##### 

MÉMOIRE DE RECHERCHE

Département des Sciences du Langage

M2 Linguistique, Informatique et Technologies du Langage (LITL)

Étude de la sous-spécification   
dans les titres d’articles scientifiques

Damien GOUTEUX

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

2018 – 2019

Table des matières

[Résumé 5](#_Toc19482450)

[Remerciements 6](#_Toc19482451)

[Introduction 7](#_Toc19482452)

[I. Exploration du corpus à la lumière de l’état de l’art 12](#_Toc19482453)

[I.1 Origine et prétraitements des données 12](#_Toc19482454)

[I.1.1 Récupération des données 12](#_Toc19482455)

[I.1.2 Étiquetage et analyse syntaxique en dépendances 13](#_Toc19482456)

[I.1.3 Segmentation des titres 14](#_Toc19482457)

[I.1.4 Sélection de la tête des segments 14](#_Toc19482458)

[I.2 Constitution d’un corpus de travail représentatif 17](#_Toc19482459)

[I.2.1 Sélection selon la structure des titres 18](#_Toc19482460)

[I.2.2 Sélection selon la nature des têtes 19](#_Toc19482461)

[I.2.3 Un corpus de travail représentatif du matériau de base 21](#_Toc19482462)

[I.3 Structures semblables des domaines et têtes spécifiques 22](#_Toc19482463)

[I.3.1 Variations de la structure en fonction du domaine 22](#_Toc19482464)

[I.3.2 Têtes spécifiques à un domaine 25](#_Toc19482465)

[II. Têtes transdisciplinaires et NSS dans notre corpus 31](#_Toc19482466)

[II.1 Sélection des têtes transdisciplinaires 31](#_Toc19482467)

[II.1.1 Principe de sélection 31](#_Toc19482468)

[II.1.2 Résultats et évaluations du résultat 31](#_Toc19482469)

[II.1.3 Études des têtes selon leurs segments et la structure segmentale du titre 33](#_Toc19482470)

[II.2 Noms sous-spécifiés et constructions spécificationnelles 34](#_Toc19482471)

[II.2.1 Définitions des noms sous-spécifiés 35](#_Toc19482472)

[II.2.2 Les constructions spécificationnelles classiques 36](#_Toc19482473)

[II.2.3 Recherche des CS classiques dans notre corpus 42](#_Toc19482474)

[II.3 Schémas récurrents d’emploi des têtes transdisciplinaires 46](#_Toc19482475)

[II.3.1 Recherche des schémas d’emplois des têtes transdisciplinaires 46](#_Toc19482476)

[II.3.2 Lexique des noms et détermination de l’emploi 49](#_Toc19482477)

[II.3.4 Transdisciplinarité des schémas 55](#_Toc19482478)

[Discussion sur nos résultats, limites et perspectives 58](#_Toc19482479)

[III.1 Éléments de discussion 58](#_Toc19482480)

[Limite de l’analyse en dépendance automatique de Talismane 58](#_Toc19482481)

[Limitations des têtes spécifiques aux domaines 58](#_Toc19482482)

[Têtes transdisciplinaires 59](#_Toc19482483)

[Listes de NSS et des lexiques scientifiques 59](#_Toc19482484)

[Opérationnalisation des NSS 59](#_Toc19482485)

[Conclusion 61](#_Toc19482486)

[Bibliographie 63](#_Toc19482487)

[A1. Distance des domaines de par leurs têtes spécifiques 69](#_Toc19482488)

[A2. Combinaisons des têtes de titres bisegmentaux 70](#_Toc19482489)

[A3. Liste des têtes transdisciplinaires 72](#_Toc19482490)

[A4. Étiquettes utilisées par Talismane et HAL 76](#_Toc19482491)

[A4.1 Catégories morphosyntaxiques de Talismane 76](#_Toc19482492)

[A4.2 Code des 27 domaines de HAL retenus 77](#_Toc19482493)

[A5. Éléments techniques 79](#_Toc19482494)

[A5.A Présentation de l’API de requêtage de notre corpus 79](#_Toc19482495)

[A5.B Description de nos données informatiques 79](#_Toc19482496)

[A5.C Analyse de 100 titres traités par Talismane 80](#_Toc19482497)

[A6. Index des tableaux 86](#_Toc19482498)

# Résumé

Nous étudions la sous-spécification des têtes transdisciplinaires dans les titres de publications scientifiques. Nous rapprochons ces têtes transdisciplinaires, très fréquentes dans de nombreux domaines scientifiques, des noms employés de façon sous-spécifiée : des noms dont la carence sémantique est comblée par le contexte, nom et contexte étant reliés par une construction spécificationnelle. Pour mener à bien ce rapprochement, nous étudions les schémas récurrents dans lesquels s’insèrent les têtes transdisciplinaires des titres. Nous étudions ces schémas pour les rapprocher des constructions spécificationnelles déjà relevées dans la littérature. Ce rapprochement syntaxique et fonctionnel nous permet de dresser une liste des têtes transdisciplinaires s’employant le plus souvent de façon sous-spécifiées. Enfin, nous mettons en rapport les têtes sous-spécifiées et les schémas récurrents trouvés avec les différents domaines scientifiques.

Mots-clés : titre, tête, schéma, patron, nom porteur, emploi sous-spécifié, sous-spécification, construction spécificationnelle, contenu spécifiant, publication, article, domaine, discipline.

# Remerciements

Je tiens à remercier mes deux codirecteurs de recherche, Mme Josette Rebeyrolle et M. Ludovic Tanguy, pour leur suivi, leur aide et leurs précieux retours sur mon travail tout au long de l’année.

Mes remerciements vont également à 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, me permettant d’intégrer le master LITL tout en conservant mon emploi salarié.

Je tiens également à remercier M. Jean Favre, M. Olivier Dunyach et M. Frédéric Gil de la société Scalian DS et M. Dimitri Cognet de la société Thales Alenia Space pour avoir accueilli mon projet de reprise d’études favorablement.

Je veux remercier aussi mes camarades de promotion du master LITL qui ont bien voulu accepter mes contraintes d’emploi du temps liés à mon double statut d’étudiant et de salarié et avec qui j’ai partagé de nombreux moments studieux et amicaux.

Enfin, je tiens à remercier ma famille et mes amis, pour leur patience et leurs encouragements, pour m’avoir eux aussi accompagné durant les deux années de ce master.

# Introduction

Un titre de document scientifique est un énoncé singulier d’une importance cruciale. D’une part, il s’agit d’un texte très court d’une dizaine de mots. D’autre part, il constitue le premier contact entre un document et ses lecteurs, suivi du résumé et enfin du document en lui-même (Whissell, 2012). Dans 92 % des cas, le lecteur s’arrêtera au titre (Mabe et Amin, 2002) : soit car le titre lui apporte la réponse qu’il cherche (Salager-Meyer et Alcaraz Ariza (2013) rapportent ainsi que certains médecins prennent des décisions cliniques fondées uniquement sur les titres), soit le plus souvent car le lecteur détermine que l’article ne l’intéresse pas. C’est donc sur la lecture du titre seul, indépendamment du document titré, que le chercheur fait son tri parmi la littérature scientifique (Goodman et al., 2001) dont la production augmente constamment en doublant tous les 12 ans (Stix, cité dans Salager-Meyer et Alcaraz Ariza, 2013).

Le tri effectué sur la seule lecture du titre soulève de nombreuses questions concernant l’information que contient un titre, les mots et les structures utilisés pour convoyer l’information et leurs évolutions au fil du temps, le fonctionnement du titre et son intégration dans le discours, la recherche de règles prescriptives pour écrire un bon titre, ou encore la définition de ce qu’est un bon titre et les différentes mesures de performances d’un titre comme le nombre de téléchargements ou de citations du document. L’intérêt des chercheurs pour ces questions s’est traduit par de nombreux articles sur les titres en anglais. Les titres en français ont été moins étudiés. On peut néanmoins citer les travaux de Ho-Dac et al. (2004), Rebeyrolle et al. (2009) et Tanguy et Rebeyrolle (à paraître).

Un consensus émerge pour accorder deux fonctions principales aux titres (Haggan, 2004 ; Hartley, 2005, 2007 ; Salager-Meyer et Alcaraz Ariza. 2013 ; Whissell, 2012). La première est d’informer le lecteur sur le contenu du document en présentant ses points principaux : son sujet et son périmètre d’étude au minimum et éventuellement la méthode employée et les résultats obtenus (Anthony, 2001 ; Cheng et al., 2012 ; Jamali et Nikzad, 2011 ; Paiva et al., 2012 ; Swales et Feak, 1994, p. 205). La seconde fonction est d’attirer l’attention du lecteur (Hartley, 2005). Dans notre travail, nous prenons en compte uniquement la fonction informationnelle du titre, considérant, comme Grant (2013), Haggan (2004) et Hartley (2005), qu’elle est la plus importante. Notamment car elle oblige à ce que les mots du titre reflètent le contenu du document pour faciliter son appréhension cognitive lors du parcours d’une liste de titres par un chercheur, une activité scientifique régulière (Soler, 2007, p. 91), ou pour faciliter sa sélection automatique par des algorithmes (Haggan, 2004 ; Hartley, 2007 ; Merrill et Knipps, 2014), même si la recherche en plein texte diminue cet aspect. La fonction informationnelle est également plus facile à analyser que la fonction d’attraction qui peut considérablement obscurcir le sens d’un titre pour susciter l’intérêt (Hartley, 2005) ou faire appel à des notions complexes pour le traitement automatique des langues comme l’humour (Sagi et Yechiam, 2008 ; Subotic et Mukherjee, 2014). Enfin, Jalilifar et al. (2010) soutiennent qu’un titre ne peut pas être uniquement attractif, car ce serait une bien piètre indication pour les lecteurs. Un titre doit toujours donner des indices informationnels sur le contenu du document, l’attractivité étant là pour le distinguer des autres titres. En suivant ces auteurs, on peut donc considérer que la fonction d’attraction est secondaire par rapport à la fonction informationnelle.

La majorité de la littérature sur les titres traitant de titres d’articles de journaux scientifiques et de communications (Goodman et al., 2001, Haggan, 2004, Soler, 2007, Wang et Bai, 2007), notre travail se limite à ce type de publication et à celles dont les titres sont construits de manière similaire : chapitres d’ouvrages collectifs et communications ou posters dans des congrès ou des conférences.

Notre travail de première année avait constitué à étudier les noms apparaissant immédiatement après un double point dans les titres de publications scientifiques. Nous disposions d’un corpus de 85 500 titres en français de différents types de publications scientifiques, dont les plus nombreux étaient les articles, les communications et les chapitres d’ouvrage, issus d’un grand nombre de domaines scientifiques (voir la partie I.1.1 Récupération des données qui reprend la méthode de constitution de ce premier corpus pour établir notre corpus de travail de cette année).

Nous avions observé trois schémas lexico-syntaxique récurrents d’emplois de ces noms couvrant 65 % des titres de notre corpus : 1) un syntagme nominal dont le noyau avait pour complément un syntagme prépositionnel enchâssant lui-même un syntagme nominal, couvrant 50 % de notre corpus, 2) un syntagme prépositionnel enchâssant le premier cas, couvrant 5 %, et 3) deux syntagmes nominaux coordonnés, couvrant 10 %. Nous avions également constaté l’utilisation récurrente et transdisciplinaire d’un lexique de noms communs placés immédiatement après le double point, dont les onze plus fréquents étaient : *étude, cas, approche, analyse, application, pratique, exemple, enjeu, perspective, modélisation, limite*.

Les noms distingués sont tous abstraits, ils ne dénotent pas un objet tangible du monde réel, et sont liés au domaine scientifique : on les retrouve tous, sauf *enjeu*, dans le lexique transdisciplinaire des écrits scientifiques (LTES) décrit par Tutin (2008). Par ailleurs, nous avons remarqué une similitude lexicale avec une classe de noms employés de façon très fréquente dans le discours académique (Flowerdew et Forest, 2015, p. 1) avec un contenu sémantique très faible spécifié par le contexte. On appelle les membres de cette classe des noms (en emploi) sous-spécifiés (NSS). Nous nous interrogeons donc sur l’éventuelle sous-spécification des noms employés dans les titres.

Pour enquêter plus largement sur ce phénomène de sous-spécification, nous souhaitons étendre notre périmètre d’étude à l’ensemble du titre, avant et après le double point, ou toute autre marque de ponctuation segmentant le titre en deux, et l’étendre également aux titres composés d’un seul segment. Nous prenons comme objet d’étude la tête, aussi appelée noyau ou racine, de ces segments. Le premier nom immédiatement après le double point étudié en première année était, à une préposition près dans le cas du deuxième schéma, la tête du second segment du titre. La redéfinition de notre objet d’étude englobe donc les noms appartenant au lexique commun trouvé en première année. La question est donc de savoir si les têtes de segments sont employées de manière sous-spécifiée.

Dans l’exemple (1) [[1]](#footnote-2) ci-dessous, le titre est constitué de deux segments, délimités par un double point. L’exemple (2) montre lui un titre constitué d’un seul segment. Nous en mettons en gras la tête de chaque segment.

(1) Un nouvel **OVNI** dans le ciel réunionnais : la **transparence** des prix

(2) **Problème** d'acquisition de données par une torpille

Ce nouveau travail doit donc commencer par le découpage de nos titres en segments en reprenant et en amendant une liste de signes de ponctuation qui segmentent les titres en anglais, établie par Anthony (2001). Ensuite, pour trouver les têtes de syntagmes, plutôt que de simplement parcourir le segment et prendre le premier nom rencontré comme nous le faisions en première année, nous avons décidé d’utiliser l’analyse syntaxique en dépendances (Tesnière, dans Schwischay, 2001) pour identifier les têtes de segments.

Ce sont les têtes transdisciplinaires, très fréquentes dans de nombreux domaines scientifiques, qui nous semblent les meilleures candidates pour être rapprochées des NSS. On prend pour hypothèse que leur capacité à apparaître très fréquemment dans la plupart des domaines n’est possible qu’à cause d’un faible contenu sémantique, au contraire de noms sémantiquement pleins bien plus connotés à un domaine particulier. *Céramique* évoque ainsi directement l’archéologie, l’art et l’histoire de l’art ou l’histoire, alors que *problème* semble sémantiquement creux. Seule la prise en compte du contexte de la tête transdisciplinaire permet d’accéder à son sens complet : de quel problème il s’agit, sur quoi porte-t-il ?

Deux difficultés apparaissent déjà : les NSS sont une classe considérée comme ouverte (Flowerdew et Forest, 2015, p. 12 ; Schmid, 2000 , p. 4) et une classe d’emploi, non une nature lexicale (Flowerdew et Forest, 2015, p. 7 ; Schmid, 2000, p. 13). On ne peut donc déterminer s’il s’agit d’un NSS qu’en fonction de son contexte où il est relié à un contenu qui le spécifie. La liaison entre le contenu spécifiant et le NSS se fait via une construction spécificationnelle (CS). Les deux constructions spécificationnelles les plus fréquemment étudiées, notamment par Schmid (2000, p. 22) pour l’anglais et transposées par Legallois (2008, p. 2) en français, sont :

1. **NSS** + *être* + proposition subordonnée conjonctive commençant par *que*:   
   Le **problème** est que je n’avais pas d’argent.
2. **NSS** + *être* +proposition subordonnée infinitive commençant par *de* :   
   Le **problème** est de ne pas avoir d’argent.

La question se pose de savoir si on retrouve nos noms dans de telles constructions dans les titres. Si le verbe *être* est optionnel pour Schmid (2000, p. 22), il n’en reste pas moins que la proposition doit contenir un verbe, conjugué dans le cas d’une proposition subordonnée conjonctive ou à l’infinitif pour une proposition subordonnée infinitive. Or, de nombreux travaux (Leech, 2000 ; Haggan, 2004 ; Soler, 2007 ; Cheng et al., 2012 ; Wang et Bai, 2007) soulignent la nature nominale des titres. Un milieu largement « averbal » comme les titres fait douter de retrouver les CS-I et CS-II dedans.

Néanmoins, Schmid (2000, p. 26) mentionne une autre CS de la forme **NSS** + de + syntagme nominal. Or, la complémentation des noms est une caractéristique de l’écriture académique (Biber et Gray, 2010), nos têtes pourraient donc s’insérer dans cette CS. Prenons les deux exemples (3) et (4), pour éclairer l’hypothèse d’un rapprochement possible entre les têtes et les NSS :

(3) Le **problème** de l'abandon de l'habitat dans la Corse médiévale

(4) Le **problème** du Paléolithique final de Haute-Normandie

*Problème* est un terme listé comme employé très fréquemment dans un emploi sous-spécifié par Flowerdew et Forest (2015) et Schmid (2000, p. 85). Selon Schmid (2018, p.118), ce qui unit les contenus désignés comme un problème est qu’il s’agit d’un *« fait étant un obstacle au progrès »* ou, citant Tuggy dans ce même article (2018, p. 122), *« une chose qui n’est pas en conformité avec quelque chose établi ou désiré »*. On peut rajouter à ces définitions, une chose qui a des conséquences négatives. Ainsi est catégorisé à chaque fois un concept temporaire créé par l’énoncé : l’abandon de l’habitat dans la Corse médiévale pour (3) et le Paléolithique final de Haute-Normandie pour (4). Le choix de catégoriser ce concept de *problème*, au lieu de *question* par exemple, indique une volonté de l’interlocuteur de souligner qu’il y a un obstacle ou du moins un imprévu dans le raisonnement scientifique. On peut également voir que *problème* crée une liaison à son contenu spécificationnel dès le titre. Il pourra également créer des références anaphoriques en étant repris avec un démonstratif, *ce problème*, si ce n’est dans le titre du fait sa trop grande concision, dans le résumé ou le texte de la publication scientifique[[2]](#footnote-3).

Outre les trois CS déjà mentionnées, il est également possible que les têtes s’intègrent à d’autres schémas d’utilisation très fréquents qui pourraient jouer le rôle de CS.

De précédents travaux ont montré qu’il existe des spécificités disciplinaires dans l’écriture des titres pour l’anglais (Anthony, 2001 ; Haggan, 2004 ; Lewison et Hartley, 2005 ; Soler, 2007, 2011 ; Nagano, 2015) et le français (Tanguy et Rebeyrolle, à paraître). Nous ne manquerons pas de déterminer dans le cadre de notre problématique s’il existe des variations des têtes et des schémas suivant les domaines. Nous pourrions ainsi mettre au jour des têtes spécifiques à certains domaines et d’autres qui seraient transdisciplinaires.

Nous voulons déterminer la proximité de fonctionnement entre têtes transdisciplinaires et NSS en identifiant :

* Une liste de têtes transdisciplinaires à rapprocher des NSS.
* Une liste de schémas récurrents dans lesquels s’inscrivent nos têtes transdisciplinaires à rapprocher des constructions spécificationnelles dans lesquelles les NSS s’inscrivent.
* Une répartition des têtes transdisciplinaires et des schémas par rapport aux domaines scientifiques.

Pour ce programme, nous utiliserons une approche se fondant sur le traitement automatique des langues et la linguistique de corpus (Biber et al., 1998 ; Cori et David, 2008 ; Williams, 2005). Pour les NSS, nous nous appuyons plus particulièrement sur Legallois (2008), sur la perspective constructionnelle et cognitiviste de Schmid (2000) et celle, discursive et constructionnelle, de Flowerdew et Forest (2015) (voir Adler et Moline, 2018 et Schmid, 2018, pour une comparaison des divergences et des points communs entre les deux approches).

Notre étude se déroulera en trois temps. Dans un premier temps, nous partons des données rassemblées pour délimiter un corpus de travail. Nous décrivons certains traits saillants de notre corpus à l’aide de différentes mesures, en faisant référence aux nombreux travaux existants. Notre démarche nous amènera à revérifier la nature éminemment nominale des titres. Nous étudierons les variations entre les différents domaines scientifiques, notamment les têtes de segments spécifiques à certains domaines. Dans un deuxième temps, nous construirons la liste des têtes de segments transdisciplinaires. Nous rappellerons les apports des travaux sur les noms sous-spécifiés et essayerons de détecter les constructions spécificationnelles dans lesquelles ils s’inscrivent généralement dans notre corpus. Nous détecterons ensuite les schémas récurrents dans lesquels s’inscrivent nos têtes transdisciplinaires pour essayer de rapprocher les schémas des constructions spécificationnelles et les têtes transdisciplinaires des noms sous-spécifiés. Nous nous appuierons notamment sur la forte fréquence et la transdisciplinarité des têtes transdisciplinaires pour fournir une liste de têtes transdisciplinaires se comportant comme des NSS. Enfin, dans un troisième temps, nous discuterons de nos résultats, des limites de notre travail et ouvrirons de nouvelles perspectives.

# I. Exploration du corpus à la lumière de l’état de l’art

## I.1 Origine et prétraitements des données

### I.1.1 Récupération des données

L’accès aux titres a été grandement facilité par la création de bases de données bibliographiques, dont celles des archives ouvertes. Les chercheurs ou les documentalistes d’un centre de recherche sont libres de déposer un document sur une archive ouverte avec l’accord de ses auteurs. Une archive ouverte présente l’avantage de centraliser l’accès aux travaux scientifiques, d’aider à leur diffusion et de les conserver de manière pérenne, par rapport au site d’une institution particulière ou le site web personnel d’un chercheur, et de façon gratuite et accessible à tous, au contraire des éditeurs.

Nous utilisons le corpus constitué par Tanguy et Rebeyrolle (à paraître) comprenant près de 340  000 titres. Pour obtenir une si grande quantité de titres français, ils se sont tournés vers l’archive ouverte Hyper Article en Ligne (HAL, <https://hal.archives-ouvertes.fr>) (Nivard, 2010). Cette archive fonctionne depuis 2001 et est gérée par le Centre pour la Communication Scientifique directe du Centre National pour la Recherche Scientifique (CNRS). Elle contient plus de 1,6 millions de références, qui sont soit des travaux dont elle possède une copie, soit des notices. HAL possède de nombreux types de documents différents : articles scientifiques mais aussi vidéo, cours, ouvrages ou thèses. Plusieurs institutions, dont le CNRS, encouragent le dépôt sur HAL des travaux produits par leurs chercheurs, garantissant un nombre important de titres issus de plusieurs domaines scientifiques. Alors que la majorité de la littérature traite des titres en anglais, HAL permet d’avoir accès à un grand corpus de titres en français. Nous veillerons dans ce premier chapitre à vérifier sur notre corpus certains enseignements tirés de l’étude des titres en anglais, notamment la nature des titres.

Notre matière de départ se restreint aux titres en français, d’articles scientifiques, de chapitre, de poster ou de communication, car nous prenons comme hypothèse qu’ils sont construits de manière similaire. Chaque titre est fourni avec cinq informations supplémentaires relatives à la publication titrée :

1. un **identifiant** unique de la publication et donc du titre ;
2. les prénoms et noms des **auteurs** de la publication dont on peut déduire le nombre d’auteurs ;
3. le **type** du document qui ne peut être qu’un article scientifique, un chapitre d’un ouvrage collectif, une communication ou un poster dans un congrès ou une conférence ;
4. l’**année** de publication ;
5. les **domaines scientifiques**, ou disciplines académiques, auxquels est associée la publication dont nous déduisons un domaine principal selon la méthode établie par Tanguy et Rebeyrolle (à paraître).

L’exemple (5) ci-dessous montre les différentes informations pour un titre donné :

(5) Villes durables et changement climatique : quelques enjeux sur le renouvellement des ressources urbaines  
 **Identifiant** 609897  
 **Auteurs** Véronique Peyrache-Gadeau et Bernard Pecqueur  
 **Type de document** Article scientifique (code ART)  
 **Année de publication**  2011  
 **Domaines scientifiques** 0.sde et 1.sde.mcg, le premier correspond aux sciences de   
 l’environnement et le second à un sous-domaine des sciences de l’environnement.

HAL permet d’attribuer plusieurs domaines à un document. Les domaines sont organisés en une taxonomie possédant quatre niveaux de profondeur. Néanmoins la granularité des branches est très variable. Sciences de l'Homme et Société est une des racines de l’arbre, regroupant sous son égide de nombreux domaines scientifiques, allant de l’histoire aux littératures, alors que toutes les sciences exactes bénéficient quant à elles d’une racine propre comme informatique ou chimie. Tanguy et Rebeyrolle (à paraître) ont proposé une méthode de recodage des domaines pour n’en garder qu’un seul, le plus important et discriminant, que nous utilisons. Dorénavant, un titre est associé à un seul domaine principal : le domaine de premier niveau pour les sciences exactes, le domaine de second niveau pour les sciences humaines et sociales.

Nous avons relevé les domaines suivants, avec en gras les sciences exactes (voir pour l’annexe **A4.2 Code des 27 domaines de HAL retenues** pour une correspondance entre les codes et les domaines scientifiques) : anthropologie, archéologie et préhistoire, architecture, art et histoire de l’art, autres, **chimie**, droit, **économie et finance quantitative**, éducation, géographie, gestion et management, histoire, **informatique**, linguistique, littératures, **mathématiques**, philosophie, **physique**, **planète et univers**, psychologie, science politique, **sciences cognitives**, **sciences de l’environnement**, sciences de l’information et de la communication, **sciences du vivant** et sociologie.

### I.1.2 Étiquetage et analyse syntaxique en dépendances

Les titres ont été analysés à l’aide du logiciel Talismane (Urieli et Tanguy, 2013 ; Urieli, 2013) qui fournit un découpage en différents tokens, mots et signes de ponctuation, et réalise un étiquetage morphosyntaxique des mots et une analyse syntaxique en dépendances des tokens. Pour chaque token du titre nous avons :

* sa **forme** dans le titre ;
* son **lemme** (pour les mots), ;
* sa **classe grammaticale/catégorie** (pour les mots, sinon nous avons “signe de ponctuation”) ;
* des **informations complémentaires**;
* son token **recteur**;
* la **relation de dépendance** qui le lie à son recteur.

Les informations complémentaires dépendent de la classe grammaticale, comme le genre pour les noms, le mode et le temps pour les verbes. Les titres étant des textes très travaillés, ils ne nécessitent pas de prétraitement pour corriger les fautes, même s’il y en a de très rares comme la concaténation d’un titre et d’un sous-titre sans token séparateur (6) ou le redoublement d’une préposition (7) :

(6) Développement stratégique du tourisme sportif de rivière par régulation corporatiste **L'**expérience du bassin de Saint Anne (Québec) appliquée aux Rivières de Provence

(7) Dispositif **de de** caractérisation simultanée de l'abondance de pucerons et de la croissance végétative d'arbres fruitiers

Il est à noter que Talismane a été conçu pour analyser des phrases beaucoup plus longues que des titres et entraîné sur de tels textes. On peut peut-être douter de sa capacité à analyser correctement les titres. Notamment, comme nous le verrons plus tard, les titres ne comportent souvent pas de verbes conjugués au contraire des phrases plus longues, ce qui pourrait pousser Talismane à reconnaître comme verbes des mots n’en étant pas. Nous avons donc décidé d’inclure une phrase de vérification de l’analyse de Talismane lors de l’étape de sélection des têtes pour vérifier son comportement.

### I.1.3 Segmentation des titres

Nous avons segmenté les titres selon la liste des signes de ponctuation segmentants établie par Anthony (2001). Nous en retranchons le tiret car il est utilisé pour lier de nombreux mots en français comme *e-commerce, semi-figement* ou *petit-déjeuner*. Nous avons pu vérifier que Talismane traitait les formes en *e-X* et *semi-X* comme un *e* ou *semi* suivi d’un tiret suivi d’un nom (voir la section **I.3.2.B) Corrections de Talismane**). Nous y ajoutons le point d’exclamation et les points de suspension dont l’absence ne nous semble pas justifiée. Nous avons donc les signes segmentants suivants :

|  |  |
| --- | --- |
| **Type de ponctuation** | **Signe de ponctuation** |
| **Ponctuation forte** | . ? ! ... |
| **Ponctuation faible** | ; : |

Tableau 1: signes de ponctuation segmentants

Il y a dans cette liste des signes de ponctuation forte, comme le point ou le point d’interrogation, et des signes de ponctuation faible comme le point-virgule ou le double-point. Le type de segmentation effectué découle directement du type de ponctuation : forte ou faible.

L’avantage d’utiliser le segment est qu’il s’agit d’une unité que nous définissons clairement à la suite d’Anthony (2001), directement applicable computationnellement, au contraire de la proposition dont la définition est selon Joseph Donato dans l’ouvrage collectif sous la direction de Mounin (1974, p. 273) *« très empirique »* et pour la laquelle la *« distinction entre syntagme et proposition n’était pas toujours très claire ni très systématique dans l’analyse des phrases spécifiques »*.

Dans notre matériau de départ, nous comptons 221 674 occurrences de signes de ponctuation segmentants. Le signe le plus fréquent est le double-point, qui compte pour 47 % de ces occurrences, suivi du point pour 36 %, du point d’interrogation pour 13 %, et du point-virgule pour près de 2 %. Ces quatre signes comptent pour 98 % des occurrences. Ces chiffres témoignent de l’importance du double point dans les titres, qui est au centre de nombreux travaux (Diers et Downs, 1994 ; Dillon, 1981, 1982 ; Townsend, 1983) et de notre travail de première année.

### I.1.4 Sélection de la tête des segments

Nous voulons ensuite récupérer la tête des segments, qui s’assimile à la notion de prédicat suivant la définition de Conrad Bureau, toujours dans Mounin (1974, p. 267) :

*« Désigne, en syntaxe, l'élément central de la phrase, celui par rapport auquel tous les autres éléments de la phrase marquent leur fonction. Est prédicat celui des éléments : 1° qui ne dépend syntaxiquement d'aucun autre élément ; 2° par rapport auquel la phrase s'organise, et 3° dont la disparition détruit l'énoncé. »*

Pour trouver les têtes et les compter, deux solutions s’offraient à nous. La première est une règle qui consiste à prendre le verbe conjugué du segment comme tête s’il y en a un, sinon une préposition si elle occupe la première position du segment et sinon le premier nom rencontré. Cette solution présente l’avantage d’être très simple mais nous avions peur de manquer des phénomènes remarquables ou de sélectionner le mauvais mot comme tête en nous fondant si fortement sur la position.

Nous avons donc opté pour la seconde solution qui consiste à utiliser l’outil Talismane pour effectuer une analyse syntaxique en dépendances du titre. Il s’agit d’une utilisation a minima de l’analyse en dépendances, uniquement pour faire émerger une tête, mais nous avons rencontré deux difficultés.

Notre but est que chaque segment ait une tête correctement identifiée mais la segmentation que nous effectuons, fondée sur des signes de ponctuation, est décorrélée de l’analyse de Talismane qui possède sa propre segmentation, que nous nommerons partition et le résultat des parties pour les distinguer de nos segments. Talismane va produire pour chaque partie un arbre avec une racine unique. Dans le cas nominal, chaque partie de Talismane correspond à un segment, et la tête de chaque segment est directement la racine de l’arbre produit par Talismane.

Mais si notre titre est constitué d’une seule partie elle-même composée de plusieurs segments, nous obtenons des segments sans tête. Nous avons décidé de nous limiter aux titres avec au maximum deux parties et deux segments car ils sont les plus nombreux dans notre matériau : nous comptons 87 % de titres avec une partie et 11 % avec deux et 58 % titres avec un segment et 37 % avec deux. On peut classer nos résultats d’analyse en trois cas :

1. Des titres ayant un segment et une tête ;
2. Des titres ayant deux segments dont un seul a une tête (soit le premier, soit le second) ;
3. Des titres ayant deux segments avec une tête dans chaque.

L’exemple (8) montre un titre à deux segments avec une segmentation faible, le double-point et l’exemple (9) montre un titre à deux segments avec une segmentation forte, le point. Les deux exemples ont pour Talismane une seule partie.

(8) L'**omniprésence** de la famille au sein de l'exploitation agricole : une *situation* de fait encouragé par les règles de droit

(9) **MODÈLES** THÉOTIQUES DE LA STRUCTURE DES JOINTS DE GRAINS.LES *MODÈLES* DE STRUCTURE DES JOINTS DE GRAINS ET LEUR UTILISATION[[3]](#footnote-4)

Dans les deux exemples précédents, *omniprésence* et *modèles* (en gras) sont reconnus comme des têtes des premiers segments mais pas *situation* et *modèles* (en italique) pour les seconds segments. Nous utilisons Talismane comme une « boîte noire » et nous ne voulons pas entrer dans les détails de sa partition des titres et de son analyse. Nous voulons néanmoins prendre en compte les spécificités des résultats donnés pour mieux les exploiter dans la perspective de notre travail : trouver des têtes aux différents segments d’un titre.

Avant d’aborder notre méthode pour résoudre le premier problème des segments sans tête, nous devons présenter le second problème de notre approche. La fiabilité de Talismane n’étant pas assurée sur des énoncés courts et généralement averbaux comme des titres, nous avons décidé d’estimer sa fiabilité. Nous avons choisi un échantillon de 20 titres aléatoirement pour chaque structure, en différenciant le cas numéro deux selon que le segment sans tête est le premier et le second. Nous avons également choisi 20 titres ayant un segment et deux têtes pour observer cet ensemble et éventuellement tenter d’en reprendre des titres. Nous avons vérifié manuellement pour ces 100 titres le choix de la tête, sa catégorisation morphosyntaxique et son lemme. Les résultats complets sont dans l’annexe **A5.C Analyse de 100 titres traités par Talismane**. Si globalement, Talismane arrive à étiqueter morphosyntaxiquement et à trouver le lemme correctement dans des énoncés aussi courts que des titres, la fiabilité pour sélectionner les têtes diffère grandement selon la structure segments-têtes.

Avant d’aborder les résultats structure par structure, nous voulons aborder un premier point : Talismane ne catégorise comme type de dépendance racine, « root » dans sa nomenclature, que les verbes. Pour les autres catégories, il reconnaît que la tête est le token racine de l’arbre de l’analyse en dépendances mais sans qualifier sa relation de dépendance de racine : il indique « \_ » au lieu de « root ».

Un second point concerne les segments sans racine dans les titres bisegmentaux. On observe dans ces segments sans racine un mot qui est uniquement régi par un mot de l’autre segment. D’après nos analyses manuelles, ce mot est le plus souvent la tête de l’autre segment. Nous avons donc développé un algorithme de sélection des têtes pour suppléer les déficiences de Talismane tout en gardant le bénéfice de l’analyse syntaxique en dépendances. Notre algorithme est présenté en détail après les résultats.

#### A. Titres avec un segment et une tête

Sur les 20 titres pris, Talismane a à chaque fois détecté la bonne tête, avec la bonne catégorie morphosyntaxique et le bon lemme, sauf une fois, où l'absence d'un accent ne lui a pas permis de retrouver le lemme à partir de la forme. On peut donc estimer que les titres qui suivent cette structure sont correctement analysés par Talismane.

#### B. Titres avec un segment et deux têtes

Sur les 20 titres pris, Talismane a analysé incorrectement 12 titres et huit ont une analyse discutable. Nous ne considérons pas le tiret et la virgule comme des caractères segmentants alors qu'ils sont clairement utilisés comme tels par un titre pour le tiret et deux titres pour la virgule. De plus, les mots composés provoquent des erreurs d'analyse dans Talismane qui désigne comme tête la partie après le tiret. Enfin, on remarque un oubli de signe de ponctuation segmentant et un crochet droit utilisé comme signe de ponctuation segmentant qui entraînent à chaque fois une mauvaise analyse.

Nous pourrions changer notre liste de caractères segmentants, mais cela reviendrait à créer potentiellement de nouvelles erreurs. Nous décidons donc de ne pas utiliser les titres ayant deux têtes dans un seul segment.

#### C. Titres avec un segment ayant une tête suivie d’un segment sans tête

Sur les 20 titres, notre algorithme permet de sélectionner une tête valide dans le segment n'en contenant pas pour 17 d'entre eux. Deux titres utilisent la virgule comme un caractère segmentant. Enfin un dernier échappe à notre algorithme de sélection d’un mot pour sa promotion en tête de segment.

#### D. Titres avec un segment sans tête suivi d’un segment avec tête

Sur les 20 titres, notre algorithme permet de sélectionner une tête valide dans le segment n'en contenant pas pour 18 d'entre eux. On note des erreurs d'analyse de Talismane liées à une mauvaise catégorisation morphosyntaxique de mots dont cinq entraînent une mauvaise sélection de la tête.

#### E. Titres avec un segment avec tête suivi d’un segment avec tête

Sur les 20 titres, 16 sont correctement analysés par Talismane qui trouve les têtes des segments. Pour trois titres la tête est mal catégorisée et pour un dernier le lemme n'est pas trouvé.

#### F. Algorithme de sélection de tête de segment

Notre algorithme pour détecter la tête d’un segment à partir du résultat de l’analyse de Talismane est le suivant :

* Soit un mot du segment sans tête est régi par la tête de l’autre 🡺 promotion de ce mot comme tête. 46 798 titres ont une tête sélectionnée de cette façon.
* Soit le premier mot du segment sans tête est régi par un mot de l’autre segment 🡺 promotion de ce mot comme tête. 8 866 titres ont une tête sélectionnée ainsi.

Nous récupérons en tout 55 664 titres, soit 98 % des 56 851 titres ayant deux segments mais une seule tête. Ces titres problématiques comptent pour 18 % de l’ensemble des titres à un ou deux segments. Cela nous permet de récupérer plus de titres valides selon notre définition qu’il doit y avoir une tête par segment et au maximum deux segments par titre.

Une fois les données récupérées et prétraitées, nous constituons notre corpus de travail. Il faut pour cela établir un périmètre qui délimitera notre corpus de travail. Il faut expliquer le choix de notre périmètre et effectuer des mesures dessus, afin de mettre en relation notre corpus de travail avec ceux étudiés précédemment dans la littérature.

## I.2 Constitution d’un corpus de travail représentatif

Un périmètre de recherche établit dans le matériau de base une dichotomie claire entre ce que nous allons étudier et ce que nous n’étudierons pas. Plus un corpus est grand, plus la confrontation d’une hypothèse à son contenu aura de validité dans une approche corpus-based. Plus un corpus est grand, plus la formulation d’une hypothèse, à partir de la généralisation des faits observés dans le corpus, aura du poids dans une approche corpus-driven (Anthony, 2001 ; Biber, 2009).

Mais plus un corpus est large, plus nous risquons de nous confronter à des hapax remettant en cause confirmations et infirmations, ou rendant l’établissement de celles-ci beaucoup plus difficile. Nous pensons que, pour notre travail, le juste milieu est d’essayer de prendre le maximum de matériel tout en écartant les cas les plus rares, suivant ainsi le principe de *From-Corpus-To-Cognition* de Schmid (2000, p. 47) qui est que *« depsite the indisputable charm of rare or exotic examples, one should mainly be interested in frequent and therefore systemically and cognitively more important items »*. Notre périmètre sera constitué sur deux points : la structure segmentale des titres et la nature des têtes.

### I.2.1 Sélection selon la structure des titres

Nous avons décidé de prendre les titres composés de seulement un ou deux segments. Nous justifions ce choix par le fait qu’il s’agit de la plus grande majorité des titres (320 561 soit 94 % des titres initiaux) et qu’ils sont plus faciles à analyser. De nombreux travaux didactiques sur l’écriture des titres (Aleixandre-Benavent et al., 2014 ; Swales et Feak, 1994 ; Gustavii, 2008) conseillent d’ailleurs d’organiser les titres en deux segments autour d’un double point soit la forme *segment 1: segment 2*.

Un autre délimiteur que nous utilisons pour établir notre périmètre, en plus du nombre de segments dans le titre, est le nombre de têtes par segments. Nous nous limiterons aux titres avec au maximum une tête par segment. L’analyse en dépendance ne devant produire qu’une tête par segment, on peut considérer les segments bicéphales comme une aberration à écarter, souvent due au fait qu’une segmentation n’a pas été détectée, comme le montre l’étude des titres dans **A5.C Analyse de 100 titres traités par Talismane**. On distingue donc deux cas : les titres composé d’un seul segment avec une tête et les titres composés de deux segments avec une tête chacun.

Il y a 171 890 titres composés d’un seul segment ayant une seule tête de segment, soit près de 51 % des données initiales, comme les exemples (10) et (11). Il y a 124 938 titres composés de deux segments, soit près de 37 % des données initiales, comme les exemples (12), (13) et (14). Nous indiquons entre indice la catégorie morphosyntaxique du lemme.

(10) L'**actualité** nom de la jurisprudence communautaire et internationale

(11) **Doit** verbe -on écouter Björk ?

(12) Un nouvel **OVNI** nom dans le ciel réunionnais : la **transparence** nom des prix

(13) La **performativité** nom de l'évidence : **analyse** nom du discours néolibéral

(14) **Traces** nom de contenus africains sur Internet : **entre** préposition homogénéité et identité

Du fait des limites entre les capacités de Talismane et notre définition des segments, certains segments n’ont pas de tête. Nous avons appliqué notre algorithme pour suppléer ces limitations.

Pour finir, nous gardons 110 785 titres composés de deux segments avec une tête dans chaque. Nous avons donc 171 890 titres monosegmentaux (61 %), 110 785 bisegmentaux (39 %), soit un corpus de travail de 282 675 titres, ce qui représente 83 % du matériau initial, les 340 000 titres collectés sur HAL. Nous avons réussi à conserver 83 % du matériau initial dans cette première étape de définition du périmètre de notre corpus de travail, néanmoins nous restreignons encore notre périmètre dans l’étape suivante pour nous intéresser à une catégorie morphosyntaxique particulière.

### I.2.2 Sélection selon la nature des têtes

#### A) Répartition des natures des têtes

Nous nous sommes interrogés sur la nature de la tête des segments pour opérer une sélection sur ce critère. Cette question est directement liée à la question de la nature des titres. D’après Schwischay (2001, p. 11), *« un nœud forme avec tous les nœuds qu’il domine (directement ou indirectement) un syntagme ; et, par convention, ce syntagme porte le nom du nœud dominant »*. Nous pouvons donc, grâce à la complémentarité du modèle de l’analyse en constituants immédiats et celui de l’analyse en dépendances, déterminer le type de syntagme de chaque segment en étudiant la catégorie morphosyntaxique de sa tête à l’aide du tableau (2). La dernière colonne indique ces valeurs sur tous les segments des titres, soit 354 168 segments, en considérant les segments des titres bisegmentaux de façon indépendante.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Catégorie morphosyntaxique de la tête du segment** | **Titres monosegmentaux** | **Titres bisegmentaux, segment 1** | **Titres bisegmentaux, segment 2** | **Sur tous les**  **segments**  **(354168 segments)** |
| **Noms communs** | 136 734 | 80 % | 82 959 | 75 % | 84 960 | 77 % | 304 653 | 86 % |
| **Noms propres** | 11 094 | 6 % | 10 406 | 9 % | 4 758 | 4 % | 26 258 | 7 % |
| **Σ Tous les noms** | **147 828 | 86 %** | **93 365 | 84 %** | **89 718 | 81 %** | **330 911 | 93 %** |
| **Verbes à l’indicatif** | 8 186 | 5 % | 3 478 | 3 % | 3 513 | 3 % | 15 177 | 4 % |
| **Verbes à l’infinitif** | 5 135 | 3 % | 6 004 | 5 % | 2 140 | 2 % | 13 279 | 4 % |
| **Σ Tous les verbes** | **15 749 | 9 %** | **10 672 | 10 %** | **6 549 | 6 %** | **32 970 | 9 %** |
| **Prépositions** | **6 792 | 4 %** | **5 456 | 5 %** | **10 456 | 9 %** | **22 704 | 6 %** |

Tableau 2: Distribution des catégories morphosyntaxiques des têtes de segments

On peut remarquer des points communs : la grande majorité des têtes sont des noms, et a fortiori des noms communs, pour toutes les configurations segmentales. Les autres catégories les plus représentées sont les verbes à l’indicatif ou à l’infinitif et les prépositions. La différence la plus notable entre les premiers et seconds segments des titres bisegmentaux est que pour les seconds segments, la seconde catégorie la plus fréquente sont les prépositions et non les verbes : les têtes prépositionnelles sont presque deux fois plus fréquentes (9 %) que dans les segments des titres monosegmentaux (4 %) et dans les premiers segments des titres bisegmentaux (5 %).

On peut ensuite s’interroger sur les combinaisons possibles dans les titres bisegmentaux entre les catégories des deux têtes de segments. Nous agrégeons les différentes catégories nominales, verbales et prépositionnelles en trois catégories : Nom, Verbe et Préposition. Le tableau (3) présente les cinq combinaisons les plus fréquentes, sur 96 en tout. L’annexe **A2. Combinaisons des têtes de titres bisegmentaux** liste l’ensemble des 96 combinaisons existantes. Les cinq combinaisons les plus fréquentes couvrent 93 % des titres bisegmentaux. On constate là-aussi que la grande majorité des titres bisegmentaux ont à chaque fois un nom pour tête de segment.

|  |  |  |
| --- | --- | --- |
| **Catégorie de la tête du premier segment** | **Catégorie de la tête du second segment** | **Nombre de titres et pourcentage** |
| Nom | Nom | 75 592 ( 68 % ) |
| Nom | Préposition | 8 996 ( 8 % ) |
| Verbe | Nom | 8 506 ( 8 % ) |
| Nom | Verbe | 5 426 ( 5 % ) |
| Préposition | Nom | 4 650 ( 4 % ) |

Tableau 3 : Combinaisons agrégées les plus fréquentes de têtes dans les titres bisegmentaux

Notons qu’il existe 409 titres dont le premier et le second segment ont le même lemme pour tête. Ce qui vient à l’esprit en regardant les exemples de (15) à (20), c’est la possibilité d’achever un effet stylistique de répétition et la possibilité d’introduire une comparaison ou un questionnement :

(15) La **crise** ? Quelle **crise** ?

(16) **Crise** du logement ? Quelle **crise** ?

(17) **Ville** de jour. **Ville** de nuit

(18) **Linux** embarqué. **Linux** Temps Réel

(19) **Feu** l'arrêt Mercier ! **Feu** l'arrêt Mercier ?

(20) **Corps** dansant. **Corps** glorieux

Cette répétition est à rapprocher de l’usage dans les titres de la figure de style du chiasme détectée par Tanguy et Rebeyrolle (à paraître, 2.2) où deux mots sont repris dans l’ordre inverse. Ici, la reprise ne joue pas sur l’ordre inversée mais sur la détermination différente dans (15) et (16), la complémentation par un groupe nominal dont les noyaux sont antonymes (17), le changement de point final (19), ou simplement une différence dans l’expansion du nom (18, 20).

#### B) La nature nominale des titres

Chercher la nature d’un titre revient à s’interroger sur la nature de ses têtes de segments. Pour les titres monosegmentaux, déterminer la nature du titre revient à prendre la nature de son unique segment. On obtient à partir du tableau (2) directement 86 % de titres nominaux. Pour les titres bisegmentaux, on peut considérer deux options. La première est qu’un titre est nominal si son premier segment l’est. On obtient alors 84 % de titres nominaux. L’autre option est de considérer qu’un titre est “purement” nominal si et seulement si les deux têtes de ses segments sont des noms. On obtient alors 68 % de titres nominaux.

Quelle que soit la solution choisie, les titres sont majoritairement constitués d’un ou plusieurs syntagmes nominaux et non d’une phrase avec un noyau verbal, ce qui rejoint les conclusions de nos prédécesseurs (Leech, 2000 ; Haggan, 2004 ; Soler, 2007 ; Cheng et al., 2012 ; Wang et Bai, 2007). Cheng et al. (2012) relèvent jusqu’à 93 % de titres nominaux pour leur corpus et Wang et Bai (2007) 99 % pour le leur. Jalilifar et al. (2010) résument les raisons de cette predominance, qui s’applique aussi bien aux titres monosegmentaux (Jalilifar, 2010, p. 45) que bisegmentaux (Jalilifar, 2010, p. 47) : *« the ability to compact information in an economical way through various pre and postmodifiers (Wang et Bai, 2007) makes noun phrase titles more informative and explanatory than other structures (Yakhontova, 2002).*

Pour notre corpus de travail, nous décidons de nous restreindre aux titres monosegmentaux dont la tête est un nom et aux titres bisegmentaux dont au moins une des têtes de ses segments est un nom, l’autre pouvant être un nom, une préposition ou un verbe. Ce choix nous permet de garder la grande majorité de nos titres et d’éliminer les cas les moins fréquents. Nous obtenons un corpus de 250 998 titres, soit 74 % du matériau initial, ce qui nous semblait important pour renforcer nos hypothèses en les établissant sur le plus grand nombre possible de faits linguistiques.

Une fois le périmètre des titres étudiés défini sur la structure segmentale des titres et la nature grammaticale de leurs têtes, nous avons constitué notre corpus de travail. Nous pouvons alors effectuer plusieurs mesures sur notre corpus et les mettre en rapport avec les mêmes mesures effectuées dans des travaux précédents, avant d’étudier plus avant les têtes de syntagmes.

### I.2.3 Un corpus de travail représentatif du matériau de base

Nous avons défini notre périmètre d’étude comme portant sur 250 998 titres constitués d’un ou deux segments. Les titres monosegmentaux (147 828 soit 59 %) ont une tête nominale, les titres bisegmentaux (103 170, 41 %) ont au moins un segment ayant une tête nominale, l’autre ayant une tête verbale, nominale ou prépositionnelle.

On notera que les différentes caractéristiques des titres ne sont pas indépendantes : Kutch (1978), Yitzhaki (1994) et Tanguy et Rebeyrolle (à paraître) ont ainsi montré que le nombre d’auteurs est corrélé positivement à la longueur du titre. Larivière et al. (2015) ont montré que le domaine est lié au nombre d’auteurs : il y a en moyenne plus d’auteurs dans les sciences exactes. Baethge (2008) a montré que le nombre d’auteurs augmente avec le temps. Tanguy et Rebeyrolle (à paraître) ont également montré, en partant des mêmes données de base et donc avec le même déséquilibre de répartition, que la longueur était très légèrement corrélée à l’année de publication.

Sur la longueur des titres, les titres monosegmentaux ont une longueur moyenne de 10,38 mots, avec une longueur minimale de 1 mot et une longueur maximale de 77 mots, tandis que les titres bisegmentaux ont une longueur moyenne de 14,45 mots, avec une longueur minimale de 2 mots et une longueur maximale de 228 mots. Les titres bisegmentaux les plus courts sont au nombre de 64. Quarante-neuf utilisent comme signe segmentateur le double point et 51 sont des chapitres d’ouvrage dont 29 sont de la forme *Entrée : NC*, indiquant une entrée dans un ouvrage de type dictionnaire ou encyclopédie. La longueur supérieure des titres bisegmentaux s’explique par la facilité de traitement qu’apporte la segmentation à l’interlocuteur : la segmentation sert à la fois de pause et d’articulation pour sa compréhension. La longueur moyenne des titres du corpus de travail est de 12,05 mots, alors que celle des données de départ est de 13,8 mots. Cette constatation est normale car il existe des titres ayant plus de deux segments que notre corpus de travail n’inclut pas.

On peut regarder comment nos corpus se répartissent en fonction du type de publication scientifique :

|  |  |  |  |
| --- | --- | --- | --- |
| **Type de publication** | **Titres monoseg.** | **Titres biseg.** | **Corpus** |
| Article | 63 993 43 % | 45 827 44 % | 109 820 44 % |
| Communication | 53 148 36 % | 35 350 34 % | 88 498 35 % |
| Chapitre d’ouvrage | 29 413 20 % | 21 221 21 % | 50 634 20 % |
| Poster | 1 274 1 % | 772 1 % | 2 046 1 % |

Tableau 4 : Distribution des structures des titres selon le type

La structure des titres n’est pas corrélée au type de publication, la distribution des deux ensembles étant presque identique. De plus, cette répartition est quasi identique à celle de l’ensemble des 340 000 titres qui constituent nos données de départ (Tanguy et Rebeyrolle, à paraître).

On peut aussi mesurer le nombre d’auteurs en fonction de la structure du titre :

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre d’auteurs** | **Titres monoseg.** | **Titres biseg.** | **Corpus** |
| 1 | 87 646 59 % | 65 199 63 % | 152 845 61 % |
| 1-4 | 135 564 92 % | 96 581 94 % | 232 145 92 % |
| 1-9 | 146 767 99 % | 102 307 99 % | 249 074 99 % |

Tableau 5 : Distribution des structures des titres selon le nombre d'auteur

On voit bien que quelle que soit la structure du titre, la répartition par le nombre d’auteurs est la même pour les deux sous-ensembles de notre corpus de travail que pour le corpus de travail pris dans sa totalité et sur l’ensemble des données où 62 % des articles avaient également un seul auteur.

On regarde également la répartition par années de publication. Pour l’ensemble du corpus, elles s’étendent de 2019 pour les sept publications les plus récentes à 1779 pour la plus ancienne. On note que 85 % des publications ont été publiées en 2000 ou après, 90 % après 1994 et 99 % après 1933. Pour l’ensemble des données, Tanguy et Rebeyrolle (à paraître) trouvent les mêmes années pour les deux premiers pourcentages et un peu plus tard, 1940, pour le dernier. Notre corpus ne peut donc pas servir pour des études diachroniques du fait de sa répartition totalement inégale sur le temps. La période qui comporte le plus de titres, de 2005 à 2017, soit 74 % du corpus, est également trop courte. La répartition est similaire pour nos deux sous-corpus, titres monosegmentaux et bisegmentaux. Nous pouvons à présent étudier comment la structure segmentaire des titres et les têtes varient selon les domaines.

## I.3 Structures semblables des domaines et têtes spécifiques

### I.3.1 Variations de la structure en fonction du domaine

Nous regardons à présent la répartition des titres par domaine pour le corpus et les deux sous-corpus. Nous rappelons que nous avons sélectionné, grâce à la méthode décrite dans Tanguy et Rebeyrolle (à paraître), un seul domaine principal pour chaque titre. Le tableau (6) présente les 27 domaines qui existent dans notre corpus. Nous avons mis en gras les domaines des sciences exactes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N°** | **Domaine** | **Corpus**  **Nb/fréq/fréq. cumul** | **Répartition entre** | |
| **Titres monosegmentaux** | **Titres bisegmentaux** |
| **01** | **Physique** | 26 559 11% 11% | **81 %** | 19 % |
| 02 | Sociologie | 23 732 9% 20% | 48 % | 52 % |
| 03 | Droit | 21 486 9% 29% | 67 % | 33 % |
| 04 | Histoire | 19 093 8% 36% | 54 % | 46 % |
| 05 | Pas de domaine associé | 18 941 8% 44% | 59 % | 41 % |
| 06 | Gestion et management | 18 318 7% 51% | 45 % | 55 % |
| **07** | **Sciences du vivant** | 17 498 7% 58% | **66 %** | 34 % |
| **08** | **Informatique** | 13 505 5% 63% | **74 %** | 26 % |
| 09 | Linguistique | 11 556 5% 68% | 52 % | 48 % |
| 10 | Littératures | 10 712 4% 72% | 52 % | 48 % |
| 11 | Archéologie et Préhistoire | 10 124 4% 76% | 61 % | 39 % |
| 12 | Science politique | 7 152 3% 79% | 46 % | 54 % |
| 13 | Éducation | 7 062 3% 82% | 50 % | 50 % |
| 14 | Art et histoire de l'art | 6 471 3% 85% | 53 % | 47 % |
| 15 | Philosophie | 6 152 2% 87% | 60 % | 40 % |
| 16 | **Sciences de l'environnement** | 5 542 2% 89% | 54 % | 46 % |
| 17 | Sciences de l'information et de la communication | 5 481 2% 91% | 46 % | 54 % |
| 18 | Anthropologie | 5 166 2% 93% | 51 % | 49 % |
| 19 | Architecture | 3 444 1% 95% | 51 % | 49 % |
| 20 | **Planète et Univers** | 2 781 1% 96% | **62 %** | 38 % |
| 21 | **Mathématiques** | 2 377 1% 97% | **81 %** | 19 % |
| 22 | **Sciences cognitives** | 2 370 1% 98% | 53 % | 47 % |
| 23 | **Chimie** | 2 185 1% 99% | **69 %** | 31 % |
| 24 | Psychologie | 2 006 1% 99% | 54 % | 46 % |
| 25 | Géographie | 860 0% 100% | 51 % | 49 % |
| 26 | **Économie et finance quantitative** | 346 0% 100% | 47 % | 53 % |
| 27 | Autres | 79 0% 100% | 54 % | 46 % |
|  | **Sciences exactes** | 73 163 29% | 72 % | 28 % |
| moyenne 65 % écart-type 0.11  écart-type relatif 18 % | |
|  | Sciences humaines et sociales | 177 835 71% | 54 % | 46 % |
| moyenne 53 %  écart-type 0.06  écart-type relatif 10 % | |

Tableau 6 : Distribution des structures selon le domaine

On compte 73 163 titres en sciences exactes, ce qui représente 29 % de notre corpus et 177 835 titres en sciences humaines et sociales, soit 71 %.

Les titres des sciences exactes ont tendance à être plus souvent monosegmentaux que les titres des sciences humaines et sociales. Si l’on regarde la moyenne des répartitions par domaine, l’écart-type relatif important nous pousse néanmoins à la prudence. Parmi les sciences exactes, les mathématiques et la physique utilisent le plus fréquemment des titres monosegmentaux, où ils représentent 81 % des titres. Ces domaines sont suivis par l’informatique, où ils représentent 74 % des titres, suivie de la chimie avec 69 %, des sciences du vivant avec 66 % et des sciences des planètes et de l’univers avec 62 %.

Les sciences humaines et sociales sont globalement plus équilibrées entre l’utilisation de titres monosegmentaux et bisegmentaux. L’écart-type relatif de 10 % montre néanmoins que cet équilibre global varie d’un domaine à l’autre. Ainsi le droit avec 67 %, l’archéologie et la préhistoire avec 61 % et la philosophie avec 60 % privilégient elles aussi le titre monosegmental.

Si on compare la répartition par domaine de notre corpus de travail par rapport à l’ensemble des données initiales, nous avons le même ordre que celui relevé par Tanguy et Rebeyrolle (à paraître). Nous notons également que la répartition entre les domaines n’est pas homogène, certains étant très peu représentés, les plus faiblement dotés étant la géographie avec 860 titres, l’économie et finance quantitative avec 346 titres, et le domaine autres avec 79 titres. D’où la nécessité de travailler en fréquence relative pour les phénomènes que nous étudierons tout en retenant qu’une fréquence relative peut dissimuler un très petit phénomène : un phénomène ayant une fréquence relative importante de 15 % dans le domaine autre, ne concernera finalement que onze titres, rendant ce calcul très sensible à l’ajout ou au retrait d’un titre dans l’ensemble considéré.

### I.3.2 Têtes spécifiques à un domaine

#### A) Définition et principe de sélection

Nous souhaitons faire émerger une liste de têtes spécifiques aux domaines et interpréter ce qu’on y trouve. Intuitivement, on peut penser retrouver les principaux objets d’étude des différents domaines. Pour chaque tête, on peut établir deux séries statistiques ayant autant de valeurs qu’il y a de domaines :

1. Les fréquences relatives de la tête dans les différents domaines :

nombre d’occurrences de la tête dans le domaine  
total des occurrences des têtes du domaine

1. La répartition relative des occurrences de la tête entre les différents domaines :

nombre d’occurrences de la tête dans le domaine  
total des occurrences de la tête dans le corpus

Nous cherchons dans cette partie indifféremment les noms communs et les noms propres. Pour un résultat plus interprétable, lorsqu’une tête de segment qui est un nom propre est suivie par un autre nom propre, ou *de* et un autre nom propre, nous concaténons cette séquence en une seule forme qui devient la nouvelle tête, pour éventuellement réunir un prénom, optionnellement la particule et un nom. Nous estimons qu’il est plus intéressant de tester si *Gustave Eiffel, Gustave Flaubert* et *Gustave Guillaume* sont des têtes spécifiques à un domaine que *Gustave* qui est beaucoup plus générique.

Pour être véritablement spécifique à un domaine, une tête doit y être très fréquente mais également apparaître le moins possible dans d’autres domaines. Nous avons décidé d’utiliser pour sélectionner les têtes spécifiques le calcul de TF\*IDF en l’adoptant à notre configuration particulière. Nous considérons en effet chaque domaine comme un seul vaste document où apparaîtraient tous les titres, le TF est alors la fréquence du terme dans le domaine. Le TF\*IDF adapté devient alors :

TF \* log10(nombre total de domaines / nombre de domaines avec ce terme)

Le calcul de têtes spécifiques à un domaine n’a pas de sens pour le domaine Autres et les titres sans domaine associé : nous gardons donc seulement 25 domaines pour nos calculs. Chaque tête aura alors une valeur de TF\*IDF par domaine, son coefficient de spécificité. Une tête présente dans les 25 domaines aura un TF\*IDF de zéro.

#### B) Corrections de Talismane

Nos résultats, fondés sur un faible nombre d’occurrences, peuvent être très sensibles à une lemmatisation ou une catégorisation erronée d’un mot. Pour les améliorer, nous avons corrigé certaines erreurs et limitations de l’étiquetage morphosyntaxique et de la lemmatisation opérés par Talismane en établissant un dictionnaire de corrections. Le tableau (7) liste nos 13 catégories d’erreurs. Pour savoir comment les corriger, nous avons regardé les différents titres concernés pour établir à chaque fois une règle ad-hoc, la colonne Nombre indiquant le nombre de corrections effectuées :

|  |  |  |  |
| --- | --- | --- | --- |
| **Erreur ou limitation** | **Correction** | **Exemples** | **Nombre** |
| **1. Faux nom propre**  Nom commun erronément catégorisé comme nom propre avec un lemme inconnu car forme avec une majuscule | Lemme ajouté, catégorie corrigée à nom commun | Effet, Adolescence, Autoformation, Approche, Cohomologie, Teneur, Polyhandicap | 228 |
| **2. Faux nom commun**  Nom propre erronément catégorisé commun nom | Catégorie corrigée à nom propre | bitcoin, crétacé | 16 |
| **3. Lemme de nom commun non reconnu**  **car erreur d’orthographe** | Lemme corrigé, catégorie corrigée pour Synth**č**se | Quanti**ﬁ**cation, év**è**nement,  indicateus (-**r**),  Synth**č**se | 17 |
| **4. Lemme de nom commun non reconnu**  **car caractère non compris** | Lemme corrigé (en écrivant oeuvre) | œuvre | 876 |
| **5. Lemme de nom commun non reconnu** | Lemme ajouté | démotorisation, maritimisation, Compactification,  Ondelettes | 849 |
| **6. Lemme de nom propre non reconnu** | Lemme corrigé | Paris, Freud | 1927 |
| **7. Lemme de nom propre non reconnu car concaténation du prénom et du nom** | Lemme corrigé  (nous concaténons prénom et nom lorsque nous avons une tête constituée d’une suite de deux noms propres : ceci n’est pas pris en compte par Talismane et n’est pas à proprement parler une erreur de celui-ci) | Jacques Androuet, Claude Perrault, Jean Cocteau | 66 |
| **8. Forme faussement reconnue comme nom** alors qu’il s’agit d’un adjectif | Forme non prise en compte et retirée de nos calculs | Cyber, Environnemental, Global | 36 |
| **9. E- et Semi- considérés comme un nom propre indépendant** | Lemme corrigé en e- ou semi- + lemme suivant | E-chronic, E-commerce, E-administration, e-inclusion, Semi-figement | 608 |
| **10. s considéré comme un nom commun à cause d’un signe de ponctuation** | On regarde à gauche et à droit du s pour trouver un nom commun ou un nom propre après un signe de ponctuation | mobilité.s, Linguistique(s),  Quel(s) avenir(s) | 404 |
| **11. Mot anglais non reconnu catégorisé à tort comme nom commun** | Forme non prise en compte et retirée de nos calculs, provenant de titres en anglais | The | 59 |
| **12. Nom commun anglais non reconnu** | Prise en considération de son lemme en français | Synthesis, risk, Treatment | 21 |
| **13. Emploi d’un nom propre au pluriel** | Lemme corrigé à la forme singulière | Venises | 1 |

Tableau 7 : Corrections opérées sur l'étiquetage et la lemmatisation

Une fois ces corrections effectuées sur notre corpus de travail, nous pouvons passer notre filtre dessus pour obtenir les têtes spécifiques à certains domaines, en les classant par leur valeur de TF\*IDF.

#### C) Évaluation des résultats

Avec la formule du TF\*IDF, toutes les têtes d’un domaine seront classées par ce facteur de spécificité. Le nombre de têtes différentes par domaine va de 272, pour l’économie et finance quantitative, à 7 005 pour l’histoire en comptant les têtes ayant un facteur d’une valeur de zéro. On compte en tout 30 410 têtes différents. Nous présentons dans le tableau (8) ci-dessous un extrait de notre résultat en prenant les dix premières têtes, selon leur score de TF\*IDF, pour nos 25 domaines. Pour tous les domaines, classés par ordre alphabétique, nous indiquons trois nombres : le nombre de lemmes de têtes, le nombre d’occurrences de têtes et le nombre de titres.

|  |  |  |
| --- | --- | --- |
|  | **Domaine** | **Têtes associées** |
| 01 | **Anthropologie**  2 579 / 6 942 / 5 166 | ethnologie, ethnologue, anthropologie, ethnographie, Népal, sépulture, pentecôtisme, François Cadic, rite, rituel |
| 02 | **Archéologie et préhistoire**  3 444 / 13 391 / 10 124 | céramique, sanctuaire, décor, nécropole, sépulture, occupation, mobilier, archéologie, vaisselle, habitat |
| 03 | **Architecture**  1 624 / 4 629 / 3 444 | ambiance, urbanisme, ville, fortification, habitat, château, photogrammétrie, quartier, Broadacre City, concepteur |
| 04 | **Art et histoire de l’art**  3 376 / 8 685 / 6 471 | vitrail, verrière, décor, musique, peinture, sculpture, notice, théâtre, artiste, peintre |
| 05 | **Chimie**  788 / 2 710 / 2 185 | catalyse, catalyseur, oxydation, ligand, polymère, spectroscopie, hydrogénation, nanoparticule, membrane, préparation |
| 06 | **Droit**  4 189 / 26 398 / 21 486 | droit, juge, clause, obligation, contentieux, chronique, garantie, cession, responsabilité, jurisprudence |
| 07 | **Économie et finance quantitative**  272 / 489 / 346 | aversion, GRP, déterminant, complexification, aluminium, assubilité, polyhandicap, Solvency II, traitement, Paul W |
| 08 | **Éducation**  1 786 / 9 445 / 7 062 | autoformation, didactique, éduction, e-inclusion, hypermédia, enseignant, informatique, scolarisation, ordinateur, TICE |
| 09 | **Géographie**  604 / 1 191 / 860 | démographie, excision, SIDA, fécondité, vigie, mutilation, écologie, appui, géomorphologie, scolarisation |
| 10 | **Gestion et management**  3 546 / 25 955 / 18 318 | comptabilité, management, finance, financement, déterminant, gouvernance, marketing, GRH, RSE, internationalisation |
| 11 | **Histoire**  7 005 / 25 671 / 19 093 | évêque, noblesse, historiographie, manuscrit, femme, guerre, notice, abbaye, protestant, italien |
| 12 | **Informatique**  3 281 / 16 241 / 13 505 | algorithme, ordonnancement, segmentation, extraction, optimisation, routage, détection, minimisation, visualisation, spécification |
| 13 | **Linguistique**  3 435 / 15 512 / 11 556 | néologie, figement, verbe, préposition, grammaticalisation, phonologie, syntaxe, prosodie, adjectif, corpus |
| 14 | **Littératures**  5 142 / 14 278 / 10 712 | littérature, roman, Proust, poétique, Perceforest, poésie, théâtre, Montaigne, René Char, poème |
| 15 | **Mathématiques**  888 / 2 745 / 2 377 | cohomologie, théorème, estimation, package, approximation, optimisation, mathématique, algorithme, compactification, Mixmod |
| 16 | **Philosophie**  2 800 / 7 856 / 6 152 | philosophie, Leibniz, Spinoza, Descartes, Bergson, Kant, Habermas, Nietzsche, Poincaré, Henri Poincaré |
| 17 | **Physique**  3 603 / 30 667 / 26 559 | antenne, optimisation, spectroscopie, spectre, laser, propagation, absorption, excitation, commande, diffraction |
| 18 | **Planète et Univers**  1 244 / 3 675 / 2 781 | ammonite, géologie, Crétacé, gisement, forage, métamorphisme, excursion, datation, bassin, massif |
| 19 | **Psychologie**  943 / 2 663 / 2 006 | autisme, psychologie, psychologue, sevrage, psychanalyse, scarification, psychodrame, psychose, clinique, hallucination |
| 20 | **Science politiques**  2 520 / 9 864 / 7 152 | élection, parti, sociologie, justice, Turquie, État, politisation, décentralisation, vote, parlement |
| 21 | **Sciences cognitives**  1 164 / 3 141 / 2 370 | précocité, proverbe, adjectif, grammaticalisation, catégorisation, phonologie, but, psychologie, prosodie, NBIC |
| 22 | **Sciences de l’environnement**  1 983 / 7 484 / 5 542 | brève, bibliographie, agriculture, karst, muraille, Pralognan, cadastre, émission, forêt, Médiaterre |
| 23 | **Sciences de l’information et de la communication**  2 053 / 7 523 / 5 481 | journalisme, média, bibliothèque, télévision, sémiotique, journaliste, SIC, communication, blog, open |
| 24 | **Sciences du Vivant**  3 800 / 22 149 / 17 498 | dosage, protéine, lait, acide, sécrétion, digestion, infection, alimentation, teneur, nutrition |
| 25 | **Sociologie**  5 268 / 32 398 / 23 732 | sociologie, ville, géographe, géographie, tourisme, nuit, quartier, socialisation, déscolarisation, territoire |

Tableau 8 : Les dix têtes les plus spécifiques de chaque domaine

On constate plusieurs faits : le premier est une mise en garde sur la limite de dix têtes choisies pour la présentation du tableau (8). Selon le nombre de titres et le nombre de lemmes différents dans le domaine, les mots sélectionnés peuvent avoir des nombres d’occurrences très variés. Plus le nombre de titres étudiés est faible, plus le nombre d’occurrences sur lequel est fondé le résultat est bas. Un résultat dépendant d’un faible nombre d’occurrences est beaucoup moins fiable quant à sa reproductibilité sur un autre corpus et donc sa généralisation.

Dans l’éducation, on constate la présence dans les dix premières spécifiques de *ordinateur* alors qu’il apparaît en 72e position en informatique. Avec seulement 12 occurrences en informatique, on peut présumer qu’il s’agit là d’un terme trop générique pour la science dont c’est le principal objet et donc délaissé dans les titres soumis à une forte contrainte informationnelle et de concision.

Les sciences cognitives, avec seulement 2 370 titres, sont à la croisée de plusieurs domaines, notamment la linguistique et la psychologie, ce qui peut expliquer la non-présence de têtes « propres ».

Seule l’évaluation des résultats permet de juger de la pertinence de notre méthode. Le premier contrôle que nous pouvons effectuer, bien que très subjectif et limité, et de parcourir nous-même ces têtes, classées par TF\*IDF pour voir si les premières semblent plus correspondre au domaine associé que les suivantes. Ce premier contrôle est positif : les têtes avec le plus haut TF\*IDF semblent effectivement les plus proches des objets d’études des domaines, comme *céramique* et *nécropole* pour l’archéologie. L’extrême majorité des têtes ayant un TF\*IDF élevé est ce que Schmid (2000, p. 15) appelle des *full-content nouns* ayant un contenu sémantique important. Nous remarquons néanmoins la tête *brève*, pour le domaine des sciences de l’environnement, qui ne désigne pas un objet d’étude mais un support de publication.

Une méthode d’évaluation possible était de comparer les têtes avec des lexiques spécialisés pour mesurer la précision et le rappel. Néanmoins, cela exige de ne sélectionner qu’une partie des têtes spécifiques à l’aide d’un filtre pour ne pas avoir trop de bruit. Ce filtre peut être un seuil appliqué à la valeur de TF\*IDF plutôt que de prendre les X premières têtes. Il faudrait dans ce cas faire attention au taux de couverture des têtes sélectionnées par un tel seuil : plus il sera élevé, plus on sera sûr d’avoir des têtes spécifiques au détriment du nombre de titres couverts.

Une autre approche est de calculer une distance entre les domaines : si on assimile les domaines à un sac de têtes, où pour chaque tête, on met la valeur de son TF\*IDF ou 0 si la tête n’apparaît pas dans le domaine, on obtient un vecteur à 30 410 dimensions. On peut ensuite calculer une distance généralisée entre les domaines pour savoir lesquels sont les plus proches et les plus éloignés en termes de têtes dont l’annexe **A1. Distance des domaines de par leurs têtes spécifiques** présente le résultat. On constate ainsi que la sociologie a pour domaine le plus proche le domaine gestion et management et comme domaine le plus éloigné la chimie. Cette représentation permet d’avoir un aperçu de comment les domaines se positionnent les uns par rapport aux autres par la spécificité de leurs têtes et ainsi vérifier la valider de notre approche.

Un lemme peut avoir une valeur distinctive de TF\*IDF dans plusieurs domaines, ce qui traduit que cet objet d’étude est partagé par les différents domaines et qu’il n’a pas de pertinence pour un grand nombre d’autres. L’importance dans chaque domaine est pondérée par la valeur de TF\*IDF. Par exemple, *femme* a une valeur de 0,0003 en géographie, sciences de l’information et de la communication et psychologie, 0,0004 en sociologie, anthropologie et littérature et 0,0007 en histoire. Néanmoins, une forte limite de cette approche est la polysémie de certaines têtes. L’*architecture* en informatique n’est pas la même que dans le domaine de l’architecture, de même qu’une *tempête* en sciences exactes et en sciences humaines et sociales.

Nous avons dans cette partie établi le périmètre délimitant notre corpus de travail et mesuré ses contours. Nous avons décidé d’étudier le cas le plus nombreux : celui des titres monosegmentaux ou bisegmentaux possédant au moins une tête nominale.

Notre corpus de travail se compose de 250 998 titres, soit 74 % du matériau initial. Notre corpus de travail est représentatif du matériau initial en ce qui concerne la répartition des titres par type de publication, nombre d’auteurs ou domaine. Nous avons démontré que les titres sont essentiellement des syntagmes nominaux à 85 % si on ne considère que le premier segment des titres bisegmentaux et les titres monosegmentaux.

Nous avons relevé que la répartition des titres par domaine n’est pas homogène, 71 % des titres se rapportent aux sciences humaines et sociales contre 39 % pour les sciences exactes. Les premières utilisent de façon à peu près égale les titres monosegmentaux et bisegmentaux alors que les sciences exactes favorisent les titres monosegmentaux. Nous avons également calculé un indice de spécificité des têtes par rapport à un domaine particulier en se fondant sur le TF\*IDF.

Nous voulons à présent étudier des têtes qui sont à l’inverse des têtes spécifiques : des têtes des fréquentes dans toutes les disciplines, que nous appellerons têtes transdisciplinaires.

# II. Têtes transdisciplinaires et NSS dans le corpus

Dans cette partie, nous nous intéressons aux têtes de nos segments. Nous avons vu que nous avons une tête par segment et de un à deux segments par titre. Cela fait donc trois sous-ensembles de notre corpus de travail : les segments des titres monosegmentaux, les premiers segments des titres bisegmentaux et les seconds segments des titres bisegmentaux. Nous allons étudier dans ces trois ensembles les têtes de segments qui sont très fréquentes dans de nombreux domaines, des têtes que nous appellerons transdisciplinaires. Ce sont ces dernières têtes, que nous voulons rapprocher des noms sous-spécifiés que nous décrivons dans la sous-partie suivante. Enfin, nous abordons les schémas récurrents dans lesquels s’insèrent nos têtes transdisciplinaires pour essayer d’en faire le rapprochement avec les constructions spécificationnelles des noms sous-spécifiés.

## II.1 Sélection des têtes transdisciplinaires

### II.1.1 Principe de sélection

Pour être véritablement transdisciplinaire, une tête ne doit pas seulement se retrouver dans de nombreux domaines. Elle doit se retrouver *fréquemment* dans de nombreux domaines. Nous nous méfions de la moyenne des fréquences relatives de la tête dans les différents domaines car elle peut cacher des situations très disparates. Nous préférons prendre les têtes qui apparaissent avec un seuil minimum dans la moitié des domaines étudiés dans notre corpus. Nous établissons donc un seuil arbitraire de 0,001 (0,1 %), que nous nommons **seuil de médiane**, au-dessus duquel nous sélectionnons nos têtes transdisciplinaires.

### II.1.2 Résultats et évaluations des têtes transdisciplinaires

Sur les 123 227 lemmes de têtes de notre corpus de travail, cela en sélectionne 94 soit 0,08 %. Elles ont en tout 94 739 occurrences, soit près de 27 % des 354 168 occurrences de têtes que compte notre corpus. Elles se répartissent ainsi :

* 40 270 des titres monosegmentaux ont une tête transdisciplinaire, soit 27 %,
* 8 147 titres bisegmentaux ont une tête transdisciplinaire dans chaque segment,
* 9 592 premiers segments de titres bisegmentaux ont une tête transdisciplinaire, soit 17 % des premiers segments,
* 28 583 seconds segments de titres bisegmentaux ont une tête transdisciplinaire, soit 36 % des seconds segments,
* 86 592 des titres ont au moins une tête transdisciplinaire, soit 34 %.

Les occurrences de ce très petit nombre de têtes transdisciplinaires concentrent plus d’un quart de toutes les têtes et plus d’un tiers de tous les titres.

Les 20 premières têtes des 94 classés par la médiane sont : *étude, analyse, cas, approche, exemple, enjeu, évolution, apport, rôle, modèle, réflexion, évaluation, outil, question, représentation, application, construction, introduction, histoire* et *développement*. La liste complète est fournie dans l’annexe **A3. Liste des têtes transdisciplinaires**.

Aucun nom propre ne figure dans cette liste ce qui semble logique aux égards des principales catégories de noms propres. Sur le plan des personnes et des titres d’œuvres, on peut difficilement imaginer un individu présent dans la plupart des domaines ou une œuvre relative à la plupart des domaines, l’*Encyclopédie* relevant plus d’un intérêt historique que scientifique de nos jours. Sur le plan des périodes historiques, elles sont surtout propres aux domaines aux études diachroniques et ne sont pas toutes les mêmes pour les différents domaines : le *Crétacé* n’intéresse par exemple que les préhistoriens. Sur le plan des lieux, on pourrait imaginer qu’un lieu soit étudié selon différents domaines des sciences humaines et sociales comme la géographie, l’histoire ou la sociologie, mais cela est plus difficilement concevable pour les sciences exactes, l’exemple forgé *la mathématique de Paris* ayant peu de sens, à moins de parler d’une école de pensée. En classant par la médiane, les deux premiers noms propres sont *Europe* avec 0,02 % et *Paris* avec 0,01 %, les deux sont présents dans 16 domaines, dont 14 sont communs et 12 font partie des sciences humaines et sociales.

Une autre remarque est qu’il s’agit de noms communs abstraits ayant un faible contenu sémantique dénotant des entités de second ou troisième ordres définis par Lyons (1977) : elles n’existent pas mais peuvent avoir une localisation spatiotemporelle (Benítez-Castro (2014, p. 96). Néanmoins, certaines têtes peuvent avoir plusieurs sens, par exemple *outil* peut dénoter une entité ayant une matérialité tangible, fabriquée dans un but précis, comme un tournevis. Mais dans une acception plus figurative et métaphorique selon le CNRTL[[4]](#footnote-5), *outil* peut désigner toute entité étant un moyen « *qui permet d’obtenir un résultat, d’agir sur quelque chose*» ce qui est très vaste. Notons que nous avons déjà remarqué ce lemme dans notre travail de première année et statué que dans ce cas, il était équivalent à d’autres lemmes comme *système* ou *dispositif* qui sont eux-aussi présents dans notre liste de têtes transdisciplinaires. On peut rapprocher les têtes transdisciplinaires des noms généraux de Halliday et Hasan (1976), des noms ayant une *« sous-spécification sémantique »* et une *« large couverture référentielle »* (Huyghe, 2018).

Le premier contrôle possible pour tester la validité de notre filtre est de compter les domaines où ces têtes sont présentes. Tutin (2008) fixe la transdisciplinarité au fait d’avoir au moins 15 occurrences dans les trois domaines étudiés par son étude, soit 100 % des domaines étudiés. Hatier et al. (2016) fixent la transdisciplinarité au fait d’apparaître dans quatre domaines sur les dix que comporte leur étude, soit 40 %. Dans notre cas, 40 % de nos 25 domaines retenus pour nos calculs donne un seuil de dix domaines. Nos 94 têtes transdisciplinaires sont au minimum présentes dans 20 domaines, soit 80 % des 25 domaines. 35 têtes transdisciplinaires sont présentes dans 100 % des 25 domaines. Le nombre moyen de domaines où les 94 têtes sont présentes est 23,95 ce qui est extrêmement élevé sachant que le nombre minimum de domaines est de 20.

Un second contrôle consiste à confronter nos têtes transdisciplinaires à des lexiques de noms transdisciplinaires préétablis par d’autres méthodes que la nôtre. Ainsi, si nous retombons sur les mêmes lemmes, nous corroborons à la fois les études précédentes et notre travail. Nous avons utilisé le lexique transdisciplinaire des écrits scientifiques (LTES) établie par Tutin (2007, 2008), qui compte 363 noms, et le lexique scientifique transdisciplinaire (LST) de Hatier (2016) qui compte 495 noms. Dans notre annexe **A3. Liste des têtes transdisciplinaires** nous indiquons pour chaque tête son appartenance au LTES et au LST.

Sur les 94 têtes transdisciplinaires, 74 sont présentes dans le LTES, soit 79 %. Les 20 têtes qui ne figurent pas dans le LTES sont : *conception, discours, dynamique, défi, émergence, enjeu, enseignement, essai, formation, histoire, impact, jeu, méthodologie, note, point, politique, regard, remarque, retour, science*. On peut se demander pourquoi des lemmes sémantiquement liés directement à la science comme *méthodologie* ou *science* ne figurent pas dans le LTES. Une raison possible est que Tutin a établi son lexique sur trois domaines seulement, médecine, économie et linguistique, où les lemmes mentionnés pourraient être moins fréquents. Les autres peuvent avoir été considérés comme trop génériques : il en effet difficile de délimiter ce qui est propre au vocabulaire scientifique, le lexique transdisciplinaire des écrits scientifiques étant considéré comme un sous-ensemble d’un lexique abstrait plus général (Tutin, 2007). De plus, Tutin (2008, p. 247) effectue un filtrage manuel et donc potentiellement subjectif.

Sur les 94 têtes transdisciplinaires, 82 sont présentes dans le LST, soit 87 %. Les têtes non présentes dans le LST sont *an, défi, enseignement, formation, gestion, histoire, jeu, modélisation, place, politique, regard, retour*. Le corpus sur lequel Hatier a construit son lexique contenait uniquement des articles de sciences humaines et sociales ce qui peut expliquer cette différence. La validation manuelle effectuée par des experts (Hatier, 2016, p. 80) sur les candidats au LST peut aussi expliquer leur non prise en compte.

On peut s’interroger sur cette non-concordance entre nos têtes transdisciplinaires, le LTES et le LST, avec huit têtes transdisciplinaires n’appartenant à aucun des deux : *défi, enseignement, formation, histoire, jeu, politique, regard, retour*. On peut l’imputer au matériau de départ, seulement trois domaines pour Tutin (2008) et des domaines exclusivement issus des sciences humaines et sociales pour Hatier (2016). On pourrait l’imputer à un seuil de médiane trop bas : *jeu* est la 93e tête sur 94 dans l’ordre décroissant des valeurs de médiane, *an* la 89e et *défi* la 88e. Néanmoins *rapport* qui est la 94e tête dans ce même classement appartient bien au LTES et au LST. A contrario, *histoire* qui n’appartient ni au LTES ni au LST est placé à la 19e place dans notre classement. Il ne s’agit donc pas d’une conséquence liée à la valeur de seuil choisie. Reste qu’aussi bien Hatier que Tutin sélectionnent un vocabulaire à la fois transdisciplinaire et *spécifique* à la science, à l’aide pour Hatier d’un corpus de contraste qui représente *« la compétence linguistique générale »* (Hatier, 2016, p. 30). De notre côté, nous affirmons par le calcul seulement la transdisciplinarité de nos têtes, pas leur appartenance à un lexique scientifique propre. Lexique scientifique propre non pas par l’exclusivité des lemmes y figurant mais par leurs fréquences plus élevées. L’hypothèse intuitive que les têtes les plus fréquentes dans des titres d’articles scientifiques seraient forcément des lemmes membres d’un lexique scientifique propre est donc à nuancer : *jeu, regard* et *retour* ne connotent rien de scientifique.

### II.1.3 Études des têtes selon leurs segments et la structure segmentale du titre

Nous avons ensuite étudié les variations des têtes transdisciplinaires entre trois sous-ensembles de notre corpus de travail : les titres monosegmentaux, les premiers segments des titres bisegmentaux, puis leurs seconds segments. Nous traitons les segments des titres bisegmentaux séparément pour essayer de déterminer d’éventuelles différences entre les deux. L’hypothèse d’une différence repose sur les modèles de titres données par Swales et Feak (1994) où chaque segment à un rôle particulier :

1. Problème : solution
2. Général : spécifique
3. Sujet : méthode
4. Majeure : mineure

D’autres auteurs ont étendu cette typologique comme Anthony (2001) avec cinq types ou Cheng et al. (2012) avec onze types, et à chaque fois, les deux segments ont un but sémantique différencié. Nous voulons savoir si cette différenciation sémantique se reflète par des différences dans les têtes transdisciplinaires émergeant des différents sous-corpus.

Pour les titres monosegmentaux, les têtes transdisciplinaires relevées sont au nombre de 81. Six seulement d’entre elles n’apparaissent pas dans les 94 têtes transdisciplinaires relevés sur tout le corpus. Les six têtes sont : *contrôle, fonction, notion, temps, transformation* et *valeur*. Les six appartiennent au LTES. Seuls les lemmes *contrôle, temps* n’appartiennent pas au LST. On peut noter que dans le corpus général, les têtes *fonction, temps* et *valeur* ont une valeur de médiane de 0,09 %, *transformation* de 0,08 % et *contrôle* et *notion* de 0,07 %. En changeant le corpus sur lequel on fait nos calculs, ces têtes passent en dessous ou au-dessus de notre seuil, fixé à 0,1 %, mais dans les deux cas, elles ont une valeur de médiane proche.

Pour le premier segment des titres bisegmentaux, nous relevons 63 têtes transdisciplinaires. Cinq têtes n’apparaissent pas dans les 94 précédemment relevées : *compte, contribution, culture, économie* et *identité*. Dans le second segment, nous relevons 99 têtes transdisciplinaires et 19 têtes n’apparaissent pas dans les 94 têtes transdisciplinaires relevés sur tout le corpus : *condition, contexte, définition, démarche, donnée, illustration, leçon, limite, mode, mythe, paradoxe, parcours, piste, problématique, réalité, revue, source, synthèse* et *voie*. Si on dénombre toutes les têtes transdisciplinaires relevées par l’étude du corpus et des trois sous-corpus, on obtient le nombre de 123. Le tableau (9) résume le nombre de têtes transdisciplinaires trouvées par corpus.

|  |  |
| --- | --- |
| **Corpus** | **Nombre de têtes transdisciplinaires** |
| Ensemble du corpus de travail | 94 |
| Titres monosegmentaux | 81 |
| Premier segment des titres bisegmentaux | 63 |
| Second segment des titres bisegmentaux | 99 |
| Fusion des quatre listes | **123** |

Tableau 9 : Nombre de têtes transdisciplinaires selon le corpus choisi

Pour le sous-corpus des seconds segments des titres bisegmentaux, on remarque que cinq têtes, *cas, exemple, étude, application* et approche représentent 13 % des têtes du sous-corpus, il y a donc une concentration remarquable sur un très petit nombre de têtes. Pour atteindre 13 % en cumulant les têtes les plus fréquentes dans le corpus général, il faut prendre 19 têtes. Individuellement, les occurrences des têtes *cas, exemple, étude, application* et *approche* représentent respectivement 4 % pour *cas*, 3 % pour *exemple* et 2 % pour les trois dernières des 95 282 occurrences de têtes du sous-corpus.

On remarque également une surreprésentation de certaines têtes dans le sous-corps des seconds segments par rapport au corpus général. Ainsi, *cas* ne compte pour 1 % des têtes du corpus général contre 4 % dans le sous-corpus, et de même pour *exemple* avec 1 % dans le corpus général et 3 % dans le sous-corpus. Cette affirmation rejoint nos résultats de première année, où nous montrions que certains lemmes privilégiaient très fortement une position après le double-point, et donc dans un deuxième segment. 97 % des occurrences de la tête *cas* apparaissent ainsi dans un deuxième segment, 93 % des occurrences de la tête *exemple* suivent la même logique. Cette répartition montre bien l’application du schéma général : spécifique de Swales et Feak (1994). Haggan (2004) appelle cette opération un resserrement *« narrowing »* (Haggan, 2004) sur l’objet de l’article. La juxtaposition de ces deux informations par le double point rend plus facile leur interprétation, soulignée par le choix d’une tête transdisciplinaire qui indique bien un point spécifique, un cas, un exemple, d’un concept plus large.

Néanmoins, il faut prêter attention ne pas faire de contresens en interprétant ce schéma sémantique. Si les exemples (21) et (22) présentent bien, dans leurs premiers segments, le concept général suivi, dans leurs seconds segments, d’un exemple spécifique, le (23) l’inverse. L’exemple (21) utilise *l’exemple de* alors que l’exemple (23) utilise *un exemple de*. La détermination de la tête oriente donc l’interprétation du schéma. L’exemple (22) utilise une troisième association qui juxtapose, via une virgule, *un exemple,* avec le point spécifique et montre bien que s’appuyer seulement sur la détermination ne suffit pas pour interpréter ce schéma.

(21) La **cohésion** ‘to-textuelle' de l'oral : l'**exemple** d'une interview de Jean-Claude Van Damme

(22) **Ressources** lexicales pour une sémantique inférentielle : un **exemple**, le mot quitter

(23) L'**internet** des mouvements transgressifs : un **exemple** de “transnationalisation” des identités militantes.

Une dernière étude possible est, au lieu d’étudier d’un côté les têtes du premier segment et de l’autre celle du second segment des titres bisegmentaux, d’étudier les couples de têtes ainsi formés. On forme ainsi des couples ordonnés de la forme (tête premier segment, tête second segment) et on peut, pareillement qu’une tête seule, calculer une fréquence relative sur l’ensemble des titres bisegmentaux.

Seuls cinq couples ont une médiane différente de zéro : *(de, exemple), (rôle, cas), (approche, cas), (apport, exemple) et (effet, cas)*. Nous rappelons que les titres bisegmentaux peuvent avoir une tête non nominale, verbe ou préposition, d’où l’apparition de la préposition *de* dans ce classement. La préposition *de* étant la plus fréquente, il est logique qu’elle apparaisse dans les couples les plus fréquents. Notre étude portant essentiellement sur les têtes nominales, nous ne retenons que les trois derniers couples. Les exemples (24), (25) et (26) montre respectivement les couples *(rôle, cas)*, 60 titres, *(apport, exemple)*, 37 titres, et *(effet, cas)*, 36 titres.

(24) Le **rôle** des confréries durant les premières années de la proscription du catholicisme : Le **cas** du fief de Shimabara

(25) L'**apport** des archives privées à l'histoire politique : l'**exemple** de Louis Costa, le notaire rouge

(26) Les **effets** pervers de la solidarité: le **cas** des femmes seules avec enfants

Du fait des têtes transdisciplinaires du second segment, *cas* et *exemple,* nous retombons sur le schéma de Swales et Feak (1994) général : spécifique.

Nous avons donc identifié 94 têtes transdisciplinaires dans notre corpus de titres, à la fois très fréquentes et ayant un très faible contenu sémantique. Nous avons identifié également des têtes transdisciplinaires émergeant si l’on ne considère que les titres monosegmentaux, le premier ou le second segment des titres bisegmentaux comme *cas* ou *problème*. Dans notre travail de première année, nous avions déjà souligné cette répartition, en faveur d’une position post-double-point, de certains lemmes assimilés à nos têtes. N’importe quel concept peut avoir des exemples ou des cas : la très forte capacité référentielle de ces deux lemmes n’est possible que parce qu’ils dénotent uniquement une relation, de l’entité qu’il référencie envers une autre, sans donner plus d’information sur leurs référents. Partant de ces deux cas exemplaires étudiés en première année, nous avons souhaité étendre notre recherche de noms fréquents avec un faible contenu sémantique à l’ensemble des têtes d’un titre, en particulier les têtes transdisciplinaires. Ces deux caractéristiques de fréquence et de sous-spécification rapprochent les têtes transdisciplinaires d’une classe d’emploi des noms, les noms sous-spécifiés, très fréquentes dans le discours académique (Flowerdew et Forest, 2015).

## II.2 Noms sous-spécifiés et constructions spécificationnelles

### II.2.1 Définitions des noms sous-spécifiés

De nombreux travaux avec différentes perspectives se sont penchés sur les noms sous-spécifiés en anglais et plus tardivement en français. Les définitions théoriques et opératoires de ces différents auteurs ne se recoupent pas exactement, ainsi que la liste des noms pouvant être employé de la sorte, ce qui se traduit par un foisonnement terminologique (Flowerdew et Forest, 2015, p. 9 ; Schmid, 2000, p. 10-11) : *container nouns* (Vendler, 1968), *signalling nouns* (Flowerdew 2003, 2006 ; Flowerdew et Forest, 2015), *type 3 vocabulary* (Winter, 1977), *metadiscursive nouns* ou *anaphoric nouns* (Francis, 1986), *enumerables* et *advance labels* (Tadros, 1994), *carrier nouns* (Ivanic, 1991), *advance labels* et *retrospective labels* (Francis, 1994), *unspecific nouns* ou *metalanguage nouns* (Winter, 1992), *shell nouns* (Hunston et Francis, 1999 ; Schmid, 2000, 2018), *noms sous-spécifiés* (Legallois, 2008) et *noms porteurs* (Huygue, 2018). Nous retenons que la plupart de ces travaux s’accordent pour définir les NSS comme un emploi particulier et non une classe lexicale : un nom peut ainsi être employé de façon sous-spécifié ou non, nous reprenons pour le 1er cas un exemple de Huygue (2018) :

Emploi sous-spécifié du nom *fait* : le fait que Pierre soit arrivé en retard à la réunion

Emploi spécifié du nom *fait* : Il n’y a aucun fait qui étaye cette supposition.

Cette différenciation entre emploi sous-spécifié et emploi sémantique plein est le cœur du problème d’un traitement automatique. Nous devons pouvoir rendre la différenciation computationnelle, c’est-à-dire déterminable par un programme. Nous nous tournons donc vers les définitions théorique et opératoire pour construire un tel algorithme.

Comme définition théorique, nous nous proposons de reprendre celle de Flowerdew (2006, p. 348) pour sa concision et sa clarté : *« potentially any abstract noun which is unspecific out of context, but specific in context ».* Un exemple d’emploi sous-spécifié pour le lemme *défi* est le suivant : *Pour les Américains, le* ***défi*** *est de marcher à nouveau sur la Lune*. Le sens complet de *défi* ne peut être appréhendé qu’en faisant référence au contexte, ici *marcher à nouveau sur la Lune*.

La définition de Flowerdew est néanmoins très générale. Pour compléter notre définition théorique, peut donner une définition fonctionnelle en rappelant les trois fonctions clés de l’emploi sous-spécifié selon Schmid (2000, p. 14 ; 2018, 112) :

* Fonction cognitive de **conceptualisation** ou d’**encapsulation** : un morceau d’information est encapsulé dans création un concept temporaire nominal. Dans notre exemple, *marcher à nouveau sur la Lune* est encapsulé dans un concept nominal, le ***défi***.
* Fonction sémantique de **catégorisation** ou de **classification** : catégorisations de concepts, il s’agit d’une mise en perspective par le locuteur d’un morceau d’information qu’il souhaite transmettre à l’interlocuteur en lui imposant son point de vue (Legallois, 20108, p. 8). Le fait d’utiliser ***défi*** et non ***problème***n’est pas neutre, comme nous le verrons plus bas, car si le NSS a un manque de contenu sémantique, il n’en est toutefois pas dénué complètement.
* Fonction textuelle ou discursive de **liaison** : capacité de référence quasi-pronominale au concept créé qui structure le texte, qui pourra être repris par exemple par l’énoncé *ce défi*. C’est cette fonction qui assure une cohérence et une continuité au texte et qui intéresse d’un point de vue discursif Flowerdew et Forest (2015, p. 2)

On peut noter, à la suite de Huyghe (2018, p. 36), que même si c’est une classe d’emplois et non une classe lexicale, la fonction d’encapsulation est néanmoins conditionnée par leurs propriétés sémantiques. Legallois (2008, p. 2) parle d’une *« interdéfinition entre lexique et grammaire »*. Schmid (2000, p.  63-73) reprend la distinction en trois de ordres de Lyons (1977) des entités dénotées par les noms. Nous traduisons ci-dessous la définition qu’en fait Benítez-Castro (2014, p. 96) :

*« Les entités de premiers ordres sont des éléments tangibles du monde réel, comme les personnes, les animaux et les objets, ayant une localisation spatiale et temporelle, les entités de second ordre, qui n’existe pas mais arrive ou prenne place, avec une localisation spatiotemporelle, comme un crime, un mouvement ou un combat, enfin les noms de troisièmes ordres convoient purement des significations abstraites, idées, propositions ou faits, comme théorie, affirmation, aspect)*. *Les second et troisième ordres partagent une nominalisation comme origine mais les entités de troisièmes ne sont pas observables et en dehors de toute dimension spatiotemporelle. »*

Schmid (2000, 63-73) considère que les emplois de NSS ne concernent que les noms qui dénotent des entités de second ou de troisième ordres. Schmid (2000, p. 85 ; 2018 p. 118) distingue également trois grandes catégories de shell nouns selon leur contenu sémantique. La première est les *prime shell nouns,* des noms utilisés de façon privilégiée et fréquente dans des emplois de NSS comme *concept, fact, issue, idea, notion, principle, problem, reason, rumour, legend* et *thing*. La seconde est celle des *good shell nouns* (Schmid, 2000, p. 86) qui sont moins interchangeables entre eux comme *order*. La troisième catégorie se compose des *less good* ou *peripheral shell nouns* comme *move, measure, reaction, situation, way, procedure.*

Comme Flowerdew et Forest (2015, p. 12), nous considèrerons les NSS comme une classe ouverte même s’ils empruntent à la classe fermée des pronoms la caractéristique d’avoir besoin d’un contenu spécifiant (Flowerdew et Forest , 2015, p. 11). Le fait qu’ils proviennent de la classe lexicale des noms leur conférent un statut intermédiaire entre la classe fermée pronominale et la classe ouverte nominale (Huyghe, 2018, p.44 ; Legallois, 2006, p. 11).

Huyghe (2018) distingue les noms généraux tels que définit par Halliday et Hasan (1976), *« a small set of nouns having generalized reference »,* servant à construire la cohérence du texte, des NSS, appelés noms porteurs dans son article, tout en reconnaissant la possibilité d’appartenir aux deux classes. Cette distinction n’est pas aussi franche pour Flowerdew et Forest (2015, p. 9) et les deux se rejoignent sur la notion de non-spécification (Schmid, 2000, p. 10). Ce qui distingue les NSS des noms généraux, c’est le focus mis sur les structures grammaticales dans laquelle ils s’insèrent et qui en devient une définition opératoire.

### II.2.2 Les constructions spécificationnelles classiques

#### A) Définitions des constructions spécificationnelles

Un NSS s’insère au sein d’une construction spécificationnelle (CS) qui va relier le NSS à un contenu qui va le spécifier ou le « remplir » (Legallois, 2008, p. 6). Cette opération est appelée spécification ou identification (Nakamura, 2017, p. 3).

Nous recensons ici les deux constructions spécificationnelles les plus fondamentales étudiées par Schmid (2000, p. 22) pour l’anglais. Legallois (2008, p. 2) a transposées ces constructions en français et elles ont été reprises par Huyghe (2018, p. 36), Roze et al. (2014, p. 4) et Nakamura (2017, p. 2) :

1. **NSS** + être + proposition subordonnée conjonctive :   
   Le problème est que l’homme abandonne son habitat.
2. **NSS** + être + de+ proposition subordonnée infinitive :   
   Le plus grand **effort** est de vaincre les passions.

Kolhatkar et Hirst (2014, p. 4) montrent que les NSS ont une préférence pour certaines constructions spécificationnelles. En se fondant sur une étude du corpus du français frWacen son état du 25 avril 2017, Huyghe (2018, p. 37) indique qu’un NSS peut n’accepter qu’une des deux constructions : ainsi le NSS *capacité* s’utilise avec une infinitive, comme dans *la capacité de sélectionner les candidats*,et non avec une conjonctive, \**la capacité que le jury sélectionne les candidats*. Cette *« compatibilité propositionnelle »* conditionne la syntaxe, le type de proposition subordonnée, mais il conditionne également la sémantique du contenu propositionnel (Huyghe, 2018, p. 38) : *action* implique la dynamicité, *propriété*, la stativité. Ainsi chaque NSS a une *« capacité de portage propositionnel »* plus ou moins étendue (Huyghe, 2018, p. 37).

Ils existent pour (CS-I) et (CS-II) à chaque fois une variante qui peut être rapprocher des pseudo-clivées en insérant, entre le NSS et être, une virgule et le pronom de reprise *ce* (Legallois et Gréa, 2006, p. 161 ; Roze et al., 2014, p. 4), variantes qui se rencontrent notamment à l’oral :

1. **NSS** + virgule + c’/ce + être + proposition subordonnée conjonctive :   
   Le problème, c’est que l’homme abandonne son habitat.
2. **NSS** + virgule + c’/ce + être + de+ proposition subordonnée infinitive :   
   Le plus grand **effort**, c’est de vaincre les passions.

De plus, Schmid propose notamment trois autres constructions spécificationnelles (Schmid, 2000, p. 22, 26) que Legallois (2008, p. 2) n’a pas repris :

1. **NSS** + proposition subordonnée conjonctive :  
   Le **problème** que l’homme abandonne son habitat n’a toujours pas été discuté en commission.
2. **NSS** + de+ proposition subordonnée infinitive :  
   Le plus grand **effort** de vaincre ses passions a été exigé de lui.
3. **NSS** + **of** + syntagme avec pour noyau un verbe au gérondif :  
   The **problem** of raising money

CS-V et CS-VI reprennent CS-I et CS-II mais sans le verbe être. Legallois (2008, p. 2) qualifie ces constructions *« d’apparentées »* et indique que les shell nouns de Schmid (2000) sont *« une catégorie plus large que les NSS »*.

La CS-VII réunit dans un syntagme nominal complexe le NSS et le contenu spécifiant, dont le noyau est un déverbal, un gérondif en anglais. Cette dernière contrainte se retrouve implicitement dans les exemples de son ouvrage (2000). Schmid lève la contrainte du gérondif de la CS-VII dans l’exemple qu’il donne dans son article de 2018 (p.115) : *« The* ***notion*** *of love »*. On remarque néanmoins que le déverbal *love*, de *to love*, est toujours un nom dénotant une entité d’ordre supérieur à un, l’action d’aimer. Le gérondif anglais n’ayant pas d’équivalent direct en français, une traduction possible vers le français est l’infinitif, choix (A), rejoignant alors la construction CS-VI ou d’un nom déverbal choix (B), rejoignant alors la construction CS-VII.

(A) The problem of raising money 🡺 le problème de lever/réunir/recueillir de l’argent/des fonds

(B) The problem of raising money 🡺 le problème de la levée / du recueil / de la réunion des fonds

En français, dans le cas d’un déverbal dénotant une action ou une activité (DDAA), la difficulté se pose pour la CS-VII de distinguer automatiquement les emplois proprement sous-spécifiés des emplois en nom plein suivi d’un complément de nom. Ainsi Roze et al. (2014, p. 8) indiquent que les énoncés *« marché de travail*, *contrat de travail »* ne sont pas des NSS alors que *travail* est le déverbal de travailler. Cela nous semble justifié par le degré de figement des ces énoncés qui fait de *contrat de travail* une locution nominale où contrat n’est pas substituable par un autre nom.

On peut se demander jusqu’à quel point cette contrainte d’avoir pour contenu spécifiant un DDAA est vérifiée. Prenons *« projet de loi »* qui est également refusé par Roze et al. (2014, p. 8) comme toutes les formes *projet de NC*. Il nous semble néanmoins possible de rapprocher les trois énoncés suivants qui reprennent les trois constructions spécificationnelles :

1. CS-I Le projet que l’État légifère contre le vapotage dans les lieux publics.
2. CS-II Le projet de légiférer contre le vapotage dans les lieux publics.
3. CS-VII Le projet de loi contre le vapotage dans les lieux publics

L’équivalence des trois peut se comprendre en sous-tendant une action implicite dans la CS-VII, celle de rédiger/émettre une loi. L’interprétation sémantique dans ce dernier cas est ambiguë, à savoir si le *projet* concerne l’acte de rédiger, on se rapproche d’un NSS, si le *projet* concerne la loi en elle-même, ou si l’on doit interpréter *projet de loi* comme une locution nominale dénotant une classe distincte, en ne dissociant pas *projet* et *loi*. Si nous rejetons le dernier emploi comme NSS, le second ouvre la possibilité d’avoir une action implicite sous-tendue par le nom, qu’un nom DDAA rend explicite.

Aktas et Cortes (2008) avaient également relevé cette construction sans contrainte particulière sur le nom la terminant. Mousavi et Rauof Moini (2014), dans une étude sur un corpus d’articles scientifiques sur l’éducation, ont constaté que cette construction était la plus fréquente.

Nakamura (2017, p. 2) reprend également la construction suivante comme CS, cette fois toujours avec le verbe être :

1. **NSS** + être + syntagme nominal :  
   Notre **objectif** majeur est la rédaction d’une proposition de loi.

On remarque le noyau du syntagme nominal droit est *rédaction*, un DDAA qui renvoie à l’action de rédiger. L’équivalence avec *notre objectif majeur est de rédiger* est ainsi directe.

Nakamura (2017, p. 5) présente également une variante pseudo-clivée pour CS-VIII à la manière dont son dérivée CS-III et CS-IV de CS-I et CS-II respectivement. Nous la nommons CS-IX :

1. **NSS** + virgule + c’/ce être + syntagme nominal :  
   Notre **objectif** majeur, c’est la rédaction d’une proposition de loi.

Les NSS étant une classe fonctionnelle et non lexicale, les différentes constructions spécificationnelles traditionnelles sont autant de définitions opératoires des NSS (Schmid 2000). L’annotation manuelle d’un grand corpus pour détecter de tels emplois n’est pas envisageable. L’annotation purement automatique à partir d’un repérage structurel semble moins difficile mais présente de sérieuse difficulté, notamment pour les CS les plus ouvertes comme la CS-V.

Nous proposons comme solution une présélection automatique suivie d’une évaluation manuelle des résultats pour déterminer s’il s’agit bien d’un NSS. La seconde étape impose que la présélection automatique soit la plus efficace possible pour restreindre le nombre de résultats à examiner manuellement.

Pour pouvoir utiliser le traitement automatique des langues en vue d’effectuer cette présélection automatique, il faut faire une analyse syntaxique préalable des différentes constructions spécificationnelles recensées. Le point qui nous semble le plus important est d’analyser la nature et la fonction du contenu spécifiant de chaque construction spécificationnelle.

#### B) Nature et fonction du contenu spécifiant dans les constructions spécificationnelles classiques

##### CS-I, CS-II, CS-III, CS-IV, CS-V et CS-VI : NSS [(, ce être) | être ] proposition en **que** ou **de**

Sur la nature du contenu spécifiant, pour CS-I et CS-II, nous suivons Legallois (2008) et Schmid (2000) en affirmant que le contenu spécifiant est avant tout une proposition subordonnée, pour CS-I une conjonctive commençant par *que*, et pour CS-II, une infinitive même si cela pose plus de questions. Roze et al. (2014) assimile la nature des CS-III et CS-IV, les pseudo-clivées, à CS-I et CS-II respectivement.

Pour déterminer la nature du contenu spécifiant des CS-II et CS-IV, la première hypothèse est de considérer qu’il s’agit d’une proposition subordonnée infinitive incluse dans un syntagme prépositionnel commençant par *de*. La seconde hypothèse de considérer l’ensemble, en y incluant *de*, comme une seule proposition subordonnée infinitive. La préposition *de* joue alors le même rôle de complémenteur subordonnant que *que* (Huot, 1981 ; Kalmbach, 2019, p. 675).

À l’appui de la seconde hypothèse, Kalmbach (2019, p. 675) indique qu’*« on peut facilement mettre en parallèle les deux types de construction »*, conjonctive et infinitive, par la paraphrase :

« le fait que le jury sélectionne les candidats » ⬄ « le fait de sélectionner les candidats ».

La principale différence soulignée par l’auteur entre les deux subordonnées est que *« le sujet [ici, le jury] n’est pas exprimé dans l’infinitive. Par rapport à la construction conjonctive, la construction infinitive prend donc une valeur impersonnelle ou générale »*. Ce sujet implicite conforte la seconde hypothèse car la définition de la proposition en grammaire traditionnelle, rappelée par Joseph Donato dans l’ouvrage collectif sous la direction de Mounin (1974), stipule qu’il s’agit d’*« un groupe de mots qui a son propre sujet et son propre prédicat »*, ici implicite. Ce même auteur rappelle également que la *« distinction entre syntagme et proposition n'était pas toujours très claire ni très systématique »*.

Ainsi, par symétrie avec la proposition subordonnée conjonctive des CS-I et CS-III, nous parlerons donc de proposition subordonnée infinitive pour la nature du contenu spécifiant des CS-II et CS-IV. Nous privilégions donc deux appellations se référant à la catégorie morphosyntaxique d’un terme distinctif de chaque construction : la conjonction de coordination que dans un cas, l’infinitif dans l’autre.

Les constructions CS-V et CS-VI, qui reprennent CS-I et CS-II mais sans le verbe être, sont rejetées par la majorité des travaux français comme constructions spécificationnelles. Schmid (2000, p. 20), adoptant le point de vue de la grammaire traditionnelle sur cet aspect, ne statue pas (p. 23) entre *noun complements* et *appositive modifiers*. Legallois (2008, p. 8) les rapproche des noms à compléments prépositionnels (NCP) de Riegel. Or Riegel (2006, p. 38) estime pour les CS-V qu’il s’agit de propositions attributives réduites, par rapport à la proposition attributive copulative avec le verbe être et qu’elles sont *« les deux réalisations syntaxiques d’un même schéma prédictif, la première sous la forme d’une construction copulative, la seconde sous celle d’une configuration propositionnelle averbale »*. On peut en effet transformer les CS-I en CS-V par l’ajout du verbe copulatif. La même chose est possible pour les CS-II en CS-VI, nous nous permettons de reprendre le terme de proposition subordonnée infinitive. Pour Riegel, la fonction est donc celle d’attribut.

##### CS-VIII et CS-IX : NN [, ce] être syntagme nominal

La nature du contenu spécifiant est syntagme nominal pour les CS CS-VIII et CS-IX et la fonction est celle de complément attribut.

##### CS-VII : NSS **de** syntagme nominal

La construction CS-VII n’a pas été reprise en français. Depuis l’anglais, si le noyau du syntagme est un verbe au gérondif, on peut la rapprocher de la construction CS-VI pour le français que nous verrons ci-dessous. Mais si son noyau est un nom de type DDAA, on considère que le contenu spécifiant est un syntagme prépositionnel avec *de* comprenant un syntagme nominal, que nous appelons syntagme prépositionnel-nominal. On peut alors la rapprocher des constructions CS-VIII et CS-IX du fait que le contenu spécifiant soit aussi un syntagme nominal mais l’articulation avec le NSS se fait par le verbe être dans CS-VIII et CS-IX.

La fonction du syntagme prépositionnel-nominal dans CS-VII se rapproche formellement de celle de complément du nom. Néanmoins, le fait que le contenu spécifiant soit de la même nature que pour CS-VII et CS-VIII, l’équivalence entre le DDAA noyau de ce syntagme et le verbe dénotant la même action et activité dans les constructions avec une proposition, le fait qu’il puisse y avoir des constructions attributives non copulatives (Riegel, 2006), le fait enfin que le contenu spécifiant remplisse le NSS, tout cela rapproche le contenu spécifiant de la fonction attribut.

Un test possible est l’appel au contenu : un véritable complément de nom peut être supprimé alors qu’un attribut est essentiel à la phrase. On peut comparer les deux suppressions qui suivent :

* « le chat de Julie est blanc » vs « le chat est blanc »
* « l’**objectif** de rédiger la loi est important » vs « l’objectif est important »

L’*« attente de spécification »* est plus forte pour le NSS *objectif* qui *« appelle un complément d’information » (Huyghe, 2018, p. 45)*. Sur le plan sémantique, *chat* dénote une entité de premier ordre, alors qu’objectif dénote une entité de troisième ordre. Néanmoins on pourrait toujours affirmer que la différence d’attente de spécification entre les deux est une affaire de degré, de quel chat parle-t-on, plutôt qu’une dichotomie franche entre attente et non-attente. Sur le plan syntaxique, *chat* est en revanche incapable d’accepter un contenu propositionnel ce qui le disqualifie comme NSS.

Cela amène à considérer le contenu spécifiant comme obligatoire. Si l’attribut est essentiel à une phrase, *\*Le problème est*, le complément du nom ne l’est pas : *Le problème de définir une nouvelle loi est complexe* vs *Le problème est complexe*. Mais ce deuxième cas n’est pas un NSS en lui-même. S‘il est précédé d’un emploi de *problème* comme NSS, il peut s’agir d’une reprise anaphorique du concept déjà formé, l’utilisation de l’adjectif démonstratif *ce* renforcerait cette reprise mais n’est pas obligatoire. Or, nous nous intéressons au moment précis où le contenu spécifiant est associé au NSS, non aux reprises qui, dans l’étroitesse d’un titre, nous semblent peu pertinentes. Nous pouvons donc, pour identifier les NSS, parler d’une obligation de complémentation du nom si l’on adopte ce point de vue.

Pour notre part, nous privilégions la relation de complément attribut car il s’agit bien de conférer une propriété à un nom, ici donc de le remplir sémantiquement, et elle doit être au moins une fois obligatoire : la fonction de cohérence devient caduque si un NSS est seulement repris en anaphore sans jamais avoir été utilisé dans une construction spécificationnelle. Le cas d’une définition extralinguistique par le contexte de communication n’est pas recevable dans le contexte des titres qui introduisent un sujet.

La question se pose de la possibilité de transformer la CS-VIII en CS-VII. Si l’on reprend l’exemple de Nakamura (2017, p. 2), *Notre objectif majeur est la rédaction d’une proposition de loi* on a :

* \*Notre objectif majeur la rédaction d’une proposition de loi
* Notre objectif majeur, la rédaction d’une proposition de loi,
* ?Notre objectif majeur de la rédaction d’une proposition de loi

La première phrase est agrammaticale par la suppression du verbe copule entre sujet et attribut. La seconde transforme l’attribut en apposition. La troisième regroupe le NSS et le contenu spécifiant en un seul syntagme nominal complexe, en utilisant la préposition *de*. On retombe alors sur le type de construction CS-VII de Schmid (2000, p. 26 ; 2018, p. 155) : *The notion of love.* Néanmoins on peut se demander si l’énoncé formé est grammatical et acceptable. Les exemples (27, 28) tirés de notre corpus montre que le NSS objectif peut s’insérer dans une CS-VII.

(27) L'objectif de **satisfaction** de victimes en droit pénal international

(28) Comment l'**objectif** de maîtrise des flux de polluants est-il traduit dans les critères de gestion à l'amont des eaux pluviales ? - Analyse des pratiques en France et à l'international

On remarque pour les deux exemples (27, 28), le noyau nominal du contenu spécifiant est également un DDAA. Nous pensons qu’il s’agit d’une bonne contrainte pour les constructions spécificationnelles CS-VIII et CS-IX, comme pour la CS CS-VII, même si des cas épineux subsistent comme dans l’exemple (38) discuté plus bas.

Pour finir, on peut donc construire le tableau d’équivalence (10) entre les différentes CS, même si nous ne plaçons pas les transformations entre CS-I, CS-III et CS-V et CS-II, CS-IV et CS-VI sur le même niveau que la paraphrase possible entre CS-VII et CS-VIII.

|  |  |
| --- | --- |
| CS-I, III **NSS** + [ce] + être + proposition subordonnée conjonctive | CS-V **NSS** + proposition subordonnée conjonctive |
| CS-II, IV **NSS** + [ce] + être + de+ proposition subordonnée infinitive | CS-VI **NSS** + de+ proposition subordonnée infinitive |
| CS-VII **NSS** + syntagme prépositionnel-nominal | CS-VIII **NSS** + être + syntagme nominal |

Tableau 10: Tableau d'équivalence entre construction copulative et réduire

Nous rassemblons donc tous les contenus spécifiants sous la bannière de la fonction complément attribut. Les quatre natures possibles pour les contenus spécifiants sont donc une proposition subordonnée conjonctive introduite par *que* (CS-I, CS-III et CS-V), une proposition subordonnée infinitive (CS-II, CS-IV et CS-VI) introduite par *de*, un syntagme prépositionnel-nominal (CS-VII) ou un syntagme nominal (CS-VIII et CS-IX). Pour ces deux derniers, le noyau nominal doit être un DDAA, ce qui revient à demander que le noyau du contenu spécifiant dénote toujours une action, soit par l’entremise du verbe conjugué de la proposition subordonnée conjonctive (CS-I, CS-III et CS-V), soit par l’entremise du verbe à l’infinitif de la proposition subordonnée infinitive (CS-II, CS-IV et CS-VI), soit par le DDAA des trois dernières constructions (CS-VII, CS-VIII et CS-IX). À présent que nous avons rappelé la définition des NSS et des CS qui les incluent et observer la nature et la fonction du contenu spécifiant, nous allons essayer de chercher les CS classiques dans notre corpus.

### II.2.3 Recherche des CS classiques dans notre corpus

L’annotation manuelle des NSS sur un grand corpus comme le nôtre n’est pas envisageable. Le seul moyen de trouver des NSS est de rechercher, à la manière de Legallois (2008) pour CS-I et CS-II et de Roze et al. (2014) pour CS-I, CS-II, CS-III et CS-IV, les occurrences de constructions spécificationnelles dans notre corpus, ce qui nous a fait adopter le terme de définition opératoire malgré la mise en garde de Schmid (2018, p. 5) de ne pas confondre définition et opérationnalisation, ce qui se traduit chez nous par la distinction faite entre définition théorique et définition opératoire.

#### A) Constructions avec une proposition subordonnée conjonctive CS-I, CS-III et CS-V

Contrairement aux travaux de Legallois (2008) et Roze et al. (2014), dans un contexte averbal comme les titres, nous ne pouvons faire l’économie de ne pas considérer les CS sans verbe être conjuguée comme dans les CS-VI et CS-VI. Nous cherchons donc les schémas suivant dans notre corpus :

***NSS[[5]](#footnote-6)*** nom commun **[(, ce** clitique sujet**[[6]](#footnote-7) être) | être** conjugué**] que** conjonction de subordination**\_**

Nous prenons l’exemple (e1) pour illustrer qu’un syntagme prépositionnel peut s’insérer entre le NSS et le contenu spécificationnel. Pour prendre en compte cette flexibilité, plutôt que des schémas linéaires, nous faisons correspondre nos schémas à des relations de dépendance.

(e1) Le **problème** de cette nouvelle présentation est qu’elle n’est pas satisfaisante.

La difficulté est ici d’écarter toutes les propositions relatives. Schmid (2000, p. 3) indique clairement qu’une proposition relative ne peut être employée dans une construction spécificationnelle et qu’il ne faut pas confondre le *que* pronom relatif du *que* conjonction de subordination. A priori, les deux ayant des étiquettes différentes dans Talismane, cela ne devrait pas poser de problème.

En recherchant ce schéma, nous trouvons 26 tires qui correspondent. Néanmoins, Talismane étiquette erronément des lemmes *que* comme conjonctions de subordination alors qu’il s’agit de pronoms relatifs. Sur un si faible nombre de résultats, nous pouvons manuellement les filtrer et ne gardons que trois titres, (29), (30) et (31) qui possèdent un nom en emploi sous-spécifié :

(29) Condamnation d’une société au paiement de ses cotisations volontaires obligatoires en l’absence de **preuve** que ces cotisations faisaient l’objet d’un emploi contraire au droit européen des aides d’État

(30) Bibliothèque implicite ou les **représentations** que les enseignants se font d'une culture humaniste

(31) Démystification de l'**idée** que le réseau d'aide informelle se délite

On remarque qu’il n’y a jamais de verbe *être* conjugué pour relier le NSS à la proposition subordonnée conjonctive. Néanmoins on peut facilement construire une telle phrase à partir du couple NSS / proposition comme par exemple *L’****idée*** *est que le réseau d’aide informelle se délite* pour valider qu’il s’agit bien d’un emploi sous-spécifié. On peut donc constater que ce schéma est très peu présent dans nos titres, même sans verbe être conjugué.

#### B) Constructions avec une proposition subordonnée infinitive CS-II, CS-IV et CS-VI

Pour ces CS, nous cherchons le schéma suivant :

**\_ *NSS*** nom commun **[(, ce** clitique sujet **être) | être** conjugué**] de** préposition **VINF\_**

L’exemple (e2) montre que la négation *ne pas* peut s’insérer entre le *de* et verbe à l’infinitif en plus d’avoir un syntagme prépositionnel entre le NSS et le contenu. Nous continuons pour cette raison à faire correspondre nos schémas à des relations de dépendance.

(e2) Le **problème** de cette nouvelle présentation est de ne pas satisfaire le client.

En cherchant ce schéma, nous trouvons 1 161 titres qui correspondent. Néanmoins, Talismane n’arrive souvent pas à correctement choisir le recteur de la préposition *de* entraînant des faux positifs comme dans l’exemple (32) où le dernier *de* devant « faire le genre » a pour recteur le premier nom *corps*.

(32) Le corps recteur des filles à l’épreuve des filières scolaires masculines. Le rôle des socialisations primaires et des contextes scolaires dans la manière de dépendant « faire le genre »

Nous ajoutons des filtres pour trouver des occurrences de véritables constructions spécificationnelles : suppression des 27 titres avec « en vue de + infinitif », correction de la mauvaise dépendance si on trouve un nom commun immédiatement avant le *de* qui précède l’infinitif, non inclusion des sept titres avec la forme *Grégoire* étiquetée comme un verbe à l’infinitif, des six titres faisant de même pour *Alexandre*, des 12 titres avec *bien-être* dont Talismane considère l’*être* comme un infinitif et des 34 titres où un nombre était considéré comme infinitif, on tombe à 1 075 résultats.

Il n’y a que neuf résultats avec le verbe être conjugué. Nous pouvons rapidement les parcourir manuellement. Parmi ces résultats, Il n’y a qu’une seule construction spécificationnelle avec une proposition subordonnée infinitive et le verbe être, l’exemple (33).

(33) Situation palestinienne : le plus grand **effort** de la CPI est de vaincre les passions

Pour les 1066 résultats sans le verbe être conjugué, nous décidons d’en tester manuellement 10 %, soit 107, pour avoir une estimation du nombre véritable de constructions spécificationnelles dans ces résultats. Sur les 107, 59 % ne sont pas des CS. Si on applique ce taux à nos 1 066 résultats, nos résultats ne comptent plus que 629 faux positifs et 437 utilisations estimées comme véritables de NSS. Parmi les CS trouvées, on peut citer les exemples (34), (35), (36) et (37).

(34) La **tentation** d’instituer des « Cours constitutionnelles régionales »

(35) **Possibilités** de réduire les émissions de gaz à effet de serre et d’autres impacts environnementaux dans les systèmes de production de viande bovine

(36) L'**obligation** de renégocier le contrat au nom de la lutte contre les gaz à effet de serre

(37) Réversibilités post-coloniales : les mobilités d’**art** de vivre à Marrakech

On peut également construire une phrase à partir du couple NSS / proposition avec le verbe être comme par exemple *La tentation est d’instituer des « Cours constitutionnelles régionales »* pour valider qu’il s’agit bien d’une construction spécificationnelle. On remarque que pour les exemples (34), (35) et (36), le NSS est également une tête de segment ce qui va dans le sens d’un rapprochement.

#### C) CS avec verbe copule et syntagme nominal, CS-VIII et CS-IX

Le schéma pour les deux constructions spécificationnelles CS-VIII et CS-IX est le suivant :

**\_ *NSS*** nom commun **[, ce** clitique sujet**] être** verbe conjugué ***DDAA*** nom commun **\_**

DDAA signifie un nom déverbal dénotant une action ou une activité, néanmoins, nous laissons ouverte la possibilité d’avoir des noms n’étant pas des DDAA mais avec une action ou une activité implicite.

En cherchant le schéma correspondant à CS-VIII et CS-IX, toujours sur les relations de dépendances pour permettre une certaine flexibilité comme la présence de déterminants ou d’adjectifs, nous trouvons 226 résultats. Manuellement, nous éliminons des résultats qui nous semblent incorrects. Si les erreurs de Talismane comme confondre le verbe être avec le point cardinal *est* sont triviales, distinguer un emploi sous-spécifié ne l’est pas toujours. Si écarter les *full content nouns* comme *remariage, miroir* ou *misère* ne pose pas de problème, d’autres exemples se révèlent plus ardus à l’inspection du jugement intuitif, à défaut d’heuristique plus déterminante. Une technique consiste à essayer de paraphraser l’énoncé en une CS-I ou CS-II qui sont plus restrictives dans leurs syntaxes, ou CS-III et CS-IV, mais cela n’est pas toujours évident.

Nous retenons, sur les 226 résultats, un seul titre utilisant le pronom de reprise :

(38) Le plus grand **danger** social, c'est le bandit imberbe. La justice des mineurs à la Belle Époque

On voit bien ici que danger va créer un concept temporaire, que le locuteur caractérise sous un jour négatif, qui encapsule *le bandit imberbe*. Une phrase équivalente en CS-III serait *Le plus grand danger social, c’est que le bandit imberbe existe/menace/rôde*. On voit bien avec la phrase équivalente que nous retombons dans le cas où le nom noyau n’est pas un DDAA mais sous-tend de manière implicite une activité, le fait que le bandit existe.

Sur nos 226 résultats, un seul titre correspond à un emploi sous-spécifié :

(39) L'activité d'évaluation et les systèmes d'information. L'**évaluation** est aussi un travail langagier, assisté, organisé

La tête *évaluation* figure parmi nos têtes transdisciplinaires.

Les exemples (40) et (41) ont été rejeté car

(40) La **connaissance** est un réseau : perspective sur l’organisation archivistique et encyclopédique

(41) La **théorie** des chances n'est pas un jeu d'esprit : le statut de la probabilité mathématique selon Cournot

On touche ici aux limites de la capacité de jugement intuitive pour cette tâche : on ne peut pas dire que *travail*, *réseau* et *un jeu d’esprit* remplisse complètement le NSS sur le plan sémantique. On effectue les tests suivants :

* Est-il possible d’utiliser le nom dans une CS-I ou une CS-II : si oui, il y a un potentiel de portage d’une proposition, mais cela ne prouve que l’emploi étudié en est un.
* Est-ce que le nom du syntagme nominal est un DDAA ?

Si l’on répond par l’affirmative à ces deux réponses, nous sommes bien présence d’un NSS. Ce qui est certain, c’est le faible nombre de résultats qui ont le potentiel de correspondre à un emploi en NSS. Sur les dix exemples de NSS que nous avons extraits, *preuve,* ***représentation****, idée, effort, tentation, possibilité, obligation, art, danger,* ***évaluation*** deux sont des têtes transdisciplinaires. Nous savons donc que 2 % de nos 94 têtes transdisciplinaires admettent un emploi sous-spécifié néanmoins cet emploi semble rare : deux occurrences sur les 94 738 que comptent les têtes transdisciplinaires. Il reste enfin la dernière CS classique.

#### D) CS avec un syntagme prépositionnel-nominal, CS-VII

Cette dernière construction utilise le schéma suivant, très généraliste, où le premier NC est employé en NSS, que nous recherchons sur les liens de dépendance pour permettre une plus grande flexibilité :

**\_*NSS*** nom commun **de** préposition ***DDAA*** nom commun**\_**

Schmid (2018, p. 115) indique que son étude de 2000 n’a pas pris la définition CS-V-N pour des raisons techniques, car elle rapporte trop de résultats et avec beaucoup de bruits, comme les relations partie-totalité comme dans *le cœur du problème*. Or nos titres étant majoritairement averbaux, nous avons bien plus de chance d’y trouver des contenus spécifiants nominaux. Ne pas retenir cette construction nous ferait perdre de nombreux exemples. De plus, nous n’avons, pour une recherche automatique, pas d’autres alternatives qu’une définition structurelle. Nous n’avons néanmoins pas de moyen d’identifier les noms de type DDAA à notre disposition. Dans notre corpus, cette requête rapporte 179 931 résultats ce qui est beaucoup trop large pour permettre ensuite une sélection manuelle.

Une possibilité, reprise de Roze et al. (2014, p. 8) est de contraindre la requête en prenant le résultat uniquement si le premier nom appartient à un lexique. Roze et al. utilisent ce procédé pour découvrir de nouvelles CS en fournissant un lexique de noms déjà identifiés dans un ensemble de CS de type CS-I, CS-II, CS-III et CS-IV. Nous voulons faire de même avec nos têtes transdisciplinaires.

## II.3 Schémas récurrents d’emploi des têtes transdisciplinaires

### II.3.1 Recherche des schémas d’emplois des têtes transdisciplinaires

Du fait que les NSS sont une classe ouverte, et que les définitions varient d’un auteur à l’autre, aucune liste définitive n’est possible. Schmid liste 670 *shell nouns* (2000, p. 381), Flowerdew et Forest (2015), 845 *signalling nouns*, et Tutin (2008, p. 3), 356 *noms sous-spécifiés*. Néanmoins, Schmid (2018, p. 118) souligne la convergence de sa liste avec celle de Flowerdew et Forest (2015) sur les NSS les plus fréquents malgré leurs différentes méthodes. Ces listes peuvent donc servir d’indices, mais en aucun cas de preuves, pour prendre en compte le potentiel d’emploi sous-spécifié de nos têtes transdisciplinaires.

Sur les 94 têtes transdisciplinaires, 23 sont reconnues comme pouvant être un NSS par Legallois (2008, p. 3), soit seulement 24 %. Cependant, la définition opératoire de Legallois repose uniquement sur les CS CS-I et CS-II, et son corpus, les articles de l’année 1995 du quotidien *Libération*, est très éloigné du nôtre. Or, une définition opératoire est toujours dépendante du corpus sur lequel elle est appliquée.

Sur les 94 têtes transdisciplinaires, 83 ont un lemme dont la traduction en anglais apparaît dans la liste de Flowerdew et Forest (2015), soit 88 %. Son corpus est beaucoup plus proche de notre matériau, puisqu’il s’agit du Flowerdew Corpus of Academic English (Flowerdew et Forest, 2015, p. 68) composé de journaux académiques, de discours et de leçons. Cela nous amène à vouloir chercher les schémas récurrents des têtes transdisciplinaires dans nos titres.

#### A) Recherche de schémas émergents via la fouille de données

L’existence de nos têtes transdisciplinaires, fréquentes, abstraites, dotées d’un faible contenu sémantique, le fait que 83 % d’entre elles apparaissent dans la liste des signalling nouns, nous pousse à nous demander s’il n’existerait pas d’autres constructions spécificationnelles, propres aux titres. Nous allons à présent essayer de rechercher des schémas récurrents dans lesquels s’inséreraient nos têtes transdisciplinaires et d’évaluer si ceux-ci pourraient jouer le rôle de construction spécificationnelle.

La question se pose de distinguer les schémas récurrents des têtes transdisciplinaires des autres. Pour cela, il faut faire un détour par la fouille de données séquentielles et ses notions de *séquences* et de *motifs*. Nous reprenons directement une méthode formulée par Roze et al. (2014, p. 8), qui s’inspiraient déjà de Quiniou et al. (2012). Tout d’abord, nous construisons des séquences de mots à partir des titres autour des têtes nominales transdisciplinaires ou non. Nous écartons les têtes non nominales ainsi que les têtes non transdisciplinaires noms propres car nous voulons comparer les têtes transdisciplinaires avec celles qui ne le sont pas, or, nos têtes transdisciplinaires sont exclusivement des noms communs. Les comparer à une autre catégorie lexico-syntaxique ferait intervenir d’autres paramètres qui sortent du périmètre de cette étude.

Une séquence est une suite d’items ou d’ensembles d’items ordonnés. Nous calculons toutes les séquences existantes autour des têtes en utilisant une taille minimale de deux mots et une taille maximale de cinq mots. Chaque séquence est constituée d’items qui peuvent être :

* l’étiquette PONCT pour les marques de ponctuation, sauf s’il s’agit d’un marque segmentatrice, où on aura SEG
* le lemme pour un mot si c’est une préposition, une conjonction de coordination, d’une conjonction de subordination, ou le verbe *être* ou *avoir*,
* son étiquette morphosyntaxique pour les autres classes de mot, selon la nomenclature de Talismane reprise dans l’annexe **A4.1 Catégories morphosyntaxiques de Talismane** dont on rappelle les principales étiquettes : NC pour nom commun, DET pour déterminant, ADJ pour adjectif, NPP pour nom propre, VPP pour verbe au participe passé.
* Lorsqu’il s’agit d’une tête transdisciplinaire, l‘item sera TransHead (forme abrégée : TT) au lieu de son étiquette NC.
* Lorsqu’il s’agit d’une tête non transdisciplinaire, l’item sera Head (forme abrégée : HH) au lieu de son étiquette NC.
* Nous ajoutons les items INIT pour le début du titre et END pour sa fin.

L’exemple (42) possède deux têtes, la première, *problème*, est transdisciplinaire, l’autre, *région,* ne l’est pas.

(42) Les **problèmes** d'environnement dans une région d'extraction pétrolière : la **région** de Nijnevartovsk situé sur le territoire Khanti-Mansi (Russie)

Nous allons extraire pour chaque tête les séquences suivantes avec un taille minimum de deux et une taille maximum de 5 :

|  |  |  |  |
| --- | --- | --- | --- |
| **Longueur** | **Séquences extraites pour *problème*** | **Longueur** | **Séquences extraites pour *région*** |
| 2 | DET TT | 2 | DET HH |
| 2 | TT de | 2 | HH de |
| 3 | INIT DET TT | 3 | SEG DET HH |
| 3 | DET TT de | 3 | DET HH de |
| 3 | TT de NC | 3 | HH de NPP |
| 4 | INIT DET TT de | 4 | ADJ SEG DET HH |
| 4 | DET TT de NC | 4 | SEG DET HH de |
| 4 | TT de NC dans | 4 | DET HH de NPP |
| 5 | INIT DET TT de NC | 4 | HH de NPP VPP |
| 5 | DET TT de NC dans | 5 | NC ADJ SEG DET HH |
| 5 | TT de NC dans DET | 5 | ADJ SEG DET HH de |
|  |  | 5 | SEG DET HH de NPP |
|  |  | 5 | DET HH de NPP VPP |
|  |  | 5 | HH de NPP VPP sur |

Tableau 11 : Séquences extraites de l'exemple (42)

Nous répartissons les séquences dans deux bases : d’un côté, la base transdisciplinaire qui contient les séquences avec une tête transdisciplinaire, et de l’autre la base commune qui contient les séquences qui ont pour pivot une tête non transdisciplinaire.

Nous calculons ensuite le support de chaque séquence. Le support de la séquence S, c’est le nombre de séquences qui la contiennent dans une base donnée. Une séquence S est contenue dans une séquence S’ si on y retrouve tous les items de S, toujours dans le même ordre, mais éventuellement de façon disjointe. Ainsi, la séquence **DET** **TT** est contenue dans la séquence **DET** **TT** de, **DET** ADJ **TT** de, **DET** NC PONCT **TT**. Le support relatif est le support divisé par le nombre de séquence dans la base. Selon Roze et al. (2014, p. 383), *« un motif fréquent est une séquence dont le support est supérieur ou égal à un seuil fixé »* que nous fixerons ultérieurement.

Nous faisons deux extractions, EX1 et EX2, de séquence et de calculs de support. Pour la première extraction, nous utilisons une taille minimale de deux et une taille maximale de trois. Pour la seconde, nous utilisons la même taille minimale et une taille maximale de quatre. Par cette taille maximale limitée, nous souhaitons avant tout étudier l’environnement immédiat des têtes. Le tableau (12) présente les résultats des deux extractions.

|  |  |  |
| --- | --- | --- |
|  | **EX1 : séquence de 2 à 3 items** | **EX2 : séquence de 2 à 4 items** |
| Nombre de séquence avec pour pivot une tête transdisciplinaire | 510 980 | 919 764 |
| Nombre de séquences avec pour pivot une tête non transdisciplinaire. | 992 815 | 1 787 067 |

Tableau 12 : Nombre de séquences pour les deux bases de chaque extraction

Les motifs émergeants sont *« les motifs dont le support augmente de manière significative d’un ensemble de données à un autre »* (Roze et al., 2014, p. 383), ce qui se traduit par un taux de croissance supérieur à une valeur *p* que nous fixons à un pour être le plus large possible. Tout motif transdisciplinaire dont le support est supérieur au support du motif commun correspondant sera sélectionné. Notre de but est de trouver des motifs qui seraient propres aux têtes transdisciplinaires et fréquents. Pour cela, nous écartons à chaque fois les motifs dont le support est inférieur à 0,1 % du nombre de séquences dans la base transdisciplinaire, soit 511 pour EX1 et 920 pour EX2.

#### B) Découvertes de schémas émergeants liés à une position particulière des têtes

Notons que la technique de la fouille de données séquentielles, applicable dans d’autres domaines scientifiques, ignore complètement le fonctionnement linguistique des items. Si la séquence S2 **DET** NC PONCT **TT** contient bien la séquence S1 **DET** **TT**, dans la première, DET est dépendant de NC alors que dans la seconde DET est dépendant de TT. La relation « est contenu dans » est vérifiée, et le support de **DET** **TT** augmenté, mais on ne peut pas dire que le syntagme nominal que l’on pourrait voir dans S1 se retrouve dans S2. De ce fait, notre exploration des résultats pour retourner des séquences et motifs de la fouille de données vers les schémas lexico-syntaxiques doit se faire à la lumière de nos connaissances linguistiques.

Dans le cadre de l’extraction EX1, où les séquences comportent entre deux et trois items, 13 motifs fréquents sont émergeants. Le tableau (13) liste les différents motifs, leurs supports et leurs taux de croissance. Lorsqu’un lemme est présent dans les items, sa catégorie est également donnée après un double double-point.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **N°** | **Longueur** | **item 1** | **item 2** | **item 3** | **Support** | **Croissance** |
| 1 | 3 | SEG | TransHead | à::P+D | 757 | 7,01 |
| 2 | 3 | SEG | TransHead | à::P | 1102 | 4,73 |
| 3 | 3 | SEG | TransHead | sur::P | 1264 | 4,48 |
| 4 | 3 | INIT | TransHead | sur::P | 1613 | 2,83 |
| 5 | 3 | TransHead | sur::P | DET | 3744 | 2,57 |
| 6 | 2 | TransHead | sur::P | *(aucun)* | 12788 | 2,20 |
| 7 | 3 | SEG | TransHead | de::P+D | 2188 | 1,43 |
| 8 | 3 | TransHead | ADJ | sur::P | 762 | 1,43 |
| 9 | 3 | SEG | TransHead | de::P | 4856 | 1,37 |
| 10 | 3 | DET | TransHead | sur::P | 664 | 1,20 |
| 11 | 3 | ADJ | SEG | TransHead | 5800 | 1,18 |
| 12 | 3 | NC | SEG | TransHead | 7994 | 1,04 |
| 13 | 3 | SEG | TransHead | NC | 633 | 1,01 |

Tableau 13 : Motifs émergeants avec une longueur maximale de séquence de trois items

Nous écartons d’emblée le 13e motif car il s’agit d’une erreur de Talismane qui pousse à reconnaître comme un nom commun ce qui est un adjectif comme dans l’exemple (43) où *archéozoologique* est reconnu à tort comme nom commun.

(43) Le Plan Saint-Jean (Brignoles) : étude archéozoologique

Pour les autres motifs, on peut remarquer un point principal : la tête transdisciplinaire est très fréquemment en début du titre (motif 4), après INIT, ou après une ponctuation segmentatrice SEG (motifs 1, 2, 3, 7, 9). Néanmoins les motifs correspondants à INIT TransHead et SEG TransHead, avec un taux de croissance respectivement 0,47 et de 0,96, ne sont pas spécifiques aux têtes transdisciplinaires. C’est la conjonction de cette position avec la préposition *sur*, pour les têtes en positions initiales (motif 4), et avec les prépositions *à*, *sur* et *de* pour les têtes en immédiatement après un signe de ponctuation segmentant (motif 1, 2, 3, 7, 9), qui rend ces motifs spécifiques aux têtes transdisciplinaires. Notons que l’utilisation de la préposition *sur* à la suite d’une tête transdisciplinaire est déjà un signe distinctif de ces têtes par rapport aux têtes non transdisciplinaires (motif 6). L’emplacement privilégiée des têtes transdisciplinaires après un signe de ponctuation segmentant se rapproche de notre travail de première année, mais par un autre cheminement.

Un comptage sur notre corpus confirme ce que nous avons trouvé par la fouille de données : 19 235 têtes transdisciplinaires sont après un signe de ponctuation. Et 18 638 de ces signes, soit près de 97 %, sont des signes segmentants.

On remarque que les prépositions sont éventuellement fusionnées avec un déterminant comme *du* ou *au* (motifs 1 et 7) Selon la syntaxe, ces motifs doivent être suivis par un nom.

Cela signifie que la tête transdisciplinaire est majoritairement en tête de son segment, soit du premier segment et donc en tête du titre (motif 4), soit du second segment (motifs 1, 2, 3, 7, 9, 11, 12). Nous donnons ci-dessous des exemples de titres avec le motif correspondant.

(44) Vers le Web Socio Sémantique **: introduction aux** ontologies sémiotiques motif 1

(45) Investitures et rapports de pouvoirs **: Réflexions sur** les symboles de la Querelle en Empire motif 2

(46) Conception Isotropique d'une morphologie parallèle **: Application à** l'usinage motif 3

(47) **Remarques sur** le polythéisme étrusque motif 4

Dans le cadre de l’extraction EX2, où les séquences comportent entre deux et quatre items, 43 motifs fréquents sont émergeants. Nous intéressons parmi ceux-ci aux 19 qui comptent quatre items, car c’est eux qui apportent des informations supplémentaires sur l’environnement immédiat des têtes transdisciplinaires, dans le tableau (14).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **N°** | **Longueur** | **item 1** | **item 2** | **item 3** | **item 4** | **Support** | **Croissance** |
| 1 | 4 | SEG | TransHead | à::P | DET | 930 | 7,75 |
| 2 | 4 | SEG | TransHead | sur::P | DET | 1219 | 4,63 |
| 3 | 4 | INIT | TransHead | sur::P | DET | 1557 | 3,36 |
| 4 | 4 | TransHead | sur::P | DET | NC | 3509 | 2,57 |
| 5 | 4 | ADJ | SEG | TransHead | de::P | 1524 | 1,70 |
| 6 | 4 | SEG | DET | TransHead | de::P+D | 4447 | 1,59 |
| 7 | 4 | SEG | TransHead | de::P+D | NC | 2047 | 1,45 |
| 8 | 4 | SEG | TransHead | de::P | DET | 2585 | 1,45 |
| 9 | 4 | NC | SEG | TransHead | de::P+D | 939 | 1,44 |
| 10 | 4 | SEG | TransHead | de::P | NC | 2120 | 1,39 |
| 11 | 4 | NC | SEG | TransHead | de::P | 2007 | 1,34 |
| 12 | 4 | de::P | NC | SEG | TransHead | 1598 | 1,31 |
| 13 | 4 | SEG | DET | TransHead | de::P | 6819 | 1,24 |
| 14 | 4 | NPP | SEG | TransHead | de::P | 925 | 1,19 |
| 15 | 4 | NC | ADJ | SEG | TransHead | 4524 | 1,19 |
| 16 | 4 | INIT | TransHead | de::P | DET | 7461 | 1,15 |
| 17 | 4 | ADJ | SEG | DET | TransHead | 6174 | 1,12 |
| 18 | 4 | NC | SEG | TransHead | ADJ | 1474 | 1,05 |
| 19 | 4 | ADJ | SEG | TransHead | ADJ | 987 | 1,05 |
| 20 | 4 | de::P+D | NC | SEG | TransHead | 1087 | 1,00 |

Tableau 14 : Motifs émergeants pour les séquences de longueur de quatre items

On voit apparaître dans cette seconde extraction que les motifs à prépositions non fusionnées sont suivis par un déterminant, ce qui appelle un nom (motifs 1, 2, 3, 4, 8, 16). On voyait déjà apparaître le déterminant lorsqu’il était fusionné avec la préposition dans EX1, comme ici avec les motifs 6, 7 et 9 . Dans certains motifs, le nom apparaît directement, sans déterminant (motif 10) ou avec (motif 4).

Il apparaît également qu’avant la ponctuation qui précède la tête transdisciplinaire on trouve des noms communs (motifs 9, 11, , 12, 15, 18, 20) ou des noms propres (14).

Une autre information est la possibilité d’avoir un adjectif qui doit qualifier la tête transdisciplinaire (motifs 18 et 19) ainsi qu’un déterminant pour la tête (motifs 6, 13, 17).

Nous donnons ci-dessous des exemples de titres avec le motif correspondant.

(48) Classification floue généralisée **: Application à la** quantification de la stéatose sur motif 1  
 des images histologiques couleurs

(49) Word Wild West **: remarques sur les** glissements de forme et de sens du mot west motif 2

(50) **Réflexions sur l**'économie cubaine motif 3

#### C) Analyse des résultats : construction de deux schémas de recherche

Les motifs émergeants ont fait apparaitre des spécificités dans l’emploi des têtes transdisciplinaires :

* La position en début de segment, que cela soit en début de titre (premier segment) ou en début de second segment après un signe de ponctuation segmentant
* Une préférence à être suivie par les prépositions *à, sur, de* dans cette position.

Nos connaissances linguistiques nous permettent d’interpréter les motifs émergeants détectés, de les combiner entre eux, par exemple pour y intégrer des tokens optionnels comme un déterminant ou un adjectif, pour en tirer les deux schémas suivants :

**Schéma 1 : INIT** [DET] **TRANSHEAD** [ADJ] ((P à, sur, de [DET]) | P+D à, de) NC

**Schéma 2 : SEG** [DET] **TRANSHEAD** [ADJ] ((P à, sur, de [DET]) | P+D à, de) NC

Un schéma est défini par une séquence de tokens qui peut comporter :

* des choix entre plusieurs tokens, un choix entre A et B est noté **A | B**,
* des tokens optionnels, l’optionalité de A est notée **[**A**]**,
* des répétitions de tokens, la répétition du token A entre i et j fois est notée Ai-j.

Un token de schéma peut être :

* une classe grammaticale (nom commun - NC, adjectif qualificatif - ADJ, préposition - P, déterminant - DET, préposition et déterminant fusionnés P+D),
* un marqueur spécial comme INIT qui indique le début du titre, SEG qui indique un signe de ponctuation segmentant ou TRANSHEAD indiquant une tête transdisciplinaire de segment,
* un lemme (*et*),
* la combinaison d’un lemme et d’une classe grammaticale, notée CLASSE lemme.

Aucun token non prévu par la définition du schéma n’est autorisé lorsqu’on le recherche dans les titres. Une séquence de tokens dans un titre, mots ou signes de ponctuation, correspond à un schéma lorsque la séquence du titre se conforme à une des réalisations possibles définies par le schéma. Les réalisations étant l’ensemble des séquences possibles à partir de sa définition.

#### D) Rapprochement des deux schémas avec la CS-VII

Le fait le plus remarquable est que les deux schémas se rapprochent de la CS-VII, NSS de NC., en prenant l’hypothèse que les têtes transdisciplinaires sont sous-spécifiés. D’un côté, les schémas sont plus contraignants, sur l’emplacement dans le titre de la tête : uniquement en début au début d’un segment, éventuellement après son déterminant. D’un autre côté les schémas élargissent la CS-VII : ils autorisent *à* et *sur* comme préposition.

Si la capacité de portage propositionnel (Huyghe, 2018) est une propriété qui distingue les NSS, l’expansion du nom par un syntagme prépositionnel complément du nom commençant par *de* ne l’est pas. La préposition *de,* peut se rapprocher du *of* anglais*,* qui est, selon Schmid (2018, p. 115), *« highly polysemous and frequently encodes a possessive or part-whole relation rathen than one of identity »*.La relation sémantique *partie-tout* peut très bien s’exprimer en français de la même manière : *le capot de la voiture*. Pour déterminer s'il s'agit d'emplois sous-spécifiés, il faut à présent s'interroger sur la sémantique du nom commun présent à la fin de chacun des deux schémas.

### II.3.2 Lexique des noms et détermination de l’emploi

Dans notre corpus de travail, 36 664 titres correspondent à notre premier schéma, 18 175 à notre second schéma et 3 065 aux deux à la fois. On remarque que toutes les têtes transdisciplinaires apparaissent dans nos résultats. Cela fait un total de 60 969 emplois de têtes transdisciplinaires à analyser pour déterminer l’éventuelle sous-spécification, ce qui est inenvisageable de faire manuellement. Nous décidons de prendre un exemple avec la tête *problème.*

#### A) L’étude de problème

Le choix d’étudier la tête transdisciplinaire *problème* repose sur deux points :

* Il est un *prime shell noun* de Schmid (2000, p. 85 ; 2018 ; p. 118) et
* dans la liste de Flowerdew et Forest (2015, p. 203), il est le 3e par ordre de fréquence normalisée.

Il y a donc une forte probabilité que *problème* puisse avoir des emplois en NSS dans notre corpus. La tête *problème* est la 52e dans l’ordre de fréquence dans nos résultats des schémas 1 et 2. Selon notre méthode de sélection, elle est la 58e tête transdisciplinaire, sur 94, en les classant par valeur de médiane.

On restreint donc nos schémas en prenant pour TRANSHEAD uniquement les occurrences de *problème*. XXX

De plus, on lève dans ce cas la contrainte que TRANS soit une tête pour estimer s’il existe des énoncés qui correspondent à cette séquence où le lemme *problème* ne serait pas une tête de segment. TRANS peut donc correspondre à deux cas :

* *Lemme de problème non-tête de segment*
* *Lemme de problème tête transdisciplinaire*

On obtient 282 couples différents (TRANS, NC) dans 370 occurrences à analyser manuellement. Pour chacun, nous nous aidons du contexte en suivant les indices déjà évoqués pour la CS-VII : capacité de paraphrase en CS-I ou CS-II, nom étant un DDAA, nom rendant implicite l’action ou l’activité.

Nous obtenons 135 occurrences pouvant être reçues comme un emploi de *problème* en NSS directement selon la construction CS-VII, soit 36 % du total des occurrences de *problème* dans les résultats du schéma 1, comme les exemples (34) et (35).

(34) Quelques **problèmes** d'analyse de la délinquance juvénile à la fin du XIXe siècle. L'exemple parisien

(35) Le **problème** du regroupement des activités dans la modélisation ABC. Une approche possible

(36) **Problèmes** de création en multimédia : marier l'expérience de l'audiovisuel et la rigueur de la qualité

On peut à chaque fois paraphraser ces exemples en utilisant une autre CS. Pour l’exemple (35), on peut paraphraser par la CS-II, avec être suivi d’une préposition et d’un infinitif : *le problème est de regrouper*. Pour les exemples (34) et (36), on peut les paraphraser avec la CS-VI, qui reprend la CS-II sans le verbe être : *Quelques problèmes d’analyser, Problèmes de créer en multimédia*. *Analyse, regroupement, création* sont des déverbaux dénotant une action ou une activité.

Sur les 370 résultats, deux occurrences correspondent au schéma sans que *problème* ne soit une tête de segment : les exemples (37) et (38). Une seule de ces deux occurrences peut être reçue comme un emploi de *problème* en NSS, l’exemple (37) qui paraphrasé devient : *Le problème est d’interpréter les enclos quadrangulaires de La Tène moyenne découverts en Flandre française.*

(37) Problème d'interprétation des enclos quadrangulaires de La Tène moyenne **découverts** en Flandre française : l'exemple de Borre (Nord)

(38) Les problèmes **viennent** du fonctionnement du système, pas des individus. Entretien avec Maurice Godelier

Cela signifie que, sur les 1 660 occurrences de *problème* dans le corpus, nous avons 613 occurrences où il s’agit d’une tête de segment et 1 047 occurrences en tant que non-tête. Cela fait, en ne considérant que le schéma 1, un taux d’utilisation en tant que NSS de 0,1 % si le lemme n’est pas une tête, et de 22 % si le lemme est une tête.

Cette utilisation en tant que NSS est liée à la position en tant que tête, ce qui était prévisible car, dans le schéma 1, le TRANS est le premier nom rencontré dans le titre, or les têtes de titres monosegmentaux ou de premier segment de titres monosegmentaux, sont à  86 % le premier nom commun du titre.

Les 135 occurrences retenues utilisent toutes la préposition *de* même s’il existe également des occurrences non gardées qui utilisent une construction de la forme *le problème posé par X* comme dans l’exemple (39) et qui semblent correspondre à un emploi NSS dans 17 cas. L’exemple (39) peut en effet être paraphrasé en *Problèmes de prédire en persan*. Néanmoins, cet emploi est propre à *problème*, on ne peut pas mettre un autre lemme à la place, nous ne le retenons dans notre tentative de trouver une détermination générale de l’emploi NSS.

(39) Problèmes posés par la prédication en persan. Approche contrastive persan

Le résultat le plus marquant est la très forte présence de DDAA dans résultats dont le trait saillant est l’utilisation d’un suffixe. On catégorise leurs formations morphologiques dans le tableau (15).

|  |  |  |  |
| --- | --- | --- | --- |
| **Suffixe** | **Compte** | **%** | **Exemples** |
| -tion | 83 | 62 % | *évaluation, communication, compréhension* |
| -ment | 14 | 11 % | *financement, réordonnancement, désencastrement* |
| -age | 6 | 5 % | *réglage, échantillonnage, démarrage, remplissage* |
| -sion | 6 | 5 % | *expansion, inversion, compréhension, prévision, conversion* |
| Suffixe zéro | 15 | 11 % | *abandon, commande, enquête, transport, groupe, nuance, contrôle, calcul, rejet, analyse* |

Tableau 15: Suffixes associés au nom noyau du contenu spécifiant

Les 10 restants, soit 8 %, sont des déverbaux avec modification du radical, comme *compromis* 🡺 *compromettre*, *livraison* (x3) 🡺 *livrer*, *émergence* 🡺 *émerger (x2)*, *résistance* 🡺 *résister*, *gouvernance* 🡺 *gouverner*, *ordre* 🡺 *ordonner* ou un gérondif anglais importé directement en français, *routing,* retrouvant ainsi les exemples initiaux de la CS-VII.

Sauf pour les 10 derniers et les 15 déverbaux avec zéro suffixe, on peut donc construire un filtre sur les terminaisons les plus fréquentes qui indiquent un DDAA : « -tion » et sa variante « -sion », « -age », « -ment » pour mieux sélectionner les titres avec un usage potentiel de NSS. De plus, on peut rétrécir le schéma à la préposition *de*, pour rapprocher notre schéma 1 le plus possible de la CS-VII, laissant de côté les variantes possibles.

Nous exécutons notre requête de façon disjointe encore une fois, mais en n’autorisant aucune autre proposition que *de* entre le TRANS et le NC. Nous obtenons 117 résultats. Ce filtre étant construit par l’ajout de contraintes sur le précédent, il ne peut pas y avoir de résultats qui n’auraient pas été couverts par le précédent. Nous pouvons donc calculer une précision, nous avons 7 faux positifs donc une précision de 94 %, et un rappel, de 81 %, de notre nouvelle définition du schéma par rapport à l’ancienne. Les 7 faux positifs bruitant nos résultats sont dus à la manière disjointe de faire correspondre l’énoncé à notre schéma pour 6 d’entre eux, comme pour l’exemple (37) et par l’utilisation d’un mot ayant une des terminaisons ciblées pour 1 d’entre eux, l’exemple (38) avec *environnement*.

(37) Quelques problèmes éditoriaux autour de L’Éducation sentimentale

(38) Les problèmes d'environnement dans une région d'extraction pétrolière : la région de Nijnevartovsk situé sur le territoire Khanti-Mansi (Russie)

Étudier *problème* nous amené à constater que :

* L’utilisation en tant que NSS semble liée au fait que le mot soit tête de segment, reste à déterminer l’éventuelle corrélation entre les deux faits.
* L’utilisation en tant que NSS peut être estimée en analysant la morphologie du nom après la préposition *de.*
* Il pourrait exister des constructions spécificationnelles propre à chaque nom, comme *le problème posé par*, comme Nakamura (2017, p. 7) signale *avoir pour/comme objectif de*.

#### B) Estimation globale à l’aide du schéma 1

À partir du schéma 1 modifié, nous pouvons estimer l’usage en tant que NSS des têtes transdisciplinaires. La détermination d’un emploi sous-spécifié dans une construction aussi large que la CS-VII étant une affaire d’interprétation sémantique, nous ne pouvons qu’admettre que notre schéma ne sera au mieux qu’une estimation large. Néanmoins, nous pouvons comparer pour chaque lemme des têtes transdisciplinaires le nombre d’emplois correspondant à notre schéma 1, c’est-à-dire où le nom suivant se termine en -tion, -sion, -age, -ment, et si cet emploi intervient si le lemme est bien une tête. Nous différencions donc à chaque fois :

* *Lemme de tête transdisciplinaire qui n’est pas tête de segment (exemple 37 et 38)*
* *Lemme de tête transdisciplinaire effectivement tête transdisciplinaire*

Nous obtenons, sur l’ensemble de notre corpus, 12 124 correspondances avec notre schéma 1. Seulement 52 sont une correspondance alors que le lemme n’est pas une tête, soit 12 072 où nous avons une tête transdisciplinaire estimée employée comme NSS. Nous comptons en tout 94 738 occurrences de têtes transdisciplinaires dont 58 003 dans le premier segment, les seules pouvant correspondre au schéma 1. Sur ces têtes transdisciplinaires de premier segment, nous en avons donc 21 % estimées employées comme NSS, soit dans 23 % des titres. Ce même emploi ne se retrouve qu’à 0,09 % lorsque le lemme n’est pas une tête. On peut également se poser la question si cette caractéristique est propre aux têtes transdisciplinaires. Nous recherchons les correspondances du schéma 1 avec cette fois-ci les têtes qui ne sont pas transdisciplinaires. On obtient 17 051 résultats, soit 11 % des têtes non transdisciplinaires. Les têtes que nous n’avons pas sélectionnées comme transdisciplinaires ont également la capacité être employée comme NSS, bien que moins fréquemment.

#### B) Estimation globale à l’aide du schéma 2

Nous refaisons le même cheminement pour le second schéma en restreignant directement la préposition à *de*. Nous obtenons les résultats suivants :

* Têtes transdisciplinaires : 2 928 emplois de NSS
* Lemmes de têtes transdisciplinaires non-têtes :  1 773 emplois de NSS

Ce résultat peut sembler paradoxal : le fait d’être employé en tant que NSS semble beaucoup moins corrélé au fait d’être une tête. Cela s’explique en analysant ces résultats, comme les exemples (39) et (40).

(39) La connaissance, un **outil** de la prévention de la délinquance

(40) Caractérisation et génération de surfaces agricoles – **Etude** de la diffraction en coordonnées non orthogonales

Les deux utilisent des marques de ponctuation non reconnues comme segmentatrices par notre définition, la virgule pour (39) et le tiret pour (40). Nous pensons qu’il s’agit pourtant bien ici de titres bisegmentaux ayant pour tête de second segment *outil* et *étude*, qui sont tous les deux des lemmes de têtes transdisciplinaires. La forte proportion d’emploi NSS de lemmes non-têtes provient donc d’une mauvaise segmentation initiale des titres.

Nous prenons comme hypothèse que le schéma 2 ne peut se trouver que dans un second segment de titre, écartant ainsi les titres mal segmentés. Si on fait ces requêtes uniquement sur les titres bisegmentaux, ce qui revient à contraindre que la marque de ponctuation soit segmentatrice, on obtient les résultats qui suivent :

* Têtes transdisciplinaires : 3 626 emplois de NSS
* Lemmes de têtes transdisciplinaires non-têtes :  1 395 emplois de NSS

Nous avons 36 731 occurrences de têtes transdisciplinaires en second segment qui ont une proportion à hauteur de 10 % à avoir un emploi de NSS. Sur les 47 335 autres occurrences de têtes nominales non transdisciplinaires de second segment, on estime à 3 166 les emplois en tant que NSS, soit une proportion de 7 %.

#### C) Tests de corrélation entre emploi NSS et le fait d’être une tête transdisciplinaire

En ne comptant que notre estimateur fondé sur les schémas 1 et 2 pour détecter les emplois NSS, car les autres CSS donnent en tout moins de 500 résultats et n’influent qu’à la marge, nous avons les informations suivantes :

|  |  |  |  |
| --- | --- | --- | --- |
| **Corpus** | **Nombre de têtes / dont emploi NSS** | **Nombre de têtes transdisciplinaires / dont emploi NSS** | **Nombre de têtes non transdisciplinaire / dont emploi NSS** |
| Nombre de titres monosegmentaux | 147 828 |  |  |
| Nombre de titres monosegmentaux dont la tête est un nom commun | 136 734 |  |  |
| Nombre de titres bisegmentaux | 103 170 |  |  |
|  |  |  |  |
| Nombre de 1er segment dont la tête est un nom commun | 136 734 + 79 959 = **216 693 /29 123** | 58 003  / 12 072  21 % | 158 690  / 17 051  11 % |
| Nombre de 2nd segment dont la tête est un nom commun | **84 066**  **/ 6 792** | 36 731  / 3 626  10 % | 47 335  / 3 166  7 % |
| **Total** | **300 759** ↑→  **/ 35 915** | **94 734**  **/ 15 698  17 %** | **206 025**  **/ 20 217  10 %** |

On peut calculer la corrélation entre deux associations :

* Têtes transdisciplinaires et estimation d’emploi NSS : 0,35
* Têtes non transdisciplinaires et estimation d’emploi NSS : 0,18

On rappelle que plus un coefficient de corrélation est proche, en valeur absolue, de 1, plus les deux variables sont liées, positivement ou négativement selon la valeur du coefficient. On voit que le fait d’être une tête transdisciplinaire augmente ce coefficient, les têtes transdisciplinaires ont plus de chance d’être NSS, mais qu’il demeure trop faible pour qu’une véritable corrélation soit établie entre les deux notions.

#### D) Tests de corrélation entre emploi NSS et le fait d’être une tête

On peut néanmoins vouloir savoir si le fait, pour un nom commun, d’être une tête est corrélée au fait d’être un emploi en NSS. Il y a 1 118 481 de noms communs dans nos titres. Nous avons qu’il y a sur ceux-ci 300 759 têtes dont 35 915 sont estimées être des emplois NSS. Nous faisons donc le coefficient de corrélation entre :

* Tête et estimation d’emploi NSS : 0,30

Là aussi, le coefficient de corrélation est trop faible pour qu’un véritable lien soit fait entre le fait pour un mot d’être tête de segment et le fait d’être employé en tant que NSS.

#### E) Tests de corrélation entre emploi NSS et problème

Devant ce manque de résultats, nous revenons à *problème*qui compte 1 660 occurrences dans nos titres sur 1 118 481 occurrences de noms communs. On regarde, indépendamment que l’occurrence soit tête ou pas, si notre estimateur le classe comme potentiel NSS :

|  |  |  |
| --- | --- | --- |
|  | Occurrences | NSS |
| *problème* | 1 660 (dont 1226 têtes) / 1 118 481 | 185 (dont 167 têtes) |

Nous calculons la corrélation entre :

* Fait pour une occurrence d’avoir pour lemme *problème* et d’être un NSS : 0,32.

Ce coefficient semble assez faible si l’on prend en compte le fait que problème soit un *prime shell nouns* et le 3e sur la liste de fréquence de Flowerdew et Forest (2015). Nous ne pouvons que constater :

* Soit notre estimateur d’emploi en NSS est mauvais, pourtant le cas d’étude sur *problème* avait montré une précision de 94 % et un rappel de 81 % en ce qui concerne le schéma 1. Nous n’avons pas estimé ces valeurs pour le schéma 2 mais la non-segmentation des titres sur certains caractères pourraient expliquer ce défaut, ainsi que la non prise en compte des déverbaux à suffixe zéro et des constructions particulières comme *le problème posé par*.
* Soit il n’y a effectivement pas de corrélation entre le fait d’être une tête transdisciplinaire et d’être employé de façon sous-spécifiée. La similitude entre la liste des têtes transdisciplinaires et les listes de NSS, 88 % des têtes transdisciplinaires apparaissent dans la liste de Flowerdew et Forest (2015), n’est que d’ordre lexicale. Ce qui est renforcé par le fait que les NSS sont une classe fonctionnelle, un emploi potentiel d’un lemme, et non une classe lexicale, même si des propriétés sémantiques en facilitent l’utilisation en tant que NSS.

### II.3.4 Transdisciplinarité des schémas

On peut chercher comment les deux schémas identifiés précédemment se répartissent dans les différents domaines de notre corpus. Nous mettons le pourcentage que cela fait par rapport aux têtes transdisciplinaires de ce domaine dans le premier segment pour le schéma 1, dans le second segment pour le schéma 2 et par rapport à toutes les têtes transdisciplinaires du domaine dans la troisième colonne. Nous mettons enfin l’ordre des domaines par rapport à ce pourcentage.

|  |  |  |  |
| --- | --- | --- | --- |
| **Domaine** | **Tête transdisciplinaire en emploi de NSS selon schéma 1** | **Tête transdisciplinaire en emploi de NSS selon schéma 2** | **Total tête transdisciplinaire en emploi de NSS** |
| **Anthropologie** | 121 22 % 17 | 73 9 % 13 | 194 11 % 20 |
| **Archéologie et Préhistoire** | 212 21 % 18 | 134 9 % 10 | 346 11 % 21 |
| **Architecture** | 125 26 % 11 | 56 9 % 12 | 181 14 % 15 |
| **Art et histoire de l'art** | 73 18 % 23 | 59 8 % 18 | 132 9 % 24 |
| **Chimie** | 116 20 % 20 | 20 7 % 23 | 136 14 % 14 |
| **Droit** | 521 19 % 21 | 124 7 % 21 | 645 13 % 18 |
| **Éducation** | 495 29 % 7 | 195 12 % 5 | 690 17 % 7 |
| **Gestion et management** | 1118 31 % 3 | 507 10 % 9 | 1625 16 % 10 |
| **Géographie** | 47 31 % 2 | 24 12 % 6 | 71 16 % 9 |
| **Histoire** | 253 19 % 22 | 199 8 % 19 | 452 9 % 23 |
| **Informatique** | 919 26 % 10 | 177 11 % 7 | 1096 19 % 2 |
| **Linguistique** | 484 26 % 9 | 272 12 % 4 | 756 14 % 13 |
| **Littératures** | 97 15 % 25 | 65 6 % 24 | 162 8 % 25 |
| **Mathématiques** | 128 20 % 19 | 20 9 % 14 | 148 15 % 12 |
| **Philosophie** | 124 18 % 24 | 53 7 % 22 | 177 10 % 22 |
| **Physique** | 2606 26 % 12 | 193 8 % 20 | 2799 21 % 1 |
| **Planète et Univers** | 165 24 % 15 | 42 9 % 15 | 207 15 % 11 |
| **Psychologie** | 139 30 % 5 | 51 12 % 3 | 190 18 % 4 |
| **Science politique** | 240 25 % 13 | 117 9 % 10 | 357 13 % 17 |
| **Sciences cognitives** | 158 32 % 1 | 67 12 % 2 | 225 18 % 3 |
| **Sciences de l'environnement** | 410 30 % 6 | 118 10 % 8 | 528 18 % 6 |
| **Sciences de l'information et de la communication** | 276 31 % 4 | 152 13 % 1 | 428 16 % 8 |
| **Sciences du Vivant** | 1414 25 % 14 | 228 8 % 16 | 1642 18 % 5 |
| **Sociologie** | 957 26 % 8 | 420 8 % 17 | 1377 13 % 16 |
| **Économie et finance quantitative** | 21 24 % 16 | 5 5 % 25 | 26 13 % 19 |

Nous constatons des écarts dans l’utilisation de ces schémas, et donc dans la présence de têtes transdisciplinaires en emploi sous-spécifié. L’étendue est de 13 et la moyenne de 14,36. La plus grande fréquence est en physique avec 21 %. On remarque des fréquences beaucoup plus faibles pour le schéma 2, ce qui indiquerait qu’il y a plus de têtes transdisciplinaires en emploi NSS dans le premier, ou le seul, segment.

Nous avons dans cette partie identifié un petit nombre de têtes transdisciplinaires, 123 en tout si on reprend tous les lemmes identifiés dans les différents sous-corpus, et 94 si on applique nos calculs globalement au corpus de travail. Les têtes transdisciplinaires sont très fréquentes et donc utilisées dans de nombreux titres de notre corpus de travail et, à 70 % pour les 123 têtes et à 79 % pour les 94 têtes, déjà relevées dans le lexique transdisciplinaire des écrits scientifiques de Tutin (2008). L’étude du second segment des titres bisegmentaux a mis en avant deux têtes transdisciplinaires qui le caractérisent tout particulièrement, *cas* et *exemple*. Les têtes transdisciplinaires sont caractérisées par une haute fréquence en tant que têtes et un haut degré d’abstraction. Nous conservons le nombre de 94 pour garder un point de vue global sur le corpus.

Nous avons ensuite rappelé le concept de NSS, un nom fréquent au faible contenu sémantique dont la particularité est d’être spécifié par son contexte à l’aide de plusieurs constructions spécificationnelles. Nous avons montré que le contenu spécifiant qui est relié au NSS joue une fonction d’attribut. Nous avons également montré que, si le NSS en a la capacité, on peut facilement passer de certaines CS à d’autres, que cela soit par l’ajout du pronom de reprise *ce* ou par l’ajout du verbe copule *être*. Nous avons également montré que, dans le cas d’un syntagme nominal comme contenu spécifiant, il faut toutefois que son nom noyau soit un déverbal qui dénote une action ou une activité.

Nous avons essayé de détecter les différentes occurrences de constructions spécificationnelles dans nos titres où une tête transdisciplinaire serait employée comme NSS. Nous nous sommes heurtés au problème que les définitions les plus contraignantes retournaient très peu de résultats, du fait qu’elles utilisent des verbes alors que les titres sont essentiellement averbaux. Nous pouvons résumer les emplois sous-spécifiés des têtes transdisciplinaires dans les constructions spécificationnelles dans le tableau (16).

|  |  |
| --- | --- |
| **Construction spécificationnelle** | **Nombre de têtes intégrant une construction spécificationnelle** |
| CS-I NSS + être + que | 3 |
| CS-II NSS + être + de + inf | 1 |
| C-III NSS + , + ce + être + que | 0 |
| C-IV NSS + , + ce être + de + inf | 0 |
| C-V NSS + que | 0 |
| CS-VI NSS + de + inf | estimé à 437 |
| CS-VII NSS + de + DDAA | 15 698 |
| CS-VIII NSS + être + DDAA | 1 |
| CS-IX NSS + , + ce + être + DDDAA | 1 |
| **Total** | 16 141 soit 6 % des 278 185 têtes nominales |

Tableau 16: Présence des constructions spécificationnelles classiques dans notre corpus

Nous n’avons trouvé que très peu de constructions spécificationnelles classiques dans notre corpus, nous avons décidé d’utiliser la fouille de données séquentielles pour mettre à jour des schémas d’utilisation récurrents des têtes transdisciplinaires. Nous avons trouvé deux schémas qui restreignent la CS-VII en la situant au début du titre ou après une marque de segmentation et qui et fonctionnent comme des constructions spécificationnelles.

Nous n’avons pas pu établir de corrélation entre les têtes transdisciplinaires, ou le simple fait d’être une tête de segment, et l’emploi sous-spécifié. Nous avons néanmoins identifié plusieurs emplois de têtes transdisciplinaires sous-spécifiés et que les têtes transdisciplinaires sont utilisées dans un emploi NSS à hauteur de 17 %, contre 10 % pour les autres mais sans qu’il y ait de corrélation systématique.

# Discussion sur nos résultats, limites et perspectives

Dans cette dernière partie nous revenons sur notre travail et nos résultats pour les mettre en perspective. Il s’agit de montrer leurs limites et éventuellement les perspectives d’améliorations pour nous en affranchir.

## III.1 Éléments de discussion

### Non-utilisation de la typologie de Schmid

Schmid définie une typologie des NSS (2000, p. 88) selon leurs caractéristiques sémantiques. Il définit six classes, *factual, mental, linguistic, modal, eventive*. Chaque classe est divisée en plusieurs sous-classes selon leurs utilisations, ainsi la classe *factual* se divise-t-elle en *Neutral, Causal, Evidential, Comparative, Partitive, Attitudinal* (Schmid, 2000, p. 92). Nous n’avons pas étudié la classification des têtes transdisciplinaires employées de façon sous-spécifiée.

### Non-étude de la relation entre préposition et tête transdisciplinaire

Nous avons remarqué que les têtes transdisciplinaires étaient fréquemment cooccurrentes avec les prépositions *à, de, sur*. Il aurait été envisageable d’étudier l’affinité des différentes têtes pour ces prépositions.

### Limite de l’analyse en dépendance automatique de Talismane

Si de prime abord Talismane a donné une très bonne satisfaction pour étiqueter morphosyntaxiquement les titres, il n’en est pas de même pour les relations en dépendance, notamment celles reposant sur la préposition *de* que Talismane relie souvent au mauvais recteur. Cela a peuplé de nombreux faux positifs nos requêtes au point où nous avons dû combiner la recherche via l’arbre de dépendances à la recherche positionnelle. Par exemple, l’énoncé A de B de C, se voit souvent attribué un arbre de dépendance où le second *de* a le même recteur que le premier, A, alors qu’il s’agit souvent de C. Ce cas peut-être très ambigu en français, mais empiriquement, nous avons fait un algorithme détectant B entre A et C et réattribuant le rôle de recteur à B. Des problèmes de liens de dépendances ayant une portée encore plus grande et fausse ont également été observés mais non quantifié. Cela nous laisse penser qu’on ne peut s’appuyer autant que nous le pensions initialement sur l’analyse en dépendance. L’utilisation d’un outil doit toujours être précautionneuse et détachée. Réaliser un post-traitement de correction des résultats en sortie pour comme nous l’avons fait, permet d’exploiter au mieux les puissants outils à notre disposition.

### Limitations des têtes spécifiques aux domaines

Pour la question de la variation des têtes par rapport au domaine, nous avons finalement optés pour attribuer une pondération à chaque tête. Nous sommes libres après de choisir un seuil minimum, un nombre minimum ou un pourcentage de têtes pour passer à une appréciation binaire du fait qu’il s’agit d’une tête spécifique ou non. Il manque surtout un moyen d’évaluer la pertinence des têtes.

Nous n’avons pas utilisé l’apprentissage automatique pour obtenir les têtes spécifiques aux domaines. Nous aurions pu soumettre les titres résumés à leurs têtes, une ou deux selon le nombre de segments, pour obtenir un arbre de classification supervisée. En parcourant celui-ci, nous aurions pu voir quelles têtes étaient les plus importantes pour pouvoir catégoriser dans un domaine un titre, et donc quelles têtes étaient le plus spécifique à un domaine donné.

Nous y avons vu néanmoins deux obstacles. Le premier était d’avoir seulement une ou deux têtes comme traits est très pauvre : l’apprentissage automatique se base sur la définition de traits plus pertinents, mais notre travail se concentrait uniquement sur les têtes. Le second était la difficulté de parcourir l’arbre pour avoir une liste linéaire et coefficientée des têtes spécifiques comme nous l’avons obtenue avec notre méthode.

L’utilisation de la liste des têtes spécifiques pour une autre approche de la catégorisation se heurte Le troisième obstacle était un obstacle d’utilisation de la  : la couverture des têtes spécifiques est assez faible selon le domaine considéré. L’utilisation de cette liste pour catégoriser des titres ne donnerait pas un bon résultat, mais elle peut être utilisée comme un trait dans un processus de catégorisation par apprentissage automatique.

### Têtes transdisciplinaires

La sélection des têtes transdisciplinaires sur un simple seuil de médiane, représentant le fait que la tête doit avoir dans au moins la moitié des domaines une fréquence supérieure à ce seuil est empirique. La définition d’une classe nominale par la statistique ou la structure syntaxique se prête très bien à l’automatisation. Néanmoins, il ne nous semble pas aussi simple de sélectionner automatiquement des noms sur des critères sémantiques, lorsqu’il s’agit d’aller plus loin qu’une liste.

### Listes de NSS et des lexiques scientifiques

Une grande difficulté a été de mettre la main sur des listes numériques des différentes acceptations des NSS et des lexiques scientifiques. Elles peuvent servir seulement d’indices, mais précieux, car les NSS sont un emploi et non une classe lexicale a priori, bien qu’il existe des propriétés lexicales a priori de capacité à pouvoir être employé comme NSS. Certains articles pointaient sur un site web n’était plus en ligne, d’autres ne prenaient même pas cette peine, ou d’autres proposaient seulement un format PDF.

Pour la linguistique de corpus, la mise à disposition pérenne des listes produites par la computation est parfois aussi importante que l’article. La capacité de stocker un article avec des pièces-jointes, parfois volumineuses, nous semble importante, notamment pour les archives ouvertes. L’ensemble de nos données et de notre code est de notre côté disponible à l’URL : <https://github.com/Xitog/tal/tree/master/master2>

### Opérationnalisation des NSS

L’opérationnalisation des NSS est ardue, surtout dans une perspective de traitement automatique des langues. L’idée de Huyghe (2018) de se retreindre au concept de nom porteur, noms capables de porter un contenu prépositionnel qui correspond aux constructions CS-I et CS-II, présente l’avantage de réduire considérablement le périmètre d’investigation pour pouvoir l’analyser plus profondément, comme il le fait pour *fait* dans son article.

Avec les constructions les moins contraintes, le bruit augmente considérablement. L’obligation d’un nom déverbal dénotant une action ou une activité permet de les restreindre. Muni d’une liste adéquate, mais il s’agit là-aussi d’une classe ouverte, ou, mieux, d’une règle de formation automatique des déverbaux, on pourrait considérablement augmenter la précision de notre recherche. Néanmoins, reste la question d’un nom non déverbal mais qui induit une action implicite.

Nous avons laissé de côté encore d’autres constructions spécificationnelles, faute de temps. Notamment Nakamura (2017) a commencé à développer des constructions attributives avec le verbe avoir : « Il a pour **objectif** de rédiger une loi » / « Il a pour **objectif** la rédaction d’une loi » / « Il a pour objectif qu’une loi soit rédigée ». Roze et al. (2016) ont également mis à jour de nouvelles CS dont une est celle proposée par Nakamura avec *pour*.

III.2 Perspectives ? XXX

# Conclusion

Nous sommes repartis du travail effectué pour notre mémoire de Master 1. L’identification de schémas récurrents après le double point dans les titres de publications scientifiques avait permis de mettre en avant une classe de noms communs abstraits, très fréquents et pluridisciplinaires. Ces caractéristiques se rapprochent d’un type d’emploi des noms, les noms sous-spécifiés. Nous avons voulu savoir si des noms étaient employés de cette façon dans les titres. Nous avons pris comme hypothèse que les têtes des segments des titres étaient les noms les plus propices à être employés de façon sous-spécifiée.

Nous avons, dans un premier temps, constitué un corpus de travail au sein du matériau initial qui compte près de 340 000 titres tirés de HAL, fournis par Tanguy et Rebeyrolle (à paraître). Puis, nous avons utilisé la lemmatisation, la catégorisation morphosyntaxique et l’analyse en dépendances syntaxiques fournies par l’outil Talismane (Urieli, 2013). Enfin, nous avons sélectionné les titres monosegmentaux ou bisegmentaux avec pour chacun une tête par segment, dont on au moins une de nature nominale. Lorsque Talismane trouvait un segment à deux têtes, nous avons écarté le titre. Lorsque Talismane trouvait un segment sans tête dans un titre à deux segments, nous avons essayé d’en trouver une en promouvant un mot qui serait régi uniquement par un mot de l’autre segment disposant déjà d’une tête. Nous avons pu conformer à notre règle « un segment une tête » près de 98 % des 56 851 titres auxquels il manquait une tête. Pour finir, nous avons constitué un corpus de travail de 250 998 titres, gardant près de 74 % du matériau initial.

Après avoir constitué notre corpus de travail et identifié toutes les têtes, nous nous sommes d’abord interrogés sur le nombre de segments des titres en fonction du domaine. Il apparaît que les sciences humaines utilisent dans les mêmes proportions titres monosegmentaux et titres bisegmentaux, tandis que les sciences exactes privilégient les titres monosegmentaux. Nous nous sommes interrogés sur la nature grammaticale des titres. Il s’est avéré que l’extrême majorité des têtes étaient des noms conférant ainsi une nature nominale aux titres : 86 % dans le cas des titres monosegmentaux. Dans le cas des titres bisegmentaux, cette majorité est très claire si on ne regarde que le premier segment, 84 %. Si on ne considère comme nominal que les titres bisegmentaux dont les deux têtes sont des noms, cette proportion tombe à 68 %. Nous pouvons donc conclure que les titres sont essentiellement des syntagmes nominaux.

Partant de cette constatation, nous avons voulu savoir s’il existait des têtes nominales spécifiques à certains domaines. Utilisant la valeur de TF\*IDF, en considérant les domaines comme un document unique et leurs titres comme autant de phrases de ce document, nous avons pondéré chaque tête par un indice de spécificité. Les têtes sélectionnées sont des noms pleins, qui révèlent les objets d’étude des différents domaines.

Nous avons également recherché les têtes transdisciplinaires, fréquentes dans tous les domaines. Nous avons identifié 94 têtes transdisciplinaires dans le corpus général, 74 sont présentes dans le lexique transdisciplinaire des écrits scientifiques (Tutin, 2008), soit 79 %, et 82 sont présentes dans le lexique scientifique transdisciplinaire (Hatier, 2016), soit 87 %.

Nous avons ensuite essayé de rapprocher les têtes transdisciplinaires, dont la fréquence et la transdisciplinarité impliquent un faible contenu sémantique, des noms sous-spécifiés qui se caractérisent par une très grande fréquence et un faible contenu sémantique également. Après avoir défini notre perception des noms sous-spécifiés, nous avons vu que leur définition opératoire est structurelle : les noms sous-spécifiés s’insèrent dans des constructions spécificationnelles dont la fonction est de lier le nom général sous-spécifié à un contenu présent dans son contexte et qui va le « remplir ».

XXX

Nous nous sommes heurté d’un côté à l’absence dans notre corpus de constructions spécificationnelles facilement identifiables, estimées à moins de 500 occurrences, et de l’autre à une structure non assez sélective malgré la mise en évidence de la nécessité que le contenu spécifiant soit lié à une action ou une activité, soit par le truchement d’un verbe conjugué s’il s’agit d’une proposition subordonnée conjonctive, soit par le truchement d’un verbe à l’infinitif s’il s’agit d’une proposition infinitive, soit, s’il s’agit d’un syntagme nominal pouvant être inclus dans un syntagme prépositionnelle, que le noyau nominal soit un déverbal dénotant une action ou une activité.

Faute de construction spécificationnelle classique, nous avons donc étudié les schémas récurrents dans lesquels s’insèrent nos têtes transdisciplinaires. Nous avons pu établir que ceux-ci sont très ramassés et averbaux ce qui est en accord avec les spécificités des titres. Les deux schémas détectés se rapprochent d’une construction spécificationnelle, la CS-VII, qui est la moins contraignante des CS étudiées. Ils ajoutent comme contrainte que le schéma doit se trouver en début de titre ou de segment. Nous avons ensuite voulu étudier si, outre la correspondance syntaxique entre les deux schémas récurrents détectés et la CS-VII, il y avait effectivement un emploi sous-spécifié des têtes transdisciplinaires.

Nous avons rencontré des problèmes pour estimer l’utilisation en emploi sous-spécifié. La détermination de l’emploi repose moins, dans la configuration de la CS-VII, sur des critères syntaxiques que des critères lexicaux et sémantiques. Pour le lexical, à défaut de liste établie des DDAA, de plus car c’est une classe ouverte, la tâche est impossible, et pour la sémantique, nous n’avons pu faire autrement qu’en faisant appel au jugement de l’intuition ce qui est insuffisant dans une perspective d’automatisation. Néanmoins, nous avons pu déterminer un trait saillant morphologique de certains DDAA : des terminaisons en -tion, -sion, -age, -ment. Nous fondant dessus, nous avons établi un estimateur qui nous a donné le nombre d’utilisations en emploi sous-spécifié des têtes transdisciplinaires.

Malheureusement, nous n’avons pu montrer une corrélation entre les têtes transdisciplinaires et les emplois sous-spécifiés, ni même entre le fait pour un lemme d’être une tête et les emplois sous-spécifiés. Néanmoins, les têtes transdisciplinaires sont légèrement plus utilisées en emploi NSS que les têtes n’étant pas transdisciplinaires, 17 % contre 10 %. Nous pensons cependant que l’amélioration de notre estimateur, la prise en compte de CS spécifique, l’obtention d’une méthode pour vérifier si un nom est un DDAA à suffixe zéro, toutes ces actions permettraient d’augmenter le rappel d’emploi NSS et ainsi mieux mesurer la corrélation entre têtes, têtes transdisciplinaires et emploi NSS.

# Bibliographie

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

Aktas, R. N. et Cortes, V. (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes, 7(1)*, 3-14.

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

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

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

Baethge, C. (2008). Publish together or perish: the increasing number of authors per article in academic journals is the consequence of a changing scientific culture. *Deutsches Arzteblatt international, 105(20)*, 380-383.

Benítez-Castro, M. Á. (2014). *Formal, syntactic, semantic and textual features of English shell nouns*. Thèse de doctorat, Universidad de Granada.

Biber, D., Conrad, S. et Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge : Cambridge University Press.

Biber, D. et Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9(1)*, 2-20.

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

Cori, M. et David, S. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages, 171*, 111-129.

Delhay, C. (2014). Pour un «complément-attribut». *Repères. Recherches en didactique du français langue maternelle, (49)*, 57-76.

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

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

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

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

Flowerdew, J. (2006). Use of signalling nouns in a learner corpus. *International Journal of Corpus Linguistics, 11(3)*, 345-362.

Flowerdew, J. & Forest, R. W. (2015). *Signalling nouns in English*. Cambridge University Press.

Francis, G. (1986). *Anaphoric nouns*. English Language Research, Department of English, University of Birmingham.

Francis, G. (1994). Labelling discourse: an aspect of nominal-group lexical cohesion. In Coulthard, M. ed, (1994), *Advances in written text analysis*, London: Routledge, 83-101.

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

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

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

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

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

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

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

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

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

Ho-Dac, L.-M., Jacques, M.-P. & Rebeyrolle, J. (2004). Sur la fonction discursive des titres. Dans S. Porhiel et D. Klingler (éds). *L’unité texte*, Pleyben, Perspectives, 125-152.

Hunston, S. & Francis, G. (1999). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins (Studies in Corpus Linguistics 4).

Huot, H. (1981). *Constructions infinitives du français: le subordonnant de* (Vol. 12). Genève : Librairie Droz.

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

Ivanic, R. (1991). Nouns in search of a context: A study of nouns with both open- and closed-system characteristics. *International Review of Applied Linguistics in Language Teaching, 2*, 93-114.

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

Jalilifar, A., Hayati, A., & Mayahi, N. (2010). An exploration of generic tendencies in Applied Linguistics titles. *Journal of Faculty of Letters and Humanities, 5(16)*, 35-57.

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

Kalmbach, J.-P. (2019). *La grammaire du français langue étrangère pour étudiants finnophones*. Repéré à <http://research.jyu.fi/grfle/675.html>

Kolhatkar, V., & Hirst, G. (2014). Resolving shell nouns. Dans *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 499-510.

Kutch, T. D. C. (1978). Relation of title length to numbers of authors in journal articles. *Journal of the American Society of Information Science, 19(4)*, 200-202.

Larivière, V., Gingras, Y., Sugimoto, C. R. and Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology, 66(7)*, 1323-1332.

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

Legallois, D. (2006). Quand le texte signale sa structure : la fonction textuelle des noms sous-spécifiés. *Corela HS-5* : <http://corela.edel.univ-poitiers.fr/index.php?id=1465>

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

Legallois, D., & Gréa, P. (2006). L’objectif de cet article est de… Construction spécificationnelle et grammaire phraséologique. *Cahiers de praxématique, (46)*, 161-186.

Lyons, J. (1977). *Semantics (Vol. 2)*. Cambridge: Cambridge university press.

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

Merrill, E. et Knipps, A. (2014). What's in a Title?. *The Journal of Wildlife Management, 78(5)*, 761-762.

Mounin, G. (dir.) (2004). *Dictionnaire de la linguistique*. Paris : PUF (Quadrige).

Mousavi, A. et Moini, M. R. (2014). A corpus study of shell nouns in published research articles of education. *Procedia-Social and Behavioral Sciences, 98*, 1282-1289.

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

Nakamura, T. (2017). Extensions transitives de constructions spécificationnelles. *Langue française, 2017 (2)*, 69-84.

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

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

Quiniou, S., Cellier, P., Charnois, T. et Legallois, D. (2012). Fouille de données pour la stylistique: cas des motifs séquentiels émergents. Dans *Journées Internationales d’Analyse Statistique des Données Textuelles (JADT’12), Liège,* 821-833.

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

Riegel, M. (2006). Grammaire des constructions attributives : avec ou sans copule. Dans *Construction, acquisition et communication : Études linguistiques de discours contemporains,* Engwall, G. (éd.). Stockholm : Université de Stockholm(Acta Universitatis Stockholmiensis Romanica Stockholmiensia 23).

Roze, C., Charnois, T., Legallois, D., Ferrari, S. et Salles, M. (2014). Identification des noms sous-spécifiés, signaux de l’organisation discursive. Dans *Proceedings of TALN 2014, 1,* 377-388.

Sagi, I., & Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science, 34(5)*, 680-687.

Salager-Meyer, F. & Alcaraz Ariza, M. Á. (2013). Titles are" serious stuff": a historical study of academic titles. *Jahr, 4(7), 257-271.*

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin: Mouton de Gruyter (Topics in English Linguistics 34).

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

Schwischay, B. (2001). Notes d’exposés sur deux modèles de description syntaxique [Document PDF]. Repéré à <http://www.home.uni-osnabrueck.de/bschwisc/archives/deuxmodeles.pdf>

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

Soler, V. (2011). Comparative and contrastive observations on scientific titles written in English and Spanish. *English for Specific Purposes, 30(2)*, 124-137.

Subotic, S. & Mukherjee, B. (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles*. Journal of Information Science, 40(1)*, 115-124.

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

Tadros, A. (1994). Predictive categories in expository text. In Coulthard, M. ed, (1994), *Advances in written text analysis*, London: Routledge, 83-96.

Tanguy, L., Rebeyrolle, J. (à paraître). Les titres des publications scientifiques en français : fouille de texte pour le réperage de schémas lexico-syntaxiques.

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

Tutin, A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue Française de Linguistique Appliquée, 12(2)*, 5-14.

Tutin, A. (2008). Sémantique lexicale et corpus : l’étude du lexique transdisciplinaire des écrits scientifiques. *Lublin studies in modern languages and litterature, 32,* 242-260.

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.

Urieli, A. et Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. *Actes de TALN*, Sables D'Olonne.

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

Williams, G. (2005*)*. *La linguistique de corpus*. Rennes : Presses universitaires de Rennes.

Winter, E. O. (1977). A clause-relational approach to English texts: a study of some predictive lexical items in written discourse. *Instructional science*, *6*(1), 1-92.

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

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

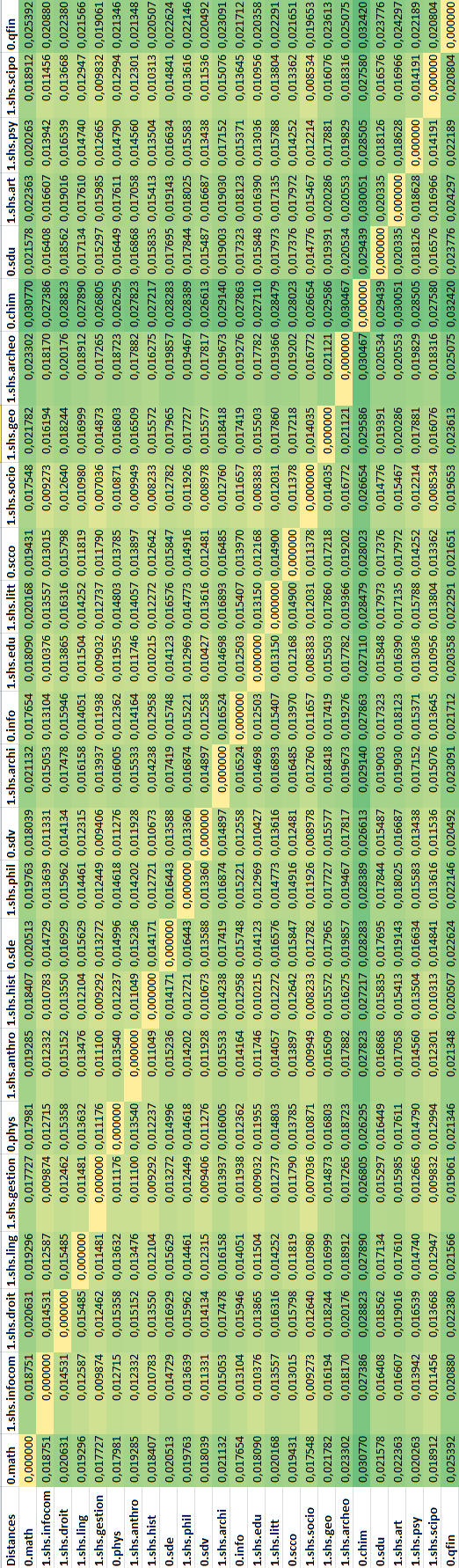
Yakhontova, T. (2002). Titles of conference presentation abstracts: A cross-cultural perspective. Dans E. Ventola, C. Shalom et S. Thompson (éds.), *The language of conferencing*, 277-300. Francfort-sur-le-Main : Peter Lang.

Yitzhaki, M. (1994). Relation of title length of journal articles to number of authors. *Scientometrics, 30(1)*, 321-332.

Yitzhaki, M. (2002). Relation of the title length of a journal article to the length of the article. *Scientometrics, 54(3)*, 435-447.

Annexes

# A1. Distance des domaines de par leurs têtes spécifiques



# A2. Combinaisons des têtes de titres bisegmentaux

Ce tableau liste toutes les combinaisons possibles des têtes de titres bisegmentaux en termes de catégories morphosyntaxiques et en agrégeant les catégories nominales, propositionnelles et verbales en une seule à chaque fois, respectivement NOUN, PREP et VERB. La requête pour obtenir ce tableau est donné en première ligne, ainsi que le nombre de titres sur lesquels la requête a été lancée : ici l’ensemble des titres bisegmentaux de notre corpus de travail.

\*\*\* ['roots.0.pos:agg', 'roots.1.pos:agg'] \*\*\* (110785)

-------------------------------------------------------

01. NOUN-NOUN 75592 68.2331 % 68.23 %

02. NOUN-PREP 8996 8.1202 % 76.35 %

03. VERB-NOUN 8506 7.6779 % 84.03 %

04. NOUN-VERB 5426 4.8978 % 88.93 %

05. PREP-NOUN 4650 4.1973 % 93.13 %

06. NOUN-CC 1209 1.0913 % 94.22 %

07. VERB-PREP 1015 0.9162 % 95.13 %

08. NOUN-ADJ 1000 0.9026 % 96.04 %

09. VERB-VERB 661 0.5967 % 96.63 %

10. ADJ-NOUN 537 0.4847 % 97.12 %

11. NOUN-PONCT 395 0.3565 % 97.47 %

12. NOUN-CS 368 0.3322 % 97.81 %

13. PREP-VERB 333 0.3006 % 98.11 %

14. PREP-PREP 310 0.2798 % 98.39 %

15. CS-NOUN 270 0.2437 % 98.63 %

16. VERB-PONCT 145 0.1309 % 98.76 %

17. VERB-CC 140 0.1264 % 98.89 %

18. NOUN-PRO 133 0.1201 % 99.01 %

19. NOUN-ADV 105 0.0948 % 99.10 %

20. VERB-ADJ 91 0.0821 % 99.18 %

21. ADJ-PREP 82 0.0740 % 99.26 %

22. PREP-CC 64 0.0578 % 99.32 %

23. PRO-NOUN 63 0.0569 % 99.37 %

24. ADJ-VERB 59 0.0533 % 99.43 %

25. VERB-CS 50 0.0451 % 99.47 %

26. NOUN-DET 42 0.0379 % 99.51 %

27. PREP-ADJ 38 0.0343 % 99.54 %

28. CS-PREP 37 0.0334 % 99.58 %

29. PREP-CS 30 0.0271 % 99.60 %

30. CS-VERB 27 0.0244 % 99.63 %

31. CC-NOUN 27 0.0244 % 99.65 %

32. NOUN-I 27 0.0244 % 99.68 %

33. VERB-ADV 24 0.0217 % 99.70 %

34. PREP-PONCT 23 0.0208 % 99.72 %

35. ADV-VERB 20 0.0181 % 99.74 %

36. NOUN-ADVWH 18 0.0162 % 99.75 %

37. PONCT-NOUN 15 0.0135 % 99.77 %

38. NOUN-ET 14 0.0126 % 99.78 %

39. ADJ-CC 14 0.0126 % 99.79 %

40. ADV-NOUN 14 0.0126 % 99.81 %

41. NOUN-CLO 14 0.0126 % 99.82 %

42. DET-NOUN 13 0.0117 % 99.83 %

43. ET-NOUN 13 0.0117 % 99.84 %

44. PRO-VERB 13 0.0117 % 99.85 %

45. VERB-ADVWH 13 0.0117 % 99.87 %

46. VERB-PRO 12 0.0108 % 99.88 %

47. CLR-NOUN 11 0.0099 % 99.89 %

48. NOUN-DETWH 11 0.0099 % 99.90 %

49. ADJ-CS 9 0.0081 % 99.90 %

50. NOUN-PROREL 8 0.0072 % 99.91 %

51. CS-PONCT 6 0.0054 % 99.92 %

52. CS-CC 6 0.0054 % 99.92 %

53. ADJ-ADJ 6 0.0054 % 99.93 %

54. PRO-PREP 5 0.0045 % 99.93 %

55. CC-PREP 5 0.0045 % 99.94 %

56. NOUN-CLS 5 0.0045 % 99.94 %

57. VERB-DET 5 0.0045 % 99.95 %

58. CS-ADJ 4 0.0036 % 99.95 %

59. PONCT-VERB 4 0.0036 % 99.95 %

60. ADJ-PONCT 3 0.0027 % 99.96 %

61. PONCT-PREP 2 0.0018 % 99.96 %

62. VERB-ET 2 0.0018 % 99.96 %

63. NOUN-CLR 2 0.0018 % 99.96 %

64. VERB-DETWH 2 0.0018 % 99.96 %

65. PREP-I 2 0.0018 % 99.96 %

66. ADJ-PRO 2 0.0018 % 99.97 %

67. VERB-I 2 0.0018 % 99.97 %

68. VERB-PROREL 2 0.0018 % 99.97 %

69. I-NOUN 2 0.0018 % 99.97 %

70. CS-CS 2 0.0018 % 99.97 %

71. I-VERB 2 0.0018 % 99.98 %

72. CLO-NOUN 2 0.0018 % 99.98 %

73. ADVWH-NOUN 2 0.0018 % 99.98 %

74. CLS-VERB 1 0.0009 % 99.98 %

75. PONCT-CC 1 0.0009 % 99.98 %

76. CC-VERB 1 0.0009 % 99.98 %

77. PRO-PONCT 1 0.0009 % 99.98 %

78. PRO-CC 1 0.0009 % 99.98 %

79. PRO-ADJ 1 0.0009 % 99.98 %

80. PREP-ADV 1 0.0009 % 99.99 %

81. PREP-CLR 1 0.0009 % 99.99 %

82. CLR-VERB 1 0.0009 % 99.99 %

83. ET-PREP 1 0.0009 % 99.99 %

84. I-PREP 1 0.0009 % 99.99 %

85. PREP-ET 1 0.0009 % 99.99 %

86. CC-CLS 1 0.0009 % 99.99 %

87. DET-PREP 1 0.0009 % 99.99 %

88. ET-VERB 1 0.0009 % 99.99 %

89. PREP-DETWH 1 0.0009 % 99.99 %

90. CLS-NOUN 1 0.0009 % 99.99 %

91. PREP-DET 1 0.0009 % 100.00 %

92. VERB-PROWH 1 0.0009 % 100.00 %

93. CC-ADJ 1 0.0009 % 100.00 %

94. ADVWH-PREP 1 0.0009 % 100.00 %

95. VERB-CLO 1 0.0009 % 100.00 %

96. PREP-PRO 1 0.0009 % 100.00 %

# A3. Liste des têtes transdisciplinaires

Le tableau suivant présente nos 123 têtes transdisciplinaires si on rassemble toutes les têtes qui correspondent à notre définition de la transdisciplinarité dans les trois sous-corpus et le corpus général de travail. Est indiqué le lemme, la catégorie du discours, si le lemme appartient aux formes du lexique transdisciplinaire des écrits scientifiques (LTES) (Tutin, 2008), au lexique scientifique transdisciplinaire (LST) (Hatier, 2016) et si le lemme appartient à la liste des signalling nouns (Flowerdew et Forest, 2015) avec la fréquence normalisée dans leur corpus. Nous notons que :

* Sur les 94 têtes transdisciplinaires relevées dans le corpus général, 74 sont présentes dans le LTES, soit 79 %, et 82 sont présentes dans le LST, soit 87 %.
* Sur les 123 têtes transdisciplinaires relevées, 86 sont présentes dans le LTES, soit 70 %, et 101 sont présentes dans le LST, soit 82 %.
* Sur les 123 têtes transdisciplinaires relevées, 110 sont également relevées par Flowerdew et Forest comme étant utilisées comme signalling nouns, soit 89 %.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **N°** | **Lemme** | **Tout le corpus** | **Titres monosegmentaux** | **1er segment des titres bisegmentaux** | **2e segment des titres bisegmentaux** | **Présence dans le LTES / LST** | **Présence dans signalling nouns** |
| 1 | activité | 1 |  | 1 |  | LTES / LST | 59 |
| 2 | an | 1 |  |  | 1 | LTES |  |
| 3 | analyse | 1 | 1 | 1 | 1 | LTES / LST | 178 |
| 4 | application | 1 | 1 | 1 | 1 | LTES / LST | 44 |
| 5 | apport | 1 | 1 | 1 | 1 | LTES / LST | 16 |
| 6 | approche | 1 | 1 | 1 | 1 | LTES / LST | 246 |
| 7 | aspect | 1 | 1 |  | 1 | LTES / LST | 78 |
| 8 | bilan | 1 |  |  | 1 | LTES / LST | 83 |
| 9 | cadre | 1 | 1 |  | 1 | LTES / LST | 31 |
| 10 | cas | 1 |  |  | 1 | LTES / LST | 890 |
| 11 | changement | 1 |  |  | 1 | LTES / LST | 209 |
| 12 | comparaison | 1 | 1 |  | 1 | LTES / LST | 44 |
| 13 | compte |  |  | 1 |  |  | 18 |
| 14 | concept | 1 |  |  | 1 | LTES / LST | 143 |
| 15 | condition |  |  |  | 1 | LTES / LST | 248 |
| 16 | conséquence | 1 | 1 |  | 1 | LTES / LST | 132 |
| 17 | construction | 1 | 1 | 1 | 1 | LTES / LST | 2 |
| 18 | contexte |  |  | 1 | 1 | LTES / LST | 73 |
| 19 | contribution | 1 | 1 |  | 1 | LTES / LST | 16 |
| 20 | contrôle |  | 1 |  |  | LTES | 3 |
| 21 | culture |  |  | 1 |  |  |  |
| 22 | défi | 1 |  |  | 1 |  | 26 |
| 23 | définition |  |  |  | 1 | LTES / LST | 68 |
| 24 | démarche |  |  |  | 1 | LTES / LST | 112 |
| 25 | développement | 1 | 1 | 1 | 1 | LTES / LST | 39 |
| 26 | dimension | 1 |  |  |  | LTES / LST | 8 |
| 27 | discours | 1 |  |  | 1 | / LST | 51 |
| 28 | dispositif | 1 |  | 1 | 1 | LTES / LST | 7 |
| 29 | donnée |  |  |  | 1 | LTES / LST | 44 |
| 30 | dynamique | 1 | 1 | 1 | 1 | / LST |  |
| 31 | économie |  |  | 1 |  |  |  |
| 32 | effet | 1 | 1 | 1 | 1 | LTES / LST | 393 |
| 33 | élément | 1 | 1 |  | 1 | LTES / LST | 33 |
| 34 | émergence | 1 | 1 |  | 1 | / LST |  |
| 35 | enjeu | 1 | 1 | 1 | 1 | / LST |  |
| 36 | enquête | 1 |  |  | 1 | / LST | 16 |
| 37 | enseignement | 1 |  |  | 1 |  |  |
| 38 | espace | 1 | 1 | 1 | 1 | / LST |  |
| 39 | essai | 1 | 1 |  | 1 | / LST | 41 |
| 40 | état | 1 | 1 | 1 | 1 | LTES / LST | 10 |
| 41 | étude | 1 | 1 | 1 | 1 | LTES / LST | 18 |
| 42 | évaluation | 1 | 1 | 1 | 1 | LTES / LST | 10 |
| 43 | évolution | 1 | 1 | 1 | 1 | LTES / LST | 39 |
| 44 | exemple | 1 |  | 1 | 1 | LTES / LST | 421 |
| 45 | expérience | 1 | 1 | 1 | 1 | LTES / LST | 3 |
| 46 | figure | 1 | 1 | 1 | 1 | / LST | 88 |
| 47 | fonction |  | 1 |  |  | LTES / LST | 150 |
| 48 | formation | 1 | 1 | 1 | 1 | / LST |  |
| 49 | forme | 1 | 1 | 1 | 1 | LTES / LST | 88 |
| 50 | gestion | 1 | 1 | 1 |  | LTES |  |
| 51 | histoire | 1 | 1 | 1 | 1 |  | 20 |
| 52 | identité |  |  | 1 |  | / LST | 3 |
| 53 | illustration |  |  |  | 1 | / LST | 33 |
| 54 | image | 1 | 1 | 1 | 1 | LTES / LST | 5 |
| 55 | impact | 1 | 1 | 1 | 1 | / LST | 96 |
| 56 | influence | 1 | 1 | 1 | 1 | LTES / LST | 44 |
| 57 | intégration | 1 | 1 | 1 |  | LTES / LST |  |
| 58 | interaction | 1 | 1 | 1 | 1 | LTES / LST | 15 |
| 59 | intérêt | 1 | 1 |  | 1 | LTES / LST | 29 |
| 60 | introduction | 1 | 1 | 1 | 1 | LTES / LST | 70 |
| 61 | jeu | 1 | 1 | 1 | 1 |  |  |
| 62 | leçon |  |  |  | 1 |  | 51 |
| 63 | lecture | 1 | 1 |  | 1 | LTES / LST | 33 |
| 64 | limite |  |  |  | 1 | LTES / LST | 10 |
| 65 | mesure | 1 | 1 | 1 |  | LTES / LST | 46 |
| 66 | méthode | 1 | 1 | 1 | 1 | LTES / LST | 280 |
| 67 | méthodologie | 1 | 1 |  | 1 | / LST | 13 |
| 68 | mode |  |  |  | 1 | LTES / LST | 11 |
| 69 | modèle | 1 | 1 | 1 | 1 | LTES / LST | 474 |
| 70 | modélisation | 1 | 1 | 1 | 1 | LTES | 3 |
| 71 | mythe |  |  |  | 1 |  | 2 |
| 72 | note | 1 | 1 | 1 | 1 | / LST | 13 |
| 73 | notion |  | 1 |  |  | LTES / LST | 73 |
| 74 | objet | 1 |  |  | 1 | LTES / LST | 3 |
| 75 | organisation | 1 | 1 | 1 |  | LTES / LST | 13 |
| 76 | outil | 1 | 1 | 1 | 1 | LTES / LST | 7 |
| 77 | perception | 1 |  |  |  | LTES / LST | 85 |
| 78 | paradoxe |  |  |  | 1 | / LST | 11 |
| 79 | parcours |  |  |  | 1 |  | 36 |
| 80 | perspective | 1 | 1 |  | 1 | LTES / LST | 36 |
| 81 | piste |  |  |  | 1 |  | 2 |
| 82 | place | 1 | 1 | 1 | 1 | LTES | 21 |
| 83 | point | 1 | 1 |  | 1 | / LST | 393 |
| 84 | politique | 1 | 1 | 1 | 1 |  | 2 |
| 85 | pratique | 1 | 1 | 1 | 1 | / LST | 73 |
| 86 | présentation | 1 | 1 | 1 | 1 | LTES / LST | 11 |
| 87 | principe | 1 |  |  | 1 | LTES / LST | 251 |
| 88 | problématique |  |  |  | 1 | / LST | 287 |
| 89 | problème | 1 | 1 |  | 1 | LTES / LST | 619 |
| 90 | processus | 1 | 1 | 1 |  | LTES / LST | 230 |
| 91 | production | 1 | 1 | 1 |  | LTES / LST | 2 |
| 92 | projet | 1 | 1 | 1 | 1 | LTES / LST | 37 |
| 93 | proposition | 1 | 1 |  | 1 | LTES / LST | 46 |
| 94 | question | 1 | 1 | 1 | 1 | LTES / LST | 313 |
| 95 | rapport | 1 |  |  | 1 | LTES / LST | 10 |
| 96 | réalité |  |  |  | 1 | LTES | 23 |
| 97 | recherche | 1 | 1 | 1 | 1 | LTES / LST | 2 |
| 98 | réflexion | 1 | 1 | 1 | 1 | LTES / LST | 16 |
| 99 | regard | 1 | 1 |  | 1 |  | 5 |
| 100 | relation | 1 | 1 | 1 | 1 | LTES / LST | 93 |
| 101 | remarque | 1 | 1 |  | 1 | / LST | 21 |
| 102 | représentation | 1 | 1 | 1 | 1 | LTES / LST | 11 |
| 103 | réseau | 1 | 1 | 1 |  | LTES / LST | 7 |
| 104 | résultat | 1 |  |  | 1 | LTES / LST | 572 |
| 105 | retour | 1 | 1 | 1 | 1 |  | 29 |
| 106 | revue |  |  |  | 1 |  | 8 |
| 107 | rôle | 1 | 1 | 1 | 1 | LTES / LST | 153 |
| 108 | science | 1 |  |  |  | / LST |  |
| 109 | source |  |  |  | 1 | LTES / LST | 10 |
| 110 | stratégie | 1 | 1 | 1 | 1 | LTES / LST | 205 |
| 111 | structure | 1 | 1 | 1 | 1 | LTES / LST | 13 |
| 112 | synthèse |  |  |  | 1 | LTES / LST | 2 |
| 113 | système | 1 | 1 | 1 | 1 | LTES / LST | 109 |
| 114 | temps |  | 1 |  |  | LTES | 184 |
| 115 | théorie | 1 | 1 |  | 1 | / LST | 494 |
| 116 | traitement | 1 | 1 |  |  | LTES / LST | 300 |
| 117 | transformation |  | 1 |  |  | LTES / LST | 2 |
| 118 | travail | 1 | 1 | 1 |  | LTES / LST | 24 |
| 119 | usage | 1 | 1 | 1 | 1 | LTES / LST | 73 |
| 120 | utilisation | 1 | 1 | 1 | 1 | / LST | 5 |
| 121 | valeur |  | 1 |  |  | LTES / LST | 13 |
| 122 | variation | 1 | 1 |  |  | LTES / LST | 15 |
| 123 | voie |  |  |  | 1 | LTES / LST | 668 |
|  | **123** | **94** | **81** | **63** | **99** | **86 / 101** | **110 / 123**  **83 / 94** |

# A4. Étiquettes utilisées par Talismane et HAL

## A4.1 Catégories morphosyntaxiques de Talismane

Ces informations sont tirées de <http://joliciel-informatique.github.io/talismane/#tagset>.

|  |  |
| --- | --- |
| **Code** | **Catégorie morphosyntaxique** |
| ADJ | Adjectif |
| ADV | Adverbe |
| ADVWH | Adverbe interrogatif |
| CC | Conjonction de coordination |
| CLO | Clitique objet |
| CLR | Clitique réflexif |
| CLS | Clitique sujet |
| CS | Conjonction de subordination |
| DET | Déterminant |
| DETWH | Déterminant interrogatif |
| ET | Mot étranger |
| I | Interjection |
| NC (que nous rassemblons dans NOUN) | Nom commun |
| NPP (que nous rassemblons dans NOUN) | Nom propre |
| P (que nous rassemblons dans PREP) | Préposition |
| P+D (que nous rassemblons dans PREP) | Préposition et déterminant combinés (“du”) |
| P+PRO (que nous rassemblons dans PREP) | Préposition et pronom combiné (“duquel”) |
| PONCT | Ponctuation |
| PRO | Pronom |
| PROREL | Pronom relatif |
| PROWH | Pronom interrogatif |
| V (que nous rassemblons dans VERB) | Verbe à l’indicatif |
| VIMP (que nous rassemblons dans VERB) | Verbe à l’impératif |
| VINF (que nous rassemblons dans VERB) | Verbe à l’infinitif |
| VPP (que nous rassemblons dans VERB) | Verbe au participe passé |
| VPR (que nous rassemblons dans VERB) | Verbe au participe présent |
| VS (que nous rassemblons dans VERB) | Verbe au subjonctif |

## A4.2 Code des 27 domaines de HAL retenus

Ces informations sont tirées de HAL : <https://hal.archives-ouvertes.fr>

|  |  |  |
| --- | --- | --- |
| 01 | 0.chim | Chimie |
| 02 | 0.info | Informatique |
| 03 | 0.math | Mathématiques |
| 04 | 0.phys | Physique |
| 05 | 0.qfin | Économie et finance quantitative |
| 06 | 0.scco | Sciences cognitives |
| 07 | 0.sde | Sciences de l'environnement |
| 08 | 0.sdu | Planète et Univers |
| 09 | 0.sdv | Sciences du Vivant |
| 10 | 1.shs.anthro | Anthropologie |
| 11 | 1.shs.archeo | Archéologie et Préhistoire |
| 12 | 1.shs.archi | Architecture |
| 13 | 1.shs.art | Art et histoire de l'art |
| 14 | 1.shs.autre | Autres |
| 15 | 1.shs.droit | Droit |
| 16 | 1.shs.edu | Éducation |
| 17 | 1.shs.geo | Géographie |
| 18 | 1.shs.gestion | Gestion et management |
| 19 | 1.shs.hist | Histoire |
| 20 | 1.shs.infocom | Sciences de l'information et de la communication |
| 21 | 1.shs.ling | Linguistique |
| 22 | 1.shs.litt | Littératures |
| 23 | 1.shs.phil | Philosophie |
| 24 | 1.shs.psy | Psychologie |
| 25 | 1.shs.scipo | Science politique |
| 26 | 1.shs.socio | Sociologie |
| 27 | NONE | Pas de domaine associé |

# A5. Éléments techniques

## A5.A Présentation de l’API de requêtage de notre corpus

Nous présentons dans cette partie notre interface de programmation de l’application (API) que nous avons développée afin d’interroger notre corpus.

Requêtes sur notre corpus pour filtrer le corpus, trouver des titres et faire des statistiques.

stat('domain')

Produit un comptage des titres selon le domaine des titres. Le résultat est un dictionnaire où la clé est le domaine et la valeur le nombre de titre dans ce domaine.

stat(('nb\_parts', 'nb\_segments'))

Produit un comptage des titres selon les combinaisons des valeurs possibles pour le nombre de parties et le nombre de segments. Le résultat est un dictionnaire où la clé est un tuple constitué d’une combinaison existante de valeurs des deux dimensions, par exemple 1 partie, 2 segments, et la valeur le nombre de titre correspondant à cette combinaison, le nombre de titres ayant 1 partie et 2 segments.

count({'nb\_parts' : 1, 'nb\_segments' : 2})

Compte le nombre de titre ayant une partie et deux segments.

t12 = select({'nb\_parts' : 1, 'nb\_segments' : 2})

Création d’un sous-corpus composé des titres ayant une partie et deux segments. On peut ensuite utiliser les requêtes stat et count sur celui-ci via une variable globale qui contient le corpus courant.

find({'nb\_roots' : 2}, nb=20)

Cherche et affiche 20 titres ayant 2 têtes.

find({'roots.0.lemma' : 'rôle', 'roots.1.lemma' : 'cas',   
 'segments.0.lemma' : '.'})

Cherche et affiche 5 titres dont la tête du premier segment est le lemme *rôle*, celle du second segment le lemme *cas* et dont le signe de ponctuation segmentant est un point. Cette requête ne marche que sur un corpus constitué de titres à au moins deux segments.

avg('nb\_segments')

minn('nb\_segments')

maxx('nb\_segments')

Obtient respectivement la moyenne des valeurs, la valeur minimum et la valeur maximum pour la clé *nb\_segments* dans le corpus actuel.

## A5.B Description de nos données informatiques

Nous avons comme données de base un ensemble de 339 687 titres ayant les caractéristiques suivantes :

* identifiant,
* année,
* type de support (article, chapitre ou communication),
* domaine,
* auteurs,
* nombre d’auteurs,
* texte du titre,
* liste de mots et de signes de ponctuation que nous appelons tokens du titre :
  + Pour chaque token :
    - forme
    - étiquette morphosyntaxique
    - lemme (toujours égale à sa forme pour un signe de ponctuation)
    - informations supplémentaires
    - token recteur
    - type de relation de dépendance
    - sa position dans le titre
* longueur du titre en nombre de tokens (mots + signes de ponctuation),
* longueur du titre en nombre de mots uniquement,
* segments :
  + Permet d’accéder aux différents segments du titre et notamment :
    - sa tête,
    - son caractère segmentant (si ce n’est pas un premier segment)
    - la position de la tête dans le titre,
    - la position du caractère segmentant s’il y en a un
* nombre de segments.

## A5.C Analyse de 100 titres traités par Talismane

Nous avons analysé 100 titres traités par Talismane pour vérifier qu’il catégorisait bien les têtes de segments. Nous prenons 20 titres pour chaque structure (nombre de segments et position des têtes dans les segments) qui nous intéresse. Nous indiquons :

* Son identifiant dont la couleur indique le résultat de l’analyse pour le titre :
  + en **vert** si le titre a été analysé correctement en ce qui concerne la détection de têtes de segments,
  + en **orange** si l’analyse de Talismane est discutable mais n’impacte pas notre analyse,
  + en **rouge** si elle est fausse en ne détectant pas la bonne tête de segment,
  + en **violet** si la promotion d’un mot en tête de segment par notre algorithme fait changer le titre de catégorie structurelle,
  + en **rose** si une tête n’a pas été détectée.
* Pour les cinq structures qui nous intéressent, un code segment- tête de la forme :
  + 1\_\_ pour un titre ayant 1 segment et 1 tête,
  + 2\_\_ pour un titre ayant 1 segment et 2 têtes,
  + 1:0 pour un titre ayant 1 tête dans son premier segment et 0 dans son second,
  + 0:1 pour l’inverse,
  + 1:1 pour un titre ayant 1 tête dans chacun de ses deux segments.
* Les têtes de segment sont en gras et :
  + en **vert** si elles sont correctement catégorisées et lemmatisées,
  + en **bleu** si le lemme est incorrect ou inconnu (lemme ignoré pour NPP),
  + en **orange** si la catégorie morphosyntaxique est incorrecte,
  + en **rouge** s’il ne s’agit pas d’une tête,
  + en **violet** si elles ne sont pas détectées par Talismane mais par notre algorithme,
  + en **rose** si elles ne sont pas détectées ni par Talismane ni par notre algorithme.

----------------------------------------------------------------------------------------------

001 **62230** 1\_\_ Un possible **modele** semiotique global de la communication

*Note 01 : L’absence d’accent fait que Talismane n’associe pas ce NC au lemme modèle.*

002 **62250** 1\_\_ L'**IMPACT** DE L'EDITION ELECTRONIQUE SUR LA CRISE DU KOSOVO

003 **460613** 1\_\_ Un **indicateur** de politique d'ouverture à l'immigration

004 **62244** 1\_\_ Le **déplacement** médiatique du débat politique

005 **110369** 1\_\_ L'**imprimerie** et sa diffusion en Extrême-Orient

006 **911256** 1\_\_ Les **enfants** d'Hygie

007 **410464** 1\_\_ **Optimisation** de la précipitation des métaux lourds en mélange

008 **911470** 1\_\_ L'**héritage** du Boiteux d'Orgemont

009 **216325** 1\_\_ **DIFFUSION** INTERGRANULAIRE ET ÉNERGIE DES JOINTS DE GRAINS

010 **760276** 1\_\_ **Dépôt** sec des aérosols à l'interface air-eau

011 **1808328** 1\_\_ **Modélisation** de la structure d'un mélange à haute dilution

012 **1015139** 1\_\_ **Analyse** écophysiologique de la nitrophilie des espèces adventices

013 **264210** 1\_\_ Un **regard** sur les approches basées sur la vision par ordinateur

014 **1759146** 1\_\_ L’**implantation** de l’abbaye de Conques dans les environs de Sainte-Foy-la-Grande

au XIe siècle

015 **215986** 1\_\_ La **persistance** du droit successoral de l'Ancien Régime dans l'Europe du XIXe

siècle

*Note 02 : On remarque que Talismane fait dépendre le* du *de* persistance *plutôt que* Europe *mais*

*cela n’affecte pas notre analyse qui se limite à la tête de segment.*

016 **162355** 1\_\_ **Faut**-il jeter la Méditerranée avec l'eau du bain ?

017 **215983** 1\_\_ La **défense** de la victime en France au XIXe et au XXe siècle

018 **110374** 1\_\_ **Rédaction** de 120 notices

019 **62249** 1\_\_ **Vers** une approche ethnographique des usages des Technologies de l'Information et

de la Communication au sein des petites et moyennes entreprises malaisiennes

*Note 03 : L’enchaînement de compléments de nom peut perde Talismane : il ne sait plus par quoi*

*est régi la préposition* de*. Ici celui avant l’Information est indiqué comme étant*

*régi par* approche *au lieu de* Technologies*. Cela n’a pas d’incidence sur notre*

*travail.*

020 **1808326** 1\_\_ **Algorithme** de construction de modèles markoviens multidimensionnels pour le

mélange des poudres

----------------------------------------------------------------------------------------------

021 **216380** 2\_\_ **DIFFUSION** AVANT ET **ARRIÈRE** D'IONS LOURDS ET MOMENTS ANGULAIRES COMPLEXES

022 **1258669** 2\_\_ **Contenu** et **exigences** du travail

*Note 04 : Talismane normalement ne désigne que le premier NC d’un schéma NC CC NC comme*

*tête. Ici, il désigne les deux NC ce qui n’est pas cohérent.*

023 **312877** 2\_\_ **Demain** la **géographie** sociale.

*Note 05 : La promotion de l’adverbe comme tête est discutable.*

024 **1015192** 2\_\_ **Évaluation** de la dispersion des propriétés mécaniques d'un matériau composite par

sous-**échantillonnage**

*Note 06 : La présence d’un tiret provoque une erreur dans Talismane.*

025 **1808361** 2\_\_ Conditionnement des boues par gel-**dégel**

*Note 07 :* dégel *est désigné comme tête alors que ce n’est clairement pas le cas à cause du*

*tiret.*

026 **264579** 2\_\_ **Institutions** [Les **humanités** et les grandes institutions du savoir en France]

*Note 08 : On peut considérer le texte entre crochets comme un segment non détecté.*

027 **1258688** 2\_\_ **Comparaison** isoenzymatique de deux populations boliviennes (altitude et plaine)

**de** Triatoma infestans (Hemiptera\, Reduviidae)

*Note 09 :* de *est désigné comme tête alors que ce n’est clairement pas le cas.*

028 **162715** 2\_\_ **Transfert** de chaleur et de masse dans une salle d'opérations conditionnée\,

**comparaison** entre deux modes de soufflage

*Note 10 : La virgule n’est pas considérée comme segmentante mais ici elle devrait l’être.*

029 **264613** 2\_\_ **Accès** à l'information et reconnaissance d'un droit à l'information

environnementale - Le nouveau **contexte** juridique international

*Note 11 : Le tiret n’est pas considéré comme segmentant mais ici il devrait l’être. Cela*

*est facilité par la présence d’une majuscule.*

030 **62420** 2\_\_ **De** l'appropriation inachevée du concept de genre (gender) **en** communication

organisationnelle

*Note 12 :* en *est désigné comme tête alors que ce n’est clairement pas le cas.*

031 **216445** 2\_\_ **APPLICATION** DES MÉTHODES STATISTIQUES AU CALCUL DES CHAMPS THERMIQUES TURBULENTS

NON **HOMOGÈNES**

*Note 13 :* HOMOGÈNES *est désigné comme tête alors que ce n’est clairement pas le cas.*

032 **960687** 2\_\_ **Amitiés**\, des **sciences** sociales aux réseaux sociaux de l'internet

033 **216532** 2\_\_ **TRANSITION** MÉTAL-**SEMICONDUCTEUR** DANS LES COMPOSÉS Cr2S3-xSex ET Cr2+εSe3

*Note 14 : La présence d’un tiret provoque une erreur dans Talismane.*

034 **1609898** 2\_\_ Les **Vigiles** **debout**

*Note 15 : Talismane ne devrait prendre que le verbe conjugué.*

035 **960764** 2\_\_ **Misère** de l'hyper-**spécialisation** et dérives du professionnalisme

*Note 16 : La présence d’un tiret provoque une erreur dans Talismane.*

036 **62668** 2\_\_ **Bibliothèques** numériques et Google-**Print**

*Note 17 :* Print *est désigné comme tête alors que ce n’est clairement pas le cas.*

037 **1559698** 2\_\_ Dispositif **de de** caractérisatioon simultanée de l'abondance de pucerons et de la

croissance végétative d'arbres fruitiers

*Note 18 : La répétition de la préposition* de *entraîne une erreur dans Talismane.*

038 **264587** 2\_\_ Le **jeu**\, une **approche** philosophique

*Note 19 : ici, la virgule a une valeur segmentante.*

039 **460685** 2\_\_ Surveillance de chorégraphies de Web Services basées sur WS-**CDL**

*Note 20 : La présence d’un tiret provoque une erreur dans Talismane.*

040 **62434** 2\_\_ **Développement** stratégique du tourisme sportif de rivière par régulation

corporatiste L'**expérience** du bassin de Saint Anne (Québec) appliquée aux Rivières

de Provence

*Note 21 : Oubli d’un point entre les deux segments du titre. La présence d’une majuscule*

*permet de bien repéréer la segmentation manquante.*

----------------------------------------------------------------------------------------------

041 **62397** 1:0 **Réinterroger** les structures documentaires : **de** la numérisation à

l'informatisation

042 **62226** 1:0 Les **temporalités** médiatiques des personnes âgées : des **évolutions** dans la

stabilité

043 **360068** 1:0 La **performativité** de l'évidence : **analyse** du discours néolibéral

*Note 22 : Le mot n’est pas rattaché à son lemme par Talismane car son statut lexical est*

*discutable.*

044 **1061179** 1:0 La **Société** de la Carte géologique de France (1869-1872) : une éphémère

**réaction** à la création du Service de la Carte géologique de la France

045 **360074** 1:0 **Dynamique** technologique controversée et débat démocratique : le **cas** des micros

et nanotechnologies

046 **62256** 1:0 **Traces** de contenus africains sur Internet : **entre** homogénéité et identité

047 **216312** 1:0 **MODÈLES** THÉOTIQUES DE LA STRUCTURE DES JOINTS DE GRAINS.LES **MODÈLES** DE

STRUCTURE DES JOINTS DE GRAINS ET LEUR UTILISATION

*Note 23 : Les deux têtes sont les mêmes.*

048 **1759477** 1:0 Les **objets** communicants\, La **problématique** des Antennes: **Dispositif** **pour**

détecter le vêlage des vaches.

*Note 24 :* pour *est détecter faussement par notre algorithme comme un mot à promouvoir en*

*Tête car* Dispositif *et* pour *sont régis par* objets*. De plus, on a une virgule*

*segmentante, la majuscule qui la suit montrant clairement le début d’un segment.*

*Il s’agit donc d’un titre à trois segments.*

049 **760329** 1:0 L'**omniprésence** de la famille au sein de l'exploitation agricole : une

**situation** de fait encouragé par les règles de droit

050 **1208785** 1:0 **SymbAphidBase** : une **base** de données nouvelle dédiée aux symbiotes de pucerons

pour stocker et visualiser les génomes séquencés en standardisant leurs

annotations

051 **264568** 1:0 **Bill** Viola : **voir** l'eau ou la transparence en mouvement

*Note 25 :* Bill *est caractérisé comme un NC au lieu d’un NPP.*

052 **1759420** 1:0 Les **objets** communicants\, La **problématique** des Antennes; **Balises** de Détresse

*Note 26 : trois problèmes dans ce titre :* problématique *est considérée comme un adjectif, la*

*virgule n’est pas segmentante mais ici elle l’est, et Balises est détecté par*

*notre algorithme. En fait, il s’agit un titre à trois segments et non deux.*

053 **460618** 1:0 **PERCEPTION** DE L'INDÉPENDANCE DE L'AUDITEUR : **ANALYSE** PAR LA THÉORIE

D'ATTRIBUTION

054 **1707597** 1:0 **Élites** maléfiques et ""complot pédophile"" : **paniques** morales autour des enfants

055 **1759142** 1:0 **Formation** et évolution des paroisses de la basse vallée du Drot : **essai** de

synthèse

056 **859899** 1:0 **Classification** floue généralisée : **Application** à la quantification de la stéatose

sur des images histologiques couleurs

057 **510693** 1:0 Les **gastroentérites** aiguës à rotavirus de l'enfant : une **priorité** de santé

publique.

058 **960530** 1:0 **Monde** pluriel : **penser** l'unité des sciences sociales

059 **659177** 1:0 **Reconnaissance** et appropriation : **pour** une anthropologie du travail

060 **62190** 1:0 **Métiers** émergents de la nouvelle économie: **identification** des compétences

attendues et typologie des métiers exercés

----------------------------------------------------------------------------------------------

061 **1660207** 0:1 Quel **pouvoir** de stabilisation à l’échelle de l’UEM : le pacte de stabilité et de

croissance **est**-il viable ?

062 **659285** 0:1 L'**Etat** et les "" autres "" : **comparer** la visibilisation de la main-d'œuvre

immigrée

063 **62609** 0:1 Le **Libre** Accès (Open Access) : **partager** les résultats de la recherche

*Note 27 :* Libre *est caractérisé comme NPP ainsi que* Accès*. On peut se poser la question si*

*ce n’est pas le syntagme nominale entier* Libre Accès *qui devrait être tête.*

064 **960680** 0:1 **De** l'apprenti footballeur **au** petit-rat de l'Opéra : comment les institutions

d'excellence **agissent** face aux dispositions sociales des apprentis ?

*Note 28 : Notre algorithme devrait se contenter de ne prendre que* de.

065 **1258715** 0:1 **Référentiels** de compétences : ce que l'instrument **fait** à la logique compétence

066 **860275** 0:1 La **question** périurbaine : la **repenser** en tenant enfin compte de ce qui motive les

périurbains

067 **62568** 0:1 **Transférabilité** des connaissances : une re-conceptualisation de la distinction

tacite / **explicite**

*Note 29 : Talismane catégorise* explicite *comme V au lieu d’ADJ. De ce fait, il désigne*

explicite *comme tête au lieu* de re-conceptualisation.

068 **264762** 0:1 **Théophile** Gautier : **Regardez**\, mais ne touchez pas (comédie)

*Note 30 : On peut se poser la question si ce n’est pas le syntagme nominal entier*

Théophile Gautier *qui devrait être pris comme tête par notre algorithme.*

069 **1015049** 0:1 Les (**il**)**légalités** ambiguës dans le travail policier : comment l'espace **devient**

prétexte

*Note 31 : l’utilisation du suffixe entre parenthèses* il *perd Talisman. Il le catégorise comme*

*CLS. Notre algorithme ensuite trouve deux mots à prendre pour têtes au lieu d’un.*

070 **1358243** 0:1 **Evolution** de l'arboricolie chez les Cercopithèques: analyse **combinée** de données

moléculaires\, morpho-anatomiques et comportementales

*Note 32 :* combinée *est choisi comme tête alors qu’*analyse *devrait l’être.*

071 **1061109** 0:1 **ImPAC** Lyon : **évaluer** l'impact environnemental et thermique de l'exploitation des

aquifères superficiels pour la climatisation

072 **1759247** 0:1 **Relation** image/**son** : de l'illustration **sonore** à la fusion multi-modale

*Note 33* : sonore *est caractérisé comme V au lieu de ADJ et comme tête alors que de*

*est* de *est un meilleur candidat. On remarque la construction* de X à Y.

*Notre algorithme propose* Relation *est bien la tête du premier segment et*

*incorrectement* son *qui est mal catégorisé : DET au lieu de NC.*

073 **760065** 0:1 **D'**une catastrophe\, l'**autre** : **vivre** avec l'atome

*Note 34 : Notre algorithme détecterait* autre *également comme tête car il est régi par vivre.*

*Mais nous limitons notre algorithme à ne prendre que le premier mot comme tête.*

074 **110247** 0:1 **Vers** une économie des fonctionnalités: **changer** nos rapports avec le produit pour

des économies d'échelle et des nouvelles logiques de responsabilités

075 **809358** 0:1 **Après** la délocalisation...les PME **doivent**-elles relocaliser ?

07 6 **460346** 0:1 Une jeune **fille** changée en jeune homme : homélie sur un miracle survenu dans le

monastère **couvent** de Qartmin\, dans le Tur Abdin

*Note 35 : Erreur classique de confondre le NC* couvent *avec le V couvrir, de plus il ne s’agit*

*pas de la tête de segment,* homélie *y prêtant plus sûrement.*

078 **1060698** 0:1 **Extension** de procédure: ""Le législateur nous **garde** de l'opportunité du juge

079 **312714** 0:1 **Mise** au point sur ""Les cathares devant l'histoire"" et retour sur ""L'histoire

du catharisme en discussion: le débat sur la charte de Niquinta n'est pas **clos**

*Note 36 :* Mise*, détecté par notre algorithme, est catégorisé comme VPP au lieu de NC.*

080 **162674** 0:1 **Communication** financière : quelles **sont** les pratiques des entreprises ?

----------------------------------------------------------------------------------------------

081 **1258625** 1:1 Un nouvel **OVNI** dans le ciel réunionnais : la **transparence** des prix

082 **62241** 1:1 **De** l'anarchisme au combat identitaire : l'**internet** comme média révolutionnaire ?

083 **62366** 1:1 **Communication** et changement organisationnel : le **concept** de chaîne   
 d'appropriation

084 **264580** 1:1 Mystique et **magie** naturelle : les **paysages** mystiques de l'Espagne

*Note 37 :* Mystique *est catégorisée comme ADJ, Talismane privilégie donc le NC* magie *comme*

*tête. Mais il aurait dû soit choisir* Mystique.

085 **216338** 1:1 **MIGRATION** DES JOINTS DE GRAINS.LA **MIGRATION** DES JOINTS INTERGRANULAIRES

*Note 38 : La capitalisation ne pose pas de problème à Talismane. Les deux têtes sont le même*

*mot.*

086 **1609872** 1:1 La **création** d'entreprise en réponse au rêve d'île : l'**ambivalence** d'une   
 attractivité fondée sur le cadre de vie.

087 **659340** 1:1 **Mise** à disposition des données géologiques de surface : **Création** d'un accès sous   
 InfoTerre

*Note 39 : la nominalisation de la locution verbale “mettre à disposition” n’est pas bien   
 catégorisée.*

088 **960668** 1:1 **Brevet** et patrimoine génétique : la **brevetabilité** des organismes génétiquement

modifiés

089 **62616** 1:1 **Projet** DigiCulture : **pour** un portrait des usages et des usagers des ressources

culturelles numériques canadiennes

090 **62386** 1:1 **PRATIQUES** ENONCIATIVES HYPERTEXTUELLES : **VERS** DE NOUVELLES ORGANISATIONS

MEMORIELLES.

091 **110466** 1:1 L**'avenir** de la Common law en français : un **point** de vue d'Europe continentale

092 **1109003** 1:1 **Estimation** des quantiles conditionnels par quantification optimale : nouveaux

**résultats**

093 **1108914** 1:1 **Présentation** d'une langue: le **hongrois**

094 **609991** 1:1 **Variation** du risque de cancer du sein en fonction de la nature de la mutation du

gène ATM. **Étude** familiale rétrospective

095 **62386** 1:1 **PRATIQUES** ENONCIATIVES HYPERTEXTUELLES : **VERS** DE NOUVELLES ORGANISATIONS

MEMORIELLES.

096 **1015246** 1:1 L'**impact** des enceintes urbaines médiévales sur le territoire et ses limites.

L'**exemple** de la Lorraine et de l'Alsace

097 **1258763** 1:1 **Phèdre** janséniste ? **retour** sur un lieu commun (2)

*Note 40 :* Phèdre *n’est pas catégorisée comme un NPP mais comme un NC.*

098 **1409780** 1:1 **Développement** et politique. Le **cas** d'une politique de santé en Géorgie.

099 **62382** 1:1 Quels **modèles** pour la publication sur le web? Le **cas** des contenus informationnels

et culturels.

*Note 41 : Talismane arrive à scinder le* ? *du mot* web.

100 **560355** 1:1 Un **tournant** participatif ? Une **mise** en perspective historique de la participation

du public dans les politiques scientifiques américaines

*Note 42 : Ici, mise est bien reconnu comme une nature nominale.*

# A6. Index des tableaux

Tableau 1: signes de ponctuation segmentants 14

Tableau 2: Distribution des catégories morphosyntaxiques des têtes de segments 19

Tableau 3 : Combinaisons agrégées les plus fréquentes de têtes dans les titres bisegmentaux 20

Tableau 4 : Distribution des structures des titres selon le type 22

Tableau 5 : Distribution des structures des titres selon le nombre d'auteur 22

Tableau 6 : Distribution des structures selon le domaine 24

Tableau 7 : Corrections opérées sur l'étiquetage et la lemmatisation 27

Tableau 8 : Les dix têtes les plus spécifiques de chaque domaine 29

Tableau 9 : Nombre de têtes transdisciplinaires selon le corpus choisi 35

Tableau 10: Tableau d'équivalence entre construction copulative et réduire 44

Tableau 11 : Séquences extraites de l'exemple (42) 50

Tableau 12 : Nombre de séquences des deux bases pour deux différentes tailles maximales de séquences 51

Tableau 13 : Motifs émergeants pour les séquences avec une longueur maximale de trois items 52

Tableau 14 : Motifs émergeants pour les séquences de longueur de quatre items 53

Tableau 15: Suffixes associés au nom noyau du contenu spécifiant 57

Tableau 16: Présence des constructions spécificationnelles classiques dans notre corpus 62

1. Tous les exemples donnés sur fond gris sont tirés de notre corpus de titres. [↑](#footnote-ref-2)
2. Une recherche dans l’archive ouverte HAL sur « Le problème » dans le titre et « Ce problème » dans le résumé donne 192 résultats. La recherche sur « Le problème » dans le titre et « Ce problème » dans le texte intégral donne 336 résultats. Ramené au nombre des 2 millions de références de HAL, cette possibilité de reprise anaphorique est donc très faiblement employée. [↑](#footnote-ref-3)
3. Dans cet exemple, il n’y a pas d’espaces autour du point qui est pourtant bien reconnu comme marque de ponctuation. [↑](#footnote-ref-4)
4. <https://www.cnrtl.fr/definition/outil> [↑](#footnote-ref-5)
5. Dans les faits nous cherchons un nom commun mais nous utilisons cette mise en forme pour mettre le signaler l’emplacement du nom sous-spécifié dans notre schéma. [↑](#footnote-ref-6)
6. Nous ajoutons une capacité à nos schémas : celle de définir conjointement un lemme et une catégorie morphosyntaxique pour un token, toujours en indice dans ce cas, lorsqu’il y a une ambigüité possible. [↑](#footnote-ref-7)