routledge.
- Hornby, A. S. (1954). *A Guide to Patterns and Usage in English*. Oxford: Oxford University Press.
- Hunston, S., & Francis, G. (2000). *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing.

- Huyghe, R. (2015). Les typologies nominales : présentation. *Langue française*, 1(185), 5-27.
- Jacques, T. S., & Sebire, N. J. (2010). The impact of article titles on citation hits: an analysis of general and specialist medical journals. *Journal of the Royal Society of Medicine Short Reports*, 1-5.
- 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.
- Legallois, D. (2006). Il-3 Pattern Grammar. Dans D. Legallois, & J. François, *Autour des grammaires de constructions et de patterns*. *Cahier du CRISCO*, 21. (pp. 33-42). Caen: Centre de Recherches Interlangues sur la Signification en Contexte.
- Legallois, D., & Tutin, A. (2013). Présentation : Vers une extension du domaine de la phraséologie. *Langages* (1), 3-25.
- 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.
- Longrée, D., & Mellet, S. (2013). Longrée, D., & Mellet, S. (2013). Le motif: une unité phraséologique englobante? Étendre le champ de la phraséologie de la langue au discours. *Langages*, (1), 65-79.
- 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. (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.
- 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.
- Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells*. Berlin: Mouton de Gruyter.
- Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- 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*.
- Swales, J. M., & Feak, C. B. (1994). *Academic Writing for Graduate Students*. Ann Arbor: University of Michigan Press.
- Townsend, M. A. (1983). Titular Colonicity and Scholarship: New Zealand Research and Scholarly Impact. *New Zealand Journal of Psychology*, 41-43.
- Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Toulouse: Doctoral dissertation, Université de Toulouse II-Le Mirail.
- 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.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

## 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 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 7 : 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) illustrent cette puissante fonctionnalité. La première demande la création d'un cache, la seconde poursuit la lecture du cache créé avec l'identifiant fourni par la réponse à la première requête, `AoFVmLIG`.

```
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 8a 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 nos données brutes pour faire 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



exemptions car nous n’opérons pas de choix dans les titres que nous retournent HAL. Si nous devons 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 9 : 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 9 : 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 10 : une ligne de notre 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 10 : une ligne de notre 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 376 Mo pour les 278 806 titres de notre corpus général. Compressé au format ZIP, il ne fait plus que 44 Mo. Notre corpus de travail dans ce même format fait 132 Mo pour 85 531 titres, soit 16 Mo compressé.

## 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<sup>23</sup>.

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
DETH	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

Tableau 44 : codes des étiquettes POS de Talismane

## A3. Index des tableaux

Tableau 1 : nombre de doubles points dans les titres .....	19
Tableau 2 : Répartition des titres par type.....	20
Tableau 3 : Répartition des titres par année.....	21
Tableau 4 : Nombres de titre par nombres d'auteurs.....	22
Tableau 5 : Nombre de titres par nombres de doubles points et longueurs moyennes .....	24
Tableau 6 : Répartition des titres par domaines .....	24
Tableau 7 : Analyse des domaines de l'exemple .....	24
Tableau 8 : Titres avec un caractère segmentant dans notre corpus .....	25
Tableau 9 : Phrase complète dans les titres en fonction du domaine de la biologie.....	26

<sup>23</sup> Plus précisément ce tableau se trouve ici : <http://joliciel-informatique.github.io/talismane/#section2.3.4>

Tableau 10 : Comptes des noms communs les plus fréquents avant et après le double point.....	26
Tableau 11 : Tableau des noms communs les plus fréquents avec pourcentage d'occurrences après le double point .....	27
Tableau 12: exemples de suites de catégories correspondant à un syntagme nominal après le double point .....	32
Tableau 13 : Séquences générées par notre patron .....	33
Tableau 14: Les séquences les plus fréquentes dans les titres .....	35
Tableau 15 : couverture des trois patrons .....	41
Tableau 16 : Tableau des fréquences de la coordination du patron SNC .....	42
Tableau 17: Tableau des fréquences des lemmes en première position.....	42
Tableau 18 : Tableau des fréquences des lemmes en seconde position du patron SNC.....	42
Tableau 19 : répartition des lemmes les plus fréquents avant et après le double point du patron SNC .....	42
Tableau 20 : fréquence des triplets avec un choix de conjonction de coordination du patron SNC...	43
Tableau 21 : fréquences des triplets les plus fréquents du patron SNC .....	44
Tableau 22 : Fréquences des couples de noms non ordonnés du patron SNC.....	45
Tableau 23 : Fréquence des prépositions du patron SN .....	46
Tableau 24 : fréquences des noms en première position du patron SN.....	47
Tableau 25 : fréquences noms en première position avec la préposition du patron SN.....	47
Tableau 26 : fréquences des noms en première position et des différentes prépositions après pour le patron SN.....	48
Tableau 27 : fréquences du nom en deuxième position pour le patron SN .....	48
Tableau 28 : Fréquences des triplets (nom 1, préposition, nom 2) .....	49
Tableau 29 : fréquences des différentes expressions pour la notion d'état des lieux dans le patron SN .....	50
Tableau 30 : fréquence des expressions pour exprimer la notion d'état des lieux du patron SN par disciplines .....	51
Tableau 31 : Répartition des lemmes en position 1 par disciplines.....	51
Tableau 32 : Répartition des lemmes en position 2 par disciplines.....	52
Tableau 33 : Probabilité du nom 2 sachant le nom 1 pour les Sciences du Vivant (sdv) .....	52
Tableau 34 : Syntagmes contenant notamment (état, de, NC, et, perspective).....	52
Tableau 35 : fréquence de la première préposition du patron SP .....	53
Tableau 36 : fréquence du nom en première position du patron SP.....	54
Tableau 37 : fréquence de la deuxième préposition du patron SP .....	54
Tableau 38 : noms les plus fréquents en deuxième position dans le patron SP .....	54
Tableau 39 : fréquence des couples de prépositions dans le patron SP .....	55
Tableau 40 : fréquence des triplets (préposition 1, nom 1, préposition 2) dans le patron SP .....	55
Tableau 41 : fréquences d'utilisation de différentes prépositions avec "application" .....	58
Tableau 42 Décomposition sémantique des titres.....	60
Tableau 43 : fréquence des noms propres dans notre corpus de travail.....	64
Tableau 44 : codes des étiquettes POS de Talismane .....	74

## A4. Index des graphiques

Figure 1 : Pourcentage d'occurrences après le double point et nombre total d'occurrences des noms communs les plus fréquents .....	28
Figure 2 : arbre d'analyse syntagmatique .....	30
Figure 3 : Syntagmes possibles pour "DET NC ADJ P NC" .....	31
Figure 4 : Structure minimale du syntagme idéal pour le patron SN .....	38
Figure 5 : structure du syntagme idéal pour le patron SP .....	39
Figure 6 : structure du syntagme idéal pour le patron "NC CC NC" .....	40
Requête 7 : Un exemple de requête avec l'API Apache Solr formaté pour plus de lisibilité .....	72
Requêtes 8a et 2b : création et consultation d'un cache .....	72
Résultat 9 : exemple d'un élément de résultats au format JSON .....	73
Résultat 10 : une ligne de notre corpus de travail .....	73

## A5. Index des logiciels, technologies et notions mentionnés

Apache Lucene .....	15	JSON .....	15
Apache Solr .....	15	langdetect .....	18
CoNLL-U .....	16	Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) .....	15
Corpus Query Language .....	62	preprint .....	14
Corpus Query Processor .....	62	protocole de transfert hypertexte .....	HTTP
CQL .....	Voir : Corpus Query Language	Python .....	7
CQP .....	Voir : Corpus Query Processor	Stanford Core Natural Language Processing .....	15
CSV .....	15	Talismane .....	15
Excel .....	7	XML .....	15
HTTP .....	15		

## A6. Définition du schéma utilisé pour les corpus

Notre corpus utilise le format XML mais son schéma n'est pas standard. Nous fournissons ici la définition de son schéma au format XSD pour permettre à l'aide d'outils comme XLST de le transformer selon n'importe quel schéma.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="notices">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="notice" minOccurs="1" maxOccurs="unbounded">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="id" type="xs:long" />
              <xs:element name="type" type="xs:string" />
              <xs:element name="date" type="xs:short" />
              <xs:element name="text" type="xs:string" />
              <xs:element name="words">
                <xs:complexType>
                  <xs:sequence>
```

```
<xs:element name="word"  
    minOccurs="1" maxOccurs="unbounded">  
    <xs:complexType>  
        <xs:sequence>  
            <xs:element name="form" type="xs:string" />  
            <xs:element name="lemma" type="xs:string" />  
            <xs:element name="pos" type="xs:string" />  
        </xs:sequence>  
    </xs:complexType>  
</xs:element>  
</xs:sequence>  
</xs:complexType>  
</xs:element>  
<xs:element name="authors">  
    <xs:complexType>  
        <xs:sequence>  
            <xs:element name="author" type="xs:string"  
                minOccurs="1" maxOccurs="unbounded"/>  
        </xs:sequence>  
    </xs:complexType>  
</xs:element>  
<xs:element name="domains">  
    <xs:complexType>  
        <xs:sequence>  
            <xs:element name="domain" type="xs:string"  
                minOccurs="1" maxOccurs="unbounded" />  
        </xs:sequence>  
    </xs:complexType>  
</xs:element>  
</xs:sequence>  
</xs:complexType>  
</xs:element>  
</xs:sequence>  
</xs:complexType>  
</xs:element>  
</xs:schema>
```