Damien Gouteux

Structures

LEXICO-SÉmantiqueS

Dans les titres scientifiques

Projet de recherche SLT0701V

Encadré par Mme Josette Rebeyrolle et M. Ludovic Tanguy

22 janvier 2018

Numéro étudiant 21704896

Master 1 LITL 2017 – 2018



# Document du 1er semestre pour le projet de recherche SLT0701V

Table des matières

[Document du 1er semestre pour le projet de recherche SLT0701V 1](#_Toc504428849)

[Introduction 1](#_Toc504428850)

[1. État de l’art 3](#_Toc504428851)

[1.1 Sujet des articles étudiées 3](#_Toc504428852)

[1.2 Caractéristiques des corpus des articles 3](#_Toc504428853)

[1.3 Informations extrinsèques supplémentaires aux titres 5](#_Toc504428854)

[1.4 Informations intrinsèques aux titres 5](#_Toc504428855)

[1.5 Propositions de typologies des titres 6](#_Toc504428856)

[3. Données de travail 7](#_Toc504428857)

[3.1 Présentation de HAL 7](#_Toc504428858)

[3.2 Requêtes et résultats 8](#_Toc504428859)

[3.2.1 Requêtes 8](#_Toc504428860)

[3.2.2 Résultats 9](#_Toc504428861)

[4. Premières analyses et premiers résultats 10](#_Toc504428862)

[4.1 Les titres comme matériel pour le TAL 11](#_Toc504428863)

[4.2 Types des textes scientifiques dans notre corpus et HAL 11](#_Toc504428864)

[4.3 Longueurs des titres et présences des caractères segmentant 12](#_Toc504428865)

[4.4 Le lemme d’après 13](#_Toc504428866)

[4.3 Quelques exemples de titres 14](#_Toc504428867)

[5. Pistes et perspectives 14](#_Toc504428868)

[5.1 Continuer à lire d’autres travaux 14](#_Toc504428869)

[5.2 Construire nos propres corpus 14](#_Toc504428870)

[5.3 Réécriture des outils 15](#_Toc504428871)

[5.4 Explorer notre patron en profondeur 15](#_Toc504428872)

[Conclusion 15](#_Toc504428873)

[Table des requêtes et résultats 17](#_Toc504428874)

[Table des tableaux 17](#_Toc504428875)

[Bibliographie 18](#_Toc504428876)

# Introduction

*« Titles consist of only a few words, but they are serious stuff. »*Swales (1990: 224)

Un titre est la porte d’entrée d’un texte scientifique, que cela soit un cours, un article de recherche, un ouvrage ou une thèse. Il s’agit généralement du premier contact qu’a le lecteur avec le texte titré et parfois du seul, s’il décide ensuite de ne pas le lire. Comme les titres de la presse généraliste, les titres oscillent entre deux fonctions pragmatiques, informer et attirer. La première se base sur le dévoilement de l’information contenu dans l’article, son champ de recherche, son sujet et parfois même ses conclusions, pour aider rapidement le lecteur à prendre sa décision. La seconde fonction vise à séduire le lecteur en l’amusant ou l’intriguant et l’amener ainsi à continuer sa lecture.

Les mécanismes de la seconde fonction, qui peuvent aller jusqu’à masquer l’information ou l’éclairer sous un jour très particulier, sont hors de portée de notre travail. De plus, notre intuition nous porte à penser que la première fonction est bien plus présente dans les titres de textes scientifiques, la seconde n’étant présente que de façon anecdotique et même considérée comme contraire aux règles de bonnes écritures d’un titre par certains chercheurs (Aleixandre-Benavent et al., 2014).

Mais la première fonction retient notre attention. Comment cette volonté d’information se traduit-elle dans la construction d’un titre ? Ceux-ci ont souvent des spécificités syntaxiques. Souvent, un titre ne comporte pas de verbe conjugué. Il se présente comme une phrase nominale, ou plusieurs, parfois étendues par des syntagmes prépositionnels. En cela, ils ne sont pas à première vue des candidats faciles pour le traitement automatique des langues (TAL). Mais ils sont aussi souvent exempts de fautes de graphie ou de frappe, d’écritures « phonétique » présentent dans d’autres sources et qui viennent rendre plus difficile les tâches du TAL. La volonté d’informer se traduit sur le plan sémantique par la volonté de définir le thème du texte, de façon bien plus condensée que dans un paragraphe ou un texte plus long.

Pour étudier cette volonté d’informer, nous voudrions dans le cadre de ce travail explorer un type de structure, un patron, très représenté dans les titres scientifiques. Celui centré sur un double point « **:** » suivi immédiatement après d’un syntagme nominal (**SN**). Le lemme qui constitue son noyau doit être abordé avec une attention toute particulière. De ce fait, notre étude mêle plusieurs niveaux d’analyse. Le niveau lexical, le niveau syntaxique, et même, si nous en avons la possibilité, nous voudrions explorer également celui de la sémantique. Nous nous inscrivons dans une approche résolument *corpus driven* (Solen et al., 2012) : nous abordons nos données sans a priori théorique, pour essayer de faire émerger de notre corpus des constructions linguistiques. Une fois que nous aurons identifié les structures qui suivent ce patron, les variations possibles, nous voudrions essayer de corréler leur utilisation avec des caractéristiques propres aux publications titrées comme leur discipline ou le type de texte par exemple.

Au fil de notre travail lors du second semestre il est possible que d’autres types de structures retiennent notre attention et puissent étendre notre champ d’étude. Dans ce document, nous nous concentrerons sur le premier que nous évoquions, le patron **« … : SN … »**. Nous commençons par dresser un état de l’art d’études portant également sur les titres. Si l’approche *corpus driven* n’a pas d’a priori théoriques, il faut néanmoins déjà avoir une idée de l’objet que nous étudions, les titres scientifiques, et des précédentes études qui en ont été faites. Nous aborderons ensuite nos données de travail constituées en corpus. Enfin, nous livrerons nos premières analyses à partir de celles-ci avant de finir par évoquer les pistes et perspectives qui s’offrent à nous pour la poursuite de notre travail au second semestre.

# 1. État de l’art

Dans cette partie nous analysons six articles qui traitent également des titres de textes scientifiques. Nous allons les présenter selon cinq points de vue : le sujet de l’article, le corpus qu’il utilise pour son étude, les caractéristiques extrinsèques et intrinsèques des titres relevées par les auteurs et enfin différentes typologies proposées pour classer les titres.

## 1.1 Sujet des articles étudiées

Nous avons étudié six articles traitant des titres scientifiques, datés de 2004 à 2014. Chacun d’eux traite d’une problématique particulière que nous détaillons dans le tableau suivant :

|  |  |
| --- | --- |
| **Articles** | **Sujet d’étude** |
| Aleixandre-Benavent et al. (2014) | Les inexactitudes dans les titres d’articles biomédicaux : utilisations des points d’exclamation et d’interrogation, des métaphores, des doubles sens, des acronymes et des expressions vagues et imprécises. |
| Haggan (2004) | Analyse des titres dans trois disciplines : science, littérature et linguistique par rapport à leur structure syntaxique résumée selon la typologie suivante :   * Titre étant une phrase complète, * Titre composé de plusieurs syntagmes nominaux (souvent avec un double point les séparant en deux parties) * Autres structures |
| Jamali et Nikzad (2011) | Constitution d’une typologie des titres et mise en corrélation de celle-ci avec le nombre de téléchargements et de citations des articles via une étude statistique, en tenant compte de la longueur du titre et de la présence d’un double point ou d’un point d’interrogation. |
| Lewison et Hartley (2005) | Analyse du nombre de mots et de la présence d’un double point dans les titres d’articles scientifiques par rapport à leur discipline et dans une perspective diachronique sur 5 années prises entre 1981 à 2001. |
| Rebeyrolle, Jacques et Péry-Woodley (2009) | Fonctionnement des titres dans la construction du modèle interprétatif du discours. Cette article traite également des intertitres, mais nous écartons cette partie de l’étude dans ce document. |
| Whissell (2004) | Mesure de l’évolution en diachronie de l’agréabilité, de l’imaginaire, de l’activation, du style – longueur et structure – et de l’utilisation des mots dans les titres d’articles. Les trois premières valeurs sont calculées à l’aide d’un dictionnaire, le *Dictionary of Affect in Language* (DAL) (Whissel, 1989), qui donne la valence affective des mots présents dans le titre. |

Tableau 1 : Sujet des articles étudiés

## 1.2 Caractéristiques des corpus des articles

Tous ces articles s’appuient sur des corpus : une sélection de titres écrits dans des conditions de productions naturelles, dans un format manipulable par des ordinateurs, représentative de ce que l’on veut étudier (Gilquin et Gries, 2009). L’aspect naturel que nous prêtons à ces productions ne doit pas tromper : un titre est très travaillé, s’écrit généralement à plusieurs scripteurs. Ce que nous voulons dire c’est qu’ils ne sont pas écrits dans le but d’être ensuite étudiés et les scripteurs ne sont pas influencés par la présence d’un chercheur-enquêteur linguiste. Les corpus étudiés ont en commun d’être constitué exclusivement de titres en anglais et de titres d’articles parus dans des revues scientifiques, sauf pour l’article de Rebeyrolle et Péry-Woodley (2009). Nous présentons les données utilisées par les chercheurs dans les différents articles dans le tableau 3 en précisant la taille du corpus ainsi que les années et les disciplines retenues.

|  |  |
| --- | --- |
| **Articles** | **Données** |
| Aleixandre-Benavent et al. (2014) | **1 505** titres recupérés auprès *PubMed* et *Web of Sciences* entre 2009 et 2011. *PubMed* est un moteur de recherche gratuit sur la base de données bibliographiques *MEDLINE*, regroupant des notices d’articles de biologie et de médecine de plus de 5000 revues biomédicales. Ils sont gérés par la Bibliothèque américaine de médecine (NLM). *Web of Sciences* est un moteur de recherche accessible par abonnement d’une institution de recherche à laquelle le chercheur doit appartenir. Il indexe 7 bases de données bibliographiques de différentes disciplines, regroupant des articles de nombreuses revues et conférences. |
| Haggan (2004) | **751** titres pris au hasard dans des journaux scientifiques publiés aux États-Unis ou au Royaume-Uni dans les trois disciplines suivantes :   * Linguistique : 207 titres, 44 journaux, longueur moyenne : 8.8 mots * Littérature : 237 titres, 40 journaux, longueur moyenne : 9.4 mots * Science : 307 titres, 40 journaux, longueur moyenne : 13.8 mots   Haggan considère qu’il y a une opposition entre littérature et science qui pourrait se traduire en termes linguistique, d’où l’intérêt d’étudier et de comparer les deux. Elle déclare que la linguistique est un moyen terme entre les deux. Les années retenues diffèrent pour chaque journal mais elles sont prises parmi 1997, 1998, 1999 et 2000. |
| Jamali et Nikzad (2011) | **2 172** titres d’articles publiés dans 6 journaux scientifiques de la *Public Library of Science (PLOS)* en 2007. Les journaux recouvrent les disciplines de biologie, médecine et génétique, avec un journal spécialisé dans les pathogènes et un autre dans la bio-informatique. Un dernier regroupe toutes les sciences, *PLOS ONE*. |
| Lewison et Hartley (2005) | **349 700** titres répartis ainsi :   * 133 200 titres internationaux sur l’oncologie, * 216 500 titres du Royaume-Uni publiés dans des journaux scientifiques en biologie, biomédecine, chimie, médecine clinique, terre et espace, ingénierie et technologie, mathématiques, physique et une dernière catégorie « autres ».   Cinq années sont étudiées : 1981, 1986, 1991, 1996, 2001. |
| Rebeyrolle, Jacques et Péry-Woodley (2009) | Pour la partie sur les titres d’articles de journaux non scientifiques : **300** titres tirés des journaux *Libération, Le Monde, La Dépêche du Midi, La Provence, Métro* et *Le Journal du Dimanche.*  Il s’agit du seul corpus de travail basé sur des titres de la presse et non pas de textes scientifiques.  Note : cet article traite également des intertitres mais nous écartons dans ce document la partie des données relatives aux intertitres. |
| Whissell (2004) | **1 515** titres du journal scientifique *Psychological Reports* traitant de psychologie :   * 1955-57, 539 titres * 1975, 471 titres * 1995, 505 titres |

Tableau 2 : Corpus des différents articles étudiés

Une caractéristique qui saute aux yeux et la taille relativement petite des corpus : un à 300 (Rebeyrolle, Jacques et Péry-Woodley, 2009), un autre à 751 (Haggan, 2004), deux autres à 1505 (Aleixandre-Benavent et al., 2014) et 1515 (Whissell, 2004), et un à 2172 (Jamali et Nikzad, 2011). Seul Lewison et Hartley (2005), avec 349 700 titres, se rapproche d’une taille conséquente, propre à contenir une grande variété de productions langagières et donc à l’observation et à la quantification de phénomènes particuliers. Notons que le paramètre de la taille n’est pas évoqué dans Gilquin et Gries (2009) pour qualifier un jeu de données de corpus et que les articles cités étant des travaux de linguistique outillée, il est compréhensible que la volumétrie de la plupart de ces corpus reste basse.

Les disciplines des textes scientifiques sont variées : biologie, médecine, psychologie, linguistique, littérature ainsi que d’autres pour *Web of Sciences* et *PLOS ONE*. On note néanmoins une forte représentation de la biologie qui est présente dans 4 études sur 6.

## 1.3 Informations extrinsèques supplémentaires aux titres

Avant d’entrer dans les caractéristiques des titres proprement dites, il faut aborder les informations supplémentaires qui ne sont pas dans le titre mais que le chercheur peut avoir à sa disposition dans les bases de données bibliographiques. Ces informations extrinsèques sont mises en relations avec des caractéristiques intrinsèques du titre dans certains articles. Nous en faisons la liste dans le tableau 4, en précisant les mises en relations déjà faites et les auteurs.

|  |  |
| --- | --- |
| **Caractéristique** | **Description** |
| Discipline(s) | Le texte titré se rattache forcément à une discipline scientifique. Des études montrent parfois une caractérisation forte des titres d’une discipline donnée, comme l’utilisation de phrases complètes dans les articles se rapportant à la biologie selon Haggan (2004). On notera que les disciplines peuvent être hiérarchisées en arbre et qu’un même article peut être étiqueté de plusieurs disciplines. |
| Année | Le texte titré a souvent été publié dans un journal scientifique ou prépublié sur une plate-forme en ligne. On peut se livrer alors à des études en diachronie sur l’évolution de certaines caractéristiques des titres. C’est ce que font Lewison et Hartley (2005) sur 20 ans en prenant 5 années comme échantillon. |
| Nombre d’auteurs | Le texte titré peut être l’œuvre d’un ou de plusieurs auteurs. Lewison et 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. |
| Nationalité | La nationalité du ou des auteurs, celle de la revue ou de la plate-forme où il a été publié peuvent être également étudiées. |
| Nombre d’accès et de téléchargement | Certaines plates-formes électroniques comptabilisent chaque visualisation de la notice de l’article, ouverture et téléchargement. Jamali et Nikzad (2011) mettent en relation le nombre de téléchargements avec la longueur du titre et la présence dans celui-ci d’un double point et d’un point d’interrogation. |
| Nombre de citation | Certaines plates-formes électroniques comptabilisent combien de fois l’article a été cité. Jamali et Nikzad (2011) mettent en relation cette caractéristique avec la longueur du titre et la présence dans celui-ci d’un double point et d’un point d’interrogation. |

Tableau 3 : Caractéristiques en dehors des titres étudiés

## 1.4 Informations intrinsèques aux titres

Nous pouvons à présent entrer dans la caractérisation du titre lui-même. Nous distinguerons les caractéristiques de surface, relevant d’une analyse sommaire du titre, de celles relevant de la lexicologie ou de la syntaxe dans le tableau 5.

|  |  |  |
| --- | --- | --- |
| **Caractéristique** | **Niveau d’analyse** | **Description** |
| Comptage des mots | Surface | Compter le nombre de mots est une première mesure évidente. Néanmoins certains articles (Jamali et Nikzad, 2011) ne comptent que les substantifs. Aleixandre-Benavent et al. (2014) considèrent un titre faisant plus de 20 mots comme trop long. |
| Lexicométries particulières | Lexicologie | Aleixandre-Benavent et al. (2014) mettent en gardent contre l’imprécision de certains titres. Cela peut se faire en comptant le nombre de mots appartenant à un corpus d’un domaine spécialisé par rapport au nombre de mots les plus courants contenus dans le titre. |
| Présence d’acronymes | Lexicologie | Aleixandre-Benavent et al. (2014) mettent en gardent contre l’utilisation d’acronymes qui obscurcissent la compréhension du titre. Cette notion peut être remise en cause : la présence d’un acronyme inconnu peut décider le lecteur à ne pas lire l’article et au contraire, l’expansion d’un acronyme connu des chercheurs d’un domaine peut être inutile. |
| Partitions internes | Surface | Un titre peut être segmenté en partitions, séparées par un signe de ponctuation. Le plus étudié est le double point (Dillon, 1981). Haggan (2004) étudie d’autres signes de ponctuation comme « . », « - » auxquels on peut rajouter selon Aleixandre-Benavent et al. (2014) « ! », « … » et « ? ». |
| Comptage des partitions | Surface | Si le titre présente des partitions, on peut alors compter le nombre de celles-ci puis le nombre de mot dans chaque partition et voir s’il existe un déséquilibre. |
| Présence d’un point d’interrogation | Surface | Le point d’interrogation, utilisé ou non pour faire une partition, mérite que l’on s’y intéresse en lui-même. Pour Jamali et Nikzad (2011), c’est un moyen de renforcer le pouvoir d’attraction d’un type. Aleixandre-Benavent et al. (2014) reconnaissent ce pouvoir d’attraction mais pour eux il s’agit d’une mauvaise pratique car en posant la question on ne parle pas explicitement de l’objet étudié. |
| Présence de citation | Surface | Aleixandre-Benavent et al. (2014) distinguent les titres avec des citations, détectées par la présence de guillemet qu’ils classent là-aussi comme défauts mais sans s’attarder dessus. |
| Structure syntaxique | Syntaxe | Un titre n’est pas forcément une phrase avec un verbe conjugué mais il peut être une phrase nominale. Leech (2000) propose la notion de C-units pour désigner ces ensembles de mots grammaticalement indépendants. Haggan (2004) note que dans son corpus de sciences, tous les titres étant des phrases complètes sont des titres de biologie. |
| Structure syntaxique des partitions | Syntaxe | L’étude de la structure syntaxique peut se faire pour chaque partition du titre. |

Tableau 4 : Caractéristiques internes des titres étudiés

## 1.5 Propositions de typologies des titres

La plupart des articles établissent un classement des titres en fonction de différents critères :

|  |  |
| --- | --- |
| **Articles** | **Classement** |
| Lewison et Hartley (2005) | Classement selon le but : informer ou attirer |
| Jamali et Nikzad (2011) | Classement selon le contenu :   * Descriptif (titre avec conclusion), * Déclaratif (titre avec le sujet), * Question (titre avec sujet sous forme de question) |
| Rebeyrolle, Jacques et Péry-Woodley (2009) | Classement selon ce que gère le titre :   * Fonctionnel (introduction/conclusion), * Gestion de référents, * Gestion de thématique   Classement selon la fonction :   * Désignation / dénomination, * Métadiscursif, * Séduction (seductive). |
| Haggan (2004) | Classement selon la syntaxe :   * Titre étant une phrase complète, * Titre étant composé de plusieurs syntagmes nominaux (souvent avec un double point les séparant en deux parties) * Autres structures |

Tableau 5 : Classements des titres

Après avoir vu les différents corpus étudiés dans la littérature et les caractéristiques extrinsèques et intrinsèques des titres relevés par les différents auteurs ainsi que leurs propositions de typologie, nous pouvons aborder nos propres données.

# 3. Données de travail

Dans cette partie nous présentons nos données de travail, leur source et les interactions que nous avons avec elle pour les récupérer et les traitements qui sont faits dessus. Il s’agit d’un premier aperçu des éléments de notre méthodologie que nous développerons avec nos premières analyses dans la partie suivante également.

## 3.1 Présentation de HAL

Nos titres sont issus de textes scientifiques stockés sur l’archive ouverte *Hyper Article en Ligne* (HAL)[[1]](#footnote-1) (Nivard, 2010). Chaque chercheur, quelle que soit sa discipline, ou documentaliste d’un centre de recherche, est libre de déposer le texte intégral d’un texte scientifique sur HAL, s’il a l’accord de ses auteurs. Il n’y a pas de contrôle par les pairs du contenu scientifique déposé, seulement un contrôle de conformité de surface pour s’assurer du bon format du document et du respect des droits des auteurs. En le déposant sur HAL, le texte est rendu publique et partagé avec la communauté scientifique. Il s’agit d’une façon supplémentaire, et parfois complémentaire, de diffuser son travail par rapport à la publication dans une revue avec un jury de pairs qui acceptent et contrôlent les publications. Les articles scientifiques déposés sur HAL avant ou sans publication dans une revue sont appelés des *preprints*. Le dépôt dans une archive ouverte est plus rapide que la publication dans une revue, mais avec moins de contrôles.

HAL est géré par le Centre pour la Communication Scientifique directe (CCSD)[[2]](#footnote-2), fondé en 2000 et rattaché au Centre National pour la Recherche Scientifique (CNRS)[[3]](#footnote-3). Il existe des sous-ensembles de HAL dédiés à une discipline spécifique, *HAL-SHS* et *MédiHAL*, ou pour un type de texte spécifique comme *Thèses en ligne*. Les avantages des archives ouvertes par rapport à un site d’une institution particulière ou le site web personnel d’un chercheur sont la centralisation de l’accès, la diffusion et le partage des connaissances et la conservation pérenne des textes. 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[[4]](#footnote-4), fondée en 1991. Un dépôt dans HAL entraîne automatiquement la création d’une notice dans arXiv si elle entre dans les disciplines couvertes par cette dernière.

Une notice est créée sur HAL lors du dépôt du texte et éventuellement dupliquée dans d’autres archives ouvertes. Une notice est un ensemble d’informations sur le texte 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 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[[5]](#footnote-5). Ce protocole est standardisé, on peut donc accéder à d’autres archives ouvertes avec, dont arXiv, mais il est assez limité dans ses possibilités de sélection : de base, seule la possibilité de sélectionner des notices en fonction de la date est offerte. Les gestionnaires d’archives peuvent définir des sélecteurs supplémentaires mais ils sont encore très limités. Ce protocole est surtout conçu pour les robots s’assurant de la réplication des notices entre différentes archives ouvertes.

On remarque qu’une archive ouverte A peut avoir la notice d’un texte scientifique hébergé sur une autre archive ouverte B, cette dernière aura à la fois la notice et le texte intégral. Dernier cas possible, il existe des textes qui ne sont pas hébergés par aucune archive ouverte mais simplement référencés par leur notice. Il s’agit généralement de textes dont les droits appartiennent à des revues payantes. La création de notice se fait par le traitement automatisé des références bibliographiques des textes déposés.

Le second protocole offert pour accéder à HAL[[6]](#footnote-6) est bien plus intéressant dans notre cas. Il permet de sélectionner finement nos données et repose sur Apache Solr[[7]](#footnote-7) (Smiley et al., 2015), le moteur de recherche du projet d’Apache Lucene. 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). L’encodage des caractères est de l’UTF-8 pour le JSON. C’est ce second protocole et le format de sortie JSON que nous utiliserons. JSON est un format simple de données structurées stockées dans un fichier texte dont les constituants primaires 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 qui peut être testée simplement dans un navigateur internet. Celui-ci affichera directement le résultat de la requête dans sa fenêtre.

Une caractéristique essentielle commune aux articles étudiés dans l’état de l’art est qu’ils ne portaient que sur des titres scientifiques en anglais, sauf un (Rebeyrolle et Péry-Woodley, 2009) qui portait sur des articles de la presse généraliste en français. Notre travail se distinguera donc déjà par cet aspect puisque notre matériel de départ est constitué de titres scientifiques et en français. De plus, ils s’agissaient de titres d’articles de revues alors qu’HAL héberge d’autres types de textes scientifiques (voir la section 4.2).

## 3.2 Requêtes et résultats

### 3.2.1 Requêtes

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

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 1 : 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 peut demander à Solr un nombre de résultats, disons X, et la création d’un cache. La première réponse de X résultats comporte à la fin un identifiant. En relançant la même requête et en remplaçant la demande de création d’un cache par l’identifiant fourni, on obtient les X résultats *suivants*. Les deux requêtes suivantes (Requêtes 2a et 2b), la première de création, la seconde de poursuite de la demande avec l’identifiant fourni par la première requête, AoFVmLIG , illustrent cette puissante fonctionnalité.

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

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

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

Nous pouvons à présent nous pencher sur les résultats de ces requêtes qui vont constituer notre corpus.

### 3.2.2 Résultats

Notons que, comme le souligne Cori et David (2008), l’étape de sélection des données pour constituer un corpus peut comporter une part de subjectivité. Dans notre cas nous nous en exemptons car nous n’opérons pas de choix dans les titres que nous retournent HAL. Si nous devions préparer plus avant nos données, en particulier en écartant certains titres pour des raisons d’invalidité technique ou de doublon, nous respecterions la précaution méthodologique que préconisent ces deux auteurs, faire *« un inventaire soigneux de toutes les décisions prises en amont »* L’encadré Résultat 3 présente un élément tiré d’un exemple de résultat de 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 3 : exemple d’un élément de résultats au format JSON

Un premier jeu de données, retravaillé à partir de requêtes à HAL pour être mis dans un format tabulaire, nous a été fourni par M. Tanguy le 20 octobre 2017. Il s’agit de notre premier corpus de travail. Il comporte 146 603 titres, accompagnées d’informations supplémentaires que nous détaillons dans le paragraphe suivant. Sa taille le rapproche bien plus de celui utilisé par Lewison et Hartley (2005) que de tous les autres, même si ce dernier reste plus de deux fois plus volumineux. C’est ce premier jeu qui nous a servi pour nos réflexions et forger notre problématique.

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, on parlera de caractéristiques extrinsèques, mais elles sont très intéressantes pour nous aider à mieux comprendre le 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. Le champ étant une « discipline racine » ou « super discipline ».

Ces métadonnées pourront ensuite être mises en relation avec une caractérise intrinsèque du titre. Des auteurs avaient déjà mis en avant certaines régularités. Haggan (2004) avait par exemple montré que l’utilisation d’une phrase complète pour titre était une caractéristique des titres en biologie. Lewison et 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. En guise d’exemple, Résultat 4 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 4 : une ligne de notre premier corpus de travail

Après cette présentation de nos données de travail, nous prouvons entrer dans le vif de celui-ci en dévoilant nos premières analyses de ces données.

# 4. Premières analyses et premiers résultats

Dans cette partie nous présentons quelques éléments d’analyse des données que nous avons déjà effectués. À travers ces analyses, nous continuons à exposer notre méthodologie en construction. Pour ce qui est du côté applicatif, nous utilisons pour nos recherches un script simple développé dans le langage de programmation Python[[8]](#footnote-8). Nous essayons d’avoir un développement itératif avec plusieurs niveaux de profondeurs. Un niveau surfacique d’étude des titres et des algorithmes simples a été privilégié dans ce document pour avoir une vue d’ensemble de notre corpus et de notre matériau de travail.

## 4.1 Les titres comme matériel pour le TAL

Du point de vue du traitement automatique des langues (TAL), l’avantage de ce type de données écrites est que la production est très travaillée par rapport à celle que l’on trouverait sur Twitter, dans des SMS ou sur des forums. Les écarts de graphie dus à une faute de frappe où une erreur du ou des scripteurs sont extrêmement rares étant donné le soin apporté à l’écriture du titre et les éventuelles relectures successives. Cela nous évitera d’avoir à les considérer ou à ajouter des traitements pour tolérer de tels écarts et erreurs dans les algorithmes. Leur brièveté est un avantage dans le sens où, en termes de temps de calculs, on peut se livrer à une étude plus poussée sur un plus grand échantillon de titres, ce qui serait plus difficile face à des textes plus longs.

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

Nous avons remarqué néanmoins certaines incohérences dans nos données : titres en double, titre se terminant par un point-virgule ou n’ayant pas un nombre de guillemets ouvrants égal à au nombre de guillemets fermants (voir la section 4.3). De plus, certains titres utilisent un retour à la ligne comme caractère de segmentation mais celui disparaît dans la notice HAL. Il faut toujours garder un fond de méfiance lorsque l’on aborde des données textuelles et cette méfiance doit se traduire par des gardes dans nos algorithmes. Intéressons-nous à présent aux caractéristiques principales de nos données.

## 4.2 Types des textes scientifiques dans notre corpus et HAL

Nous étudierons les titres en partant de leurs métadonnées puis de leur surface. Nous nous intéressons tout d’abord au type de textes scientifiques dont nous avons le titre. Pour cela il suffit d’interroger les métadonnées. Nous obtenons le tableau 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type (traductions[[9]](#footnote-9))** | **Nb dans corpus** | **% dans corpus** | **Nb dans HAL** | **% dans HAL** |
| Article (Article dans une revue) | 45 355 | 30.94 % | 681 988 | 49.65 % |
| Report (Rapport) | 3 558 | 2.43 % | 25 144 | 1.83 % |
| Other (Autre publication) | 7 803 | 5.32 % | 31 462 | 2.29 % |
| ConferenceObject (Communication dans un congrès) | 35 860 | 24.46 % | 393 978 | 28.68 % |
| Preprint (Pré-publication, document de travail) | 2 878 | 1.96 % | 37 462 | 2.73 % |
| BookPart (Chapitre d’ouvrage) | 23 747 | 16.2 % | 105 565 | 7.69 % |
| Lecture (Cours) | 291 | 0.2 % | 1 102 | 0.08 % |
| Book (Livre) | 10 831 | 7.39 % | 24 705 | 1.8 % |
| Patent (Brevet) | 323 | 0.22 % | 3 179 | 0.23 % |
| DoctoralThesis (Thèse) | 15 954 | 10.88 % | 68 927 | 5.02 % |
| MasterThesis (Mémoire d’étudiant) | 2 | <0.1 % | 26 | <0.1 % |
| **Total** | 146 603 | 100 % | 1 373 538 | 100 % |

Tableau 6 : Répartition des titres par type

Nous avons choisi de mettre également les nombres pour l’ensemble de l’archive ouverte HAL. Cela dans un but d’essayer de mesurer la représentativité de notre corpus par rapport à l’ensemble des données. Un bon corpus se doit d’être représentatif ce qui est globalement le cas. Les livres sont néanmoins surreprésentés (~7% contre ~2%) dans notre corpus, les chapitres d’ouvrage (~16% contre ~8%) ainsi que les thèses (~11% contre ~5%). Les articles sont par contre sous représentés (~31% contre ~50%). 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 les disparités que nous constations avec l’ensemble des données de HAL nous mettent en garde contre toutes généralisations hâtives des conclusions que nous pourrions découvrir sur notre corpus. De plus, si nous devions associer une caractéristique quelconque des titres avec une caractéristique extrinsèque particulière, comme le type de texte par exemple, nous construirions des corpus monotypes pour vérifier nos hypothèses.

## 4.3 Longueurs des titres et présences des caractères segmentant

Une seconde tâche est de compter les mots, encore faudra-t-il décider si on les compte tous : Jamali et Nikzad (2011) ne comptent par exemple que les substantifs. Une troisième tâche est d’évaluer le nombre et le type des partitions internes dans les titres. Le caractère segmentant peut-être un double point mais aussi un point, un point d’interrogation, un point d’exclamation ou d’autres signes de ponctuation repérés par Haggan (2004). Nous avons déjà donné quelques coups de sonde dans notre corpus sur certaines caractéristiques dont nous présentons les résultats dans le tableau 7. Une remarque importante est de pas confondre le point final, d’interrogation ou d’exclamation à l’intérieur d’un titre, qui le partitionne, de celui pouvant éventuellement le terminer.

|  |  |  |
| --- | --- | --- |
| **Caractéristique** | **Valeur** | |
| Nombre de titres | 146 603 | |
| Estimation de la longueur moyenne | 13 mots / titre | |
| **Nombre de titres avec au moins un…** |  | Dont en dernière position |
| Double point | 53 169 (36,27%) | 0 (0%) |
| Point | 26 121 (17.82 %) | 7 230 (27.68%) |
| Point d’interrogation | 11 949 (8,15%) | 7 064 (59 .12%) |
| Guillemets : «, ", » | « 6902  » 6939[[10]](#footnote-10)  " 5813 |  |
| Point-virgule | 1 256 (0.86 %) | 5 (0.4%) |
| Point d’exclamation | 671 (0,46%) | 264 (39.34%) |

Tableau 7 : Présence de quelques caractères segmentant dans notre corpus

Il faut lire ce tableau ainsi : il y a 11 949 titres avec au moins un point d’interrogation. Parmi ceux-ci, il y en a 7 064, environ 60%, qui l’utilise comme symbole terminal. Les plus intéressants pour l’étude des partitions sont bien sûr les 40% restant qui utilisent le point d’interrogation à l’intérieur du titre, comme caractère segmentant.

On voit que la structure qui nous intéresse, utilisant un double point, est très bien représentée dans ce corpus : plus d’un tiers des titres de celui-ci en comporte un. De plus, il n’est jamais utilisé comme symbole terminal, ce qui semble logique. Nous regarderons lexicalement les mots situés immédiatement à droite du deux points, en faisant tout d’abord une approche lexicométrique.

## 4.4 Le lemme d’après

Nous avons voulu observer quelles formes et quels lemmes nous trouvions à proximité du double point sur sa droite. Nous avons amélioré notre script pour en obtenir une première vision, avec une fenêtre de 3 formes à droite. Bien sûr, les articles sont parmi les plus présents, ainsi que des coordinations et des prépositions. Si on les élimine pour ne garder que les 15 substantifs les plus employés, on obtient le tableau 8. Nous avons additionné les différentes formes d’un même lemme :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Forme** | **Compte** | **Forme** | **Compte** | **Forme** | **Compte** |
| Cas (m) | 1758 | analyse (f) | 838 | état (m) | 283 |
| étude (f) | 1210 | enjeu (m) | 669 | effet (m) | 269 |
| application (f) | 1087 | apport (m) | 404 | réflexion (f) | 256 |
| exemple (m) | 1019 | histoire (f) | 345 | modèle (m) | 255 |
| approche (f) | 1006 | rôle (m) | 321 | outil (m) | 251 |

Tableau 8 : lemmes les plus présents après un double point

Nous avons donc effectué une cooccurrence non pas entre deux lemmes mais entre une marque de ponctuation et un lemme. Nous avons effectué la lemmatisation à l’aide de règles simples, les suppressions des -s et des -x finaux, et sélectionné les substantifs à la main. Nous envisageons éventuellement dans une itération ultérieure d’utiliser des ressources supplémentaires comme un lexique morphologique et syntaxique tel que le Lexique des Formes Fléchies du Français (Sagot, 2010). Le LEFFF a été développé au sein de l’Institut National de Recherche en Informatique et en Automatique (INRIA) puis du Laboratoire Bordelais de Recherche en Informatique (LARBI) depuis 2003 par Lionel Clément et Benoît Sagot. En accès libre et gratuit, il donne, pour de nombreuses formes, son lemme de rattachement et la catégorie syntaxique de celui-ci. De plus, nous restons pour le moment en surface, sans chercher à comprendre syntaxiquement le syntagme nominal à droite du double point en construisant un arbre de constituants ou un graphe de dépendances. Pour cela, nous devrions faire appel à des outils plus puissants.

À ce tableau, nous devons ajouter le cas de *nouvelle* qui compte 348 occurrences sous les formes *nouvelle* et *nouvelles*. Notre algorithme ne distinguant par les catégories syntaxiques, il nous est difficile de savoir s’il s’agit du lemme nouvelle : nom commun ou du lemme nouveau, nouvel, nouvelle : adjectif. Pour comprendre cela, nous avons cherché à visualiser ses cooccurrences. Après les avoir observées, nous pouvons affirmer que l’extrême majorité des 348 occurrences concernent l’adjectif.

Sémantiquement, les mots retournés sont très intéressants : on pourrait les qualifier intuitivement d’objets de la recherche scientifique. Nous colorons ceux se rattachant à ce trait en vert. Les 4 derniers en orange, *enjeu, histoire, rôle* et *état* peuvent se rattacher à certaines disciplines ou à notre premier trait selon le contexte. Nous proposons à présent quelques exemples de titres.

## 4.3 Quelques exemples de titres

Nous avons ici sélectionné quelques exemples de titre tirés de notre corpus correspondant à des réalisations de la structure lexico-syntaxique qui nous intéresse :

* **Loches. Première campagne de fouilles sur la forteresse : Rapport de sondages archéologiques** [Ligne 24082, Identifiant hal-01269678, Archéologie]

Ce titre est un exemple quasi-prototypique de la structure qui nous intéresse. Son seul écart par rapport au prototype est d’avoir trois partitions au lieu de deux. Le double point est sinon immédiatement suivi d’un seul syntagme nominal dont le noyau est le mot *rapport* qui indique la nature du document. On note également qu’il y a une autre partition au début, qui précise un lieu géographique.

* **Le promontoire du "Malpas" à Soyons (Ardèche) : un habitat fortifié à l'âge du Fer sur les rives du Rhône** [Ligne 95974, Identifiant halshs-01359782, Archéologie]

Ce titre utilise encore une fois la structure qui nous intéresse. Il comporte deux partitions. Cette fois-ci, le noyau du syntagme nominal à la droite du double point n’apporte pas une précision sur la nature du document mais sur son sujet. Notons également l’utilisation des guillemets pour encadrer une indication géographique. Leur utilisation est désapprouvée par Aleixandre-Benavent et al. (2014).

* **Qualité et consommateurs : nouvelles approches de la segmentation** [Ligne XXX, Identifiant hal-01197932, Économie]

Ce titre est un exemple prototypique de notre patron. On remarque une utilisation de l’adjectif *nouveau, nouvel, nouvelle* qui qualifie le noyau du syntagme nominal, le lemme *approche* avec un complément.

* **Dynamique des structures : méthodes approchées, cinématiques ; Analyse Modale ; Recalage de Modèle** [Ligne 32, Identifiant cel-00611705]

Ce titre est intéressant car il fait survenir un nouvel élément : l’utilisation de points-virgules. Ce titre est composé de 4 partitions, ce qui très important. Il est intéressant car on peut estimer qu’il y a une hiérarchie des symboles de partition. Intuitivement, le point-virgule semble établir la partition la plus forte, le double point crée une partition moyenne et enfin la virgule qui ponctue une énumération et non une partition.

# 5. Pistes et perspectives

Dans cette partie, nous évoquons plusieurs pistes et perspectives pour la poursuite de notre travail au second semestre.

## 5.1 Continuer à lire d’autres travaux

Même si ce document constitue la plus grande partie de l’état de l’art de notre mémoire de recherche, nous ne doutons pas de n’avoir lu qu’une fraction de ce qui est disponible. Nous aimerions continuer à explorer des articles d’autres auteurs en construisant notamment des graphes de références croisées. Si trois articles A, B et C que nous avons lus citent l’article D, il sera intéressant pour nous de lire par exemple. L’état de l’art final bénéficiera ainsi d’une amélioration continue même si le noyau dur est à présent constitué.

## 5.2 Construire nos propres corpus

Nous avons exploré les possibilités d’Apache Solr pour récupérer depuis HAL des notices bibliographiques au format JSON mais notre corpus de travail est actuellement un fichier au format tabulaire donné par M. Tanguy. Nous aimerions vérifier que nous sommes capables de reconstituer nous-mêmes la chaîne de traitement ayant servi à obtenir ce fichier au format tabulaire. Nous souhaiterions utiliser cette chaîne pour produire des corpus spécialisés : par discipline ou par type de texte scientifique par exemple.

## 5.3 Réécriture des outils

Nous avons écrit rapidement pour supporter ce document un script simple en Python. Armé de ce que nous savons et de ce que nous voulons savoir, nous aimerions réécrire ce script pour qu’il soit plus intelligent et plus fin dans les mesures et les structures qu’ils ramènent. Éventuellement en l’associant à la tâche de création de corpus et d’autres outils (TreeTagger) ou références (LEFFF) disponibles. Nous pensons par exemple à la détection du lemme substantif qui constitue le noyau du syntagme nominal à la droite du double-point. Nous aimerions également qu’ils soient capables de gérer des ensembles de résultats : exemple, regrouper toutes les structures suivant notre patron qui ont pour lemme noyau « approche ». Éventuellement, procéder également à l’analyse en constituant du syntagme nominal.

## 5.4 Explorer notre patron en profondeur

La poursuite de notre travail doit également permettre d’approfondir complètement notre connaissance du patron que nous avons choisi et ses réalisations. Notre chaîne de traitement doit faire émerger les variations du patron « … : SN … ». Pour cela, nous devons détecter à droite le syntagme nominal et ses membres. Éviter de faire toute l’analyse syntaxique du titre serait un grand avantage. Nous savons le caractère délimitant à gauche le syntagme nominal : un double point. Il nous faut à présent réfléchir à comment le délimiter sur la droite. Les travaux de Quiniou et al. (2012) nous mettent sur la piste de chercher des motifs émergents de syntagmes nominaux. Intuitivement, on peut penser qu’il se caractérise d’un article (optionnel), d’une suite d’adjectifs, éventuellement modifiés par un adverbe, du substantif noyau et d’adjectifs postposés, éventuellement complémentés, ce que ne sont jamais les adjectifs postposés. Mais il serait plus intelligent de s’arrêter dès le noyau trouvé. Des sous-patrons pourraient alors émerger.

# Conclusion

Durant ce premier semestre, nous avons défini notre problématique : faire une étude corpus-driven, mêlant les niveaux syntaxique et lexical, de la structure lexico-sémantique **« … : SN »**. Éventuellement, nous nous réservons la possibilité d’ajouter le niveau sémantique à notre étude, voir d’autres structures lexico-syntaxique, si nous avons le temps sans trop diluer le focus de notre projet de recherche.

Nous avons exploré des travaux précédents ayant les titres pour objet d’étude. Aucun, parmi ceux que nous avons étudiés, n’abordait les titres de textes scientifiques en français. Un seul abordait des titres en français, mais seulement tirés de la presse généraliste. Notre étude apportera déjà cette originalité, celle d’étudier des titres scientifiques en français dans une très grande variété de disciplines.

Pour constituer nos données, nous avons étudié les possibilités d’accès à l’archive ouverte HAL à l’aide la plateforme Apache Solr. Nous avons mis au point nos premiers scripts Python pour interroger en surface nos données. Ces interrogations ont permis de voir émerger des découvertes et de nouvelles interrogations pour préciser les premières. Nous avons commencé à développer une approche itérative pour raffiner nos interrogations et obtenir peu à peu une compréhension plus fine des phénomènes observés dans notre corpus.

Dans les pistes et perspectives, nous interrogeons sur les suites à donner à notre travail pour aboutir au mémoire du projet de recherche du second semestre. Une refonte de l’outillage et une exploration en profondeur de notre patron, en explorant sa variation et ses sous-patrons et au programme.

# Table des requêtes et résultats

[Requête 1 : Un exemple de requête avec l’API Apache Solr formaté pour plus de lisibilité 9](#_Toc504428837)

[Requêtes 2a et 2b : création et consultation d’un cache 9](#_Toc504428838)

[Résultat 3 : exemple d’un élément de résultats au format JSON 10](#_Toc504428839)

[Résultat 4 : une ligne de notre premier corpus de travail 10](#_Toc504428840)

# Table des tableaux

[Tableau 1 : Sujet des articles étudiés 3](#_Toc504428841)

[Tableau 2 : Corpus des différents articles étudiés 4](#_Toc504428842)

[Tableau 3 : Caractéristiques en dehors des titres étudiés 5](#_Toc504428843)

[Tableau 4 : Caractéristiques internes des titres étudiés 6](#_Toc504428844)

[Tableau 5 : Classements des titres 7](#_Toc504428845)

[Tableau 6 : Répartition des titres par type 12](#_Toc504428846)

[Tableau 7 : Présence de quelques caractères segmentant dans notre corpus 12](#_Toc504428847)

[Tableau 8 : lemmes les plus présents après un double point 13](#_Toc504428848)

# Bibliographie

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

Bray, T. (2017). The javascript object notation (json) data interchange format. IETF.

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

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

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.

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

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

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

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

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

Nivard, J. (2010). Les archives ouvertes. *La Lettre de l'École des hautes études en sciences sociales*, n°34.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012, June). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. In Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12) (pp. 821-833).

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

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

Schmid, H. (2013, November, 1ed 1994). Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing* (p. 154).

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

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

Swales, John M., 1990. Genre Analysis. English in Academic and Research Settings. Cambridge University Press.

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.

1. <https://hal.archives-ouvertes.fr/> [↑](#footnote-ref-1)
2. <https://www.ccsd.cnrs.fr/> [↑](#footnote-ref-2)
3. <http://www.cnrs.fr/> [↑](#footnote-ref-3)
4. <http://arxiv.org/> [↑](#footnote-ref-4)
5. <http://www.openarchives.org/OAI/openarchivesprotocol.html> [↑](#footnote-ref-5)
6. <https://api.archives-ouvertes.fr/docs> [↑](#footnote-ref-6)
7. <https://lucene.apache.org/solr/> [↑](#footnote-ref-7)
8. <https://www.python.org/> [↑](#footnote-ref-8)
9. Les traductions et les valeurs ont été prises depuis cette page, consultée le 21 janvier 2018 : <https://hal.archives-ouvertes.fr/browse/doctype> Nous avons ajouté documents et notices pour faire un total. [↑](#footnote-ref-9)
10. Les guillemets français « » présente l’avantage d’être différenciés entre l’ouvrant et le fermant. **Nous remarquons que 293 titres, soit 0,19% ont des problèmes de cohérence : présence d’un ouvrant sans le fermant ou présence d’un fermant sans l’ouvrant.** Le nombre de titres concernés est néanmoins très faible. [↑](#footnote-ref-10)