MÉMOIRE DE RECHERCHE

Département des Sciences du Langage

M2 Linguistique, Informatique et Technologies du Langage (LITL)

Transdisciplinarité et sous-spécification des têtes de segments des titres de documents scientifiques

Damien GOUTEUX

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

2018 – 2019

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 le document et les lecteurs et, dans 92 % des cas, le seul : le lecteur ne lira ni le résumé ni l'article après avoir lu le titre (Mabe et Amin, 2002). C'est sur la lecture du titre seul, indépendamment du document titré, que le chercheur fait son tri parmi la littérature scientifique (Goodman et al., 2001). La production scientifique augmente constamment en doublant tous les 12 ans (Stix, cité dans Salager-Meyer et al. 2013). Ce tri effectué sur la lecture du titre soulève la question de l'information qu'il contient et les mots et les structures utilisés pour convoyer cette information. Cet intérêt s'est traduit par de nombreux articles sur les titres en anglais, mais les titres en français ont été moins étudiés, on peut néanmoins citer les travaux de Ho-Dac et al. (2001), Rebeyrolle et al. (2009) et Tanguy et Rebeyrolle (à paraître).

Nous prenons en compte dans notre travail le titre uniquement dans sa fonction informationnelle, considérant qu'elle est la plus importante, soutenu en cela par Haggan (2004) et Hartley (2005). Cette dimension est également la plus facile à analyser. Nous laissons donc de côté la fonction d'attraction qui peut considérablement obscurcir le sens d'un titre (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).

Nous devons en premier lieu revenir sur notre travail effectué durant la première année de master sur les titres de publications scientifiques en français. Nous avions étudié trois schémas fréquents dans un corpus de titres de publications scientifiques. Par schéma, nous entendons une séquence d'éléments pouvant comporter des choix entre plusieurs éléments et des répétitions d'éléments. Un élément peut être une classe grammaticale (nom - N, adjectif qualificatif - ADJ, préposition - P, conjonction de coordination CC…), un sous-ensemble d'une classe (nom commun - NC), un lemme (*et*) ou un signe de ponctuation comme le double point ou le point-virgule. Par exemple, le schéma **NC ADJ** encode la séquence un nom commun suivi d'un adjectif qualificatif.

On dit qu'une séquence de mots et de signes de ponctuation dans un énoncé correspond à un schéma lorsqu'elle se conforme à la séquence décrite par le schéma : il y a correspondance entre la séquence énonciative et la séquence schématique. Les éléments décrits par le schéma peuvent être alors individuellement associés à un mot ou un signe de ponctuation de la séquence énonciative correspondante, on dit que les mots et signes peuplent le schéma. Ainsi la séquence énonciative *Villes durables et changement climatique* correspond, entre autres, au schéma **NC ADJ CC NC ADJ** ainsi qu'au schéma **NC ADJ et NC ADJ**. Le premier n'utilise que des classes grammaticales comme éléments, le second utilise quatre classes grammaticales et un lemme, *et*, comme éléments. La séquence énonciative *union parfaite ou mariage impossible* correspond au premier schéma mais pas au second : *ou* peut être associé à CC mais pas à *et*. Pour les deux schémas, leur premier élément, NC, est associé au mot *Villes* pour le premier exemple et *union* pour le second.

Les trois schémas étudiés dans notre travail précédent étaient :

- un double point suivi d'un un syntagme nominal dont le nom est complémenté par un syntagme prépositionnel soit le schéma **: NC P NC**
- un double point suivi d'un syntagme prépositionnel dont le nom est complémenté par un syntagme prépositionnel soit le schéma **: P NC P NC**
- un double point suivi d'un syntagme nominal constitué de deux noms coordonnés soit le schéma **: NC CC NC**

Nous laissions la possibilité d'avoir des adjectifs qualificatifs pour les noms de chaque schéma mais, par souci de simplification, nous écartons cette possibilité ici. Dans notre corpus de 85 500 titres, le premier schéma couvrait 50% des titres, le deuxième 5% et le dernier 10%, soit une couverture totale de 65 % de notre corpus. Nous avions ensuite étudié les noms et les couples de noms les plus fréquents peuplant ces schémas. Nous avions constaté l'utilisation récurrente et transdisciplinaire de noms abstraits, dont les onze plus fréquents étaient :

- *étude, cas, approche, analyse, application, pratique, exemple, enjeu, perspective, modélisation, limite.*

Tous ces noms semblent liés au domaine scientifique et, sauf *enjeu*, on les retrouve dans le lexique transdisciplinaire des écrits scientifiques (LTES) décrit par Tutin (2008).

Nous avions remarqué que ces noms sont des noms généraux tels que définis par Halliday et Hasan (1976), *« a small set of nouns having generalized reference »*, servant à construire la cohérence du texte. Lexicalement, ces noms appartiennent aux listes de noms généraux fréquemment employés dans un emploi sous-spécifié (NGSS) telles qu'elles ont été définies par Schmid (2000) et Flowerdew et Forest (2015). Un NGSS est *« un nom abstrait dont le sens complet peut seulement être spécifié en référence à son contexte »* (Flowerdew, 2006). Un point important est que le NGSS est un emploi et non une nature lexicale, même si certains noms ont une appétence pour cet emploi, Schmid (2000) les nomme les « *primes »*. Un NGSS possède la particularité d'avoir un faible contenu sémantique et une très large application référentielle. La fréquence et la transdisciplinarité, qui plaident pour un faible contenu sémantique des noms que nous avions repérés, jouent en faveur de l'hypothèse d'un rapprochement possible avec les noms sous-spécifiés.

Pourtant, l'utilisation de cet emploi, dont le trait caractéristique est un faible contenu sémantique du nom, soulève des questions dans un espace comme le titre où chaque mot est compté. De plus, l'emploi de noms de façon sous-spécifiée repose sur leur inclusion dans des constructions spécificationnelles (CS) (Legallois, 2008) qui mettent en rapport le nom sous-spécifié avec un contenu spécificationnel. Les travaux sur les NGSS ont mis en avant deux constructions spécificationnelles fréquemment étudiées (Schmid, 2000 pour l'anglais et Legallois, 2008 pour l'adaptation au français) qui repose toutes deux sur l'utilisation conjuguée du verbe être. Or, de nombreux travaux (Leech, 2000 ; Haggan, 2004 ; Soler, 2007 ; Cheng et al., 2012 ; Wang et Bai, 2007) soulignent la nature nominale des titres. On ne retrouverait donc pas ces CS dans les titres ce qui un argument en défaveur de notre hypothèse : la classe de nom ayant émergé de notre premier travail se rapprocherait-elle de ces NGSS ?

Pour confirmer ou infirmer notre hypothèse, nous voulons étudier un ensemble de caractéristiques qui permettraient de les rapprocher des NGSS. En l'absence de constructions spécificationnelles dans les titres, nos noms s'intègrent-ils néanmoins dans des schémas d'utilisation très

fréquents qui pourraient jouer ce rôle ? Pour répondre à ces questions, nous utiliserons une approche se basant sur le traitement automatique des langues et la linguistique de corpus (Cori et David, 2008).

Tout d'abord, Nous pensons que la classe de nom ayant émergé dans notre premier travail peut gagner à être redéfinie par une autre approche, indépendante de sa position immédiatement après le double point. Nous avions écarté également dans notre précédente étude toute la partie avant le double point et les phénomènes récurrents pouvant y survenir, perdant ainsi des découvertes potentielles. Or, nous faisons l'hypothèse, soutenue par notre intuition et notre connaissance du précédent corpus, que le premier nom que nous étudions immédiatement après le double point est le noyau (ou tête) du syntagme de premier niveau du segment après le double point et donc la tête du segment. Nous redéfinissons donc notre cible d'étude comme les têtes de segment et nous élargissons cette étude, en ne regardant plus seulement le segment après le double point, mais également le segment avant. Nous élargissons également notre étude aux titres à un seul segment et aux titres à deux segments séparés par un autre signe de ponctuation que le double point. Notre étude portera donc sur toutes les têtes nominales des segments des titres à un ou deux segments. Dans l'exemple (1) ci-dessous, le titre est constitué de deux segments, délimité par le double point, avec en gras la tête de chaque segment :

(1) Titre n°1258625 : Un nouvel **OVNI** dans le ciel réunionnais : la **transparence** des prix

Il faut donc commencer ce nouveau travail par découper 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) qui produit un arbre dont la racine est la tête du segment. Une racine dans le cadre de l'analyse syntaxique en dépendances est un mot uniquement régisseur et jamais régi. Ce sont ces têtes dont nous voulons étudier le rapprochement possible avec les noms généraux sous-spécifiés. Pour cela, nous voulons caractériser ces têtes et les schémas récurrents dans lesquels elles s'insèrent, dans un corpus de titres de publications scientifiques, par rapport respectivement aux noms généraux sous-spécifiés et à leurs constructions spécificationnelles.

Nous gardons à l'esprit l'existence de spécificités disciplinaires dans l'écriture des titres pour l'anglais (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 disciplines : il existe en effet des têtes spécifiques à certaines disciplines et d'autres transdisciplinaires. Ce sont ces dernières qui nous semblent potentiellement rapprochables des NGSS.

Notre étude se déroulera en quatre temps. Dans un premier temps, nous délimitons, à partir des données rassemblées, et décrivons notre corpus de travail à l'aide de différentes mesures, en faisant référence aux nombreux travaux existants. Nous nous réassurerons de la nature éminemment nominale des titres. Dans un deuxième temps, nous construisons la liste des têtes de segments propres à certaines disciplines et d'autres qui sont transdisciplinaires. Dans un troisième temps, nous rappelons les apports des travaux sur les noms généraux sous-spécifiés. Nous essayons de détecter les constructions spécificationnelles dans lesquelles ils s'inscrivent généralement avant de montrer les schémas récurrents

effectivement présent dans le corpus. Nous essayons ensuite d'établir une liste de facteur de rapprochement entre nos têtes transdisciplinaires et les emplois en noms sous-spécifiés, entre les constructions spécificationnelles et les schémas récurrents que nous avons. Nous nous appuierons notamment sur leur forte fréquence et leur transdisciplinarité. Nous détaillerons les schémas récurrents au niveau syntaxique et sémantique, que cela soit pour le contenu ou le fonctionnement discursif. Enfin, dans un quatrième temps, nous discutons de nos résultats, des limites de notre travail et ouvrons de nouvelles perspectives.

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

I.1 Origine des données et prétraitement 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. Chaque chercheur, quelle que soit sa discipline, ou documentaliste d'un centre de recherche, est libre de déposer un document sur HAL 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, soit de travaux dont elle possède une copie, soit par le biais d'une notice. Plusieurs institutions, dont le CNRS, encourage le dépôt sur HAL des travaux produits par leurs chercheurs, garantissant un nombre important de titres issus de plusieurs disciplines. 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.

Dans notre matière de départ, 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 peut-être un article scientifique, un chapitre d'un ouvrage collectif ou une communication 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 (2) ci-dessous montre les différentes informations pour un titre donné :

(2) Titre n°609897 : Villes durables et changement climatique : quelques enjeux sur le renouvellement des ressources urbaines ; 2011 ; ART ; 0.sde,1.sde.mcg ; Véronique Peyrache-Gadeau, Bernard Pecqueur.

HAL possède de nombreux types de documents différents. La majorité de la littérature traitant des titres d'articles de journaux scientifiques, notre corpus se limite à ce type de publication et à celles

dont les titres sont construits de manière similaire : chapitres d'ouvrages collectifs et communications dans des conférences.

HAL permet d'attribuer plusieurs domaines à un document. Les domaines sont organisés en un arbre 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 nombreuses disciplines scientifiques, allant de l'histoire aux littératures, alors que toutes les sciences exactes bénéficient elles d'une racine propre comme informatique ou chimie. Tanguy et Rebeyrolle (à paraître) propose 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.

I.1.2 Étiquetage et analyse syntaxique en dépendances

Les titres ont été analysé à l'aide du logiciel Talismane (Urieli et Tanguy, 2013 ; Urieli, 2013) qui fournit un découpage en différents éléments, mots et signes de ponctuation, et réalise un étiquetage morphosyntaxique des mots et une analyse syntaxique en dépendances des éléments. Pour chaque élément 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 élément **régisseur**,
- le **type de dépendance** qui le lie à son régisseur.

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 l'oubli d'un point (3) ou le redoublement d'une préposition (4) :

(3) Titre n°62434 Développement stratégique du tourisme sportif de rivière par régulation corporatiste **L'**expérience du bassin de Saint Anne (Québec) appliquée aux Rivières de Provence

(4) Titre n°1559698 : 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 textes beaucoup plus longs que des titres et entraîné sur de tels textes. On peut donc 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 de textes plus longs, ce qui pourrait pousser Talismane à reconnaître comme verbes des mots n'en étant pas. Nous avons décidé d'inclure une phrase de vérification de l'analyse de Taslimane lors de l'étape de sélection des racines.

I.1.3 Segmentation des titres

Nous avons segmenté les titres selon la liste des signes de ponctuation segmentant établie par Anthony (2001). Nous en retranchons le tiret car il est utilisé pour lier de nombreux mots en français

comme *e-commerce* ou *petit-déjeuner*. 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 segmentant suivant :

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

Tableau 1: signes de ponctuation segmentant

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. Nous qualifions de segmentation forte une segmentation reposant sur une ponctuation forte et de segmentation faible une segmentation reposant sur une ponctuation faible.

L'analyse syntaxique en dépendances effectuée par Talismane ne va pas se comporter pareillement selon le type de segmentation. Une segmentation forte produit en effet deux phrases alors qu'une segmentation faible ne produit qu'une seule phrase. Talismane est bien plus capable de reconnaître une racine dans un segment qui est une phrase que dans un segment qui est une partie de phrase même s'il n'en est pas toujours capable. L'exemple (5) montre une défaillance dans un titre avec segmentation faible et l'exemple (6) montre une défaillance dans un titre avec une segmentation forte :

(5) Titre n°760329 : L'**omniprésence** de la famille au sein de l'exploitation agricole : une *situation* de fait encouragé par les règles de droit

(6) Titre n°216312: **MODÈLES** THÉOTIQUES DE LA STRUCTURE DES JOINTS DE GRAINS.LES *MODÈLES* DE STRUCTURE DES JOINTS DE GRAINS ET LEUR UTILISATION

Dans les deux exemples précédents, *omniprésence* et *modèles* (en gras) sont bien reconnus comme des têtes des premiers segments mais pas *situation* et *modèles* (en italique) pour les seconds segments. En sortie de l'analyse en dépendances de Talismane, nous avons des segments avec ou sans tête, voire plusieurs têtes en cas de comportements anormaux de l'analyseur qui est censé produire un arbre avec une racine unique : dans le cas d'un titre respectant le schéma "NC1 et NC2" par exemple, la racine choisit par Talismane est NC1, *et* dépend de NC1 et NC2 dépend de *et*.

I.1.4 Sélection de la racine des segments

Une fois les segments délimités, il nous faut trouver leur tête. Pour les trouver 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 basant 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. Cette analyse doit ramener pour chaque segment une

tête correctement identifiée. Il s'agit d'une utilisation "a minima" de l'analyse en dépendances pour faire émerger une tête mais cela n'a toutefois pas été sans problème. La segmentation que nous effectuons, basée sur des signes de ponctuation segmentant, est décorrélée de l'analyse de l'outil, nous obtenons donc des segments sans tête. Nous avons décidé de nous limiter aux titres avec au maximum deux segments car ils sont les plus nombreux. On peut classer nos résultats en trois structures segments-racines :

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

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 deux selon que le segment sans racine est le premier et le second. Nous avons également choisi 20 titres ayant un segment et deux racines pour observer cet ensemble et éventuellement tenter d'en reprendre des titres. Nous avons vérifié manuellement pour ces 100 titres le choix de la racine, sa catégorisation morphosyntaxique et son lemme. Les résultats complets sont dans l'annexe A4.B 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 sélectionnner la racine diffère grandement selon la structure segments-racines.

Avant d'aborder les résultats structure par structure, un premier point émerge : 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 l'élément racine de l'arbre de l'analyse en dépendances mais sans qualifier son type de dépendance de racine : il indique « _ » au lieu de « root ». Le second point qui émerge concerne les segments sans racine dans les titres ayant deux segments : on constate l'existence d'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 racine

Sur les 20 titres pris, Talismane a à chaque fois détecté la bonne racine, 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 racines

Sur les 20 titres pris, Talismane a analysé incorrectement 12 titres et 8 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 racine suivie d'un segment sans racine

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 racine de segment.

D. Titres avec un segment sans racine suivi d'un segment avec racine

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 racine suivi d'un segment avec racine

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 lorsque Talismane n'y arrive pas est le suivant :

```
si un mot du segment sans racine est régi par la racine de l'autre
   segment alors
       sélection du premier mot correspondant à cette définition comme
                 tête
sinon
       si le premier mot du segment sans racine est régi par un mot de
          l'autre segment alors
                 sélection du ce mot comme tête
       sinon
                 impossible de trouver une tête
       fin si
fin si
```

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 Description des données et mesures du corpus

I.2.1 Description des données des titres

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,
- énoncé,
- liste de mots et de signes de ponctuation que nous appelons éléments du titre :
    - Pour chaque élément :
        - forme
        - étiquette morphosyntaxique
        - lemme (toujours égale à sa forme pour un signe de ponctuation)
        - informations supplémentaires
        - élément régisseur
        - type de relation de dépendance
        - sa position dans le titre
- longueur du titre en nombre d'éléments (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.

On notera que les différentes données 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. Après avoir décrit nos données nous établissons le périmètre qui délimitera notre corpus de travail.

I.2.2 Sélection des données selon la structure et donc la nature des titres

Établir un périmètre établit dans le matériau de base une dichotomie claire entre ce que nous allons étudier et ce que nous n'étudierons pas. Plus il est large, plus il donne une fondation solide pour la confirmation ou l'infirmation d'hypothèses dessus. Mais plus il est large, plus nous risquons de nous

confronter à des hapax, des phénomènes extrêmement rares 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. Notre périmètre sera constitué sur deux points : la structure des titres, segmentale et racinaire, et la nature des têtes.

A. Structures 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, et le nombre de têtes par segments. Nous nous limiterons aux titres avec au maximum une tête par segment. 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.

A.1 Titres composés d'un seul segment

Exemples de titres :

(7) Titre n°360059 :

1	2	3	4	5	6	7	8
L'	**actualité**	de	la	jurisprudence	communautaire	et	internationale
DET	NC	P	DET	NC	ADJ	CC	ADJ
2	0	2	5	3	5	6	7
det	_	dep	det	prep	mod	coord	dep_coord

(8) Titre n°1258610 :

1	2	3	4	5
Doit	-on	écouter	Björk	?
V	CLS	VINF	NPP	PONCT
0	1	1	3	4
root	suj	obj	obj	ponct

Il y a 171 890 titres composés d'un seul segment ayant une seule tête de segment, soit près de 51 % des titres récupérés initialement.

A.2 Titres composés de deux segments

Exemples de titres :

(9) Titre n°1258625 :

1	2	3	4	5	6	7	8	9	10	11	12

```
Un  nouvel OVNI dans le  ciel réunionnais :     la  transparence des prix
DET ADJ    NC   P    DET NC   ADJ         PONCT DET NC           P+D NC
3   3      0    3    6   4    6           7     10  0            10  11
det mod    _    dep  det prep mod         ponct det _            dep prep


(10) Titre n°360068 :
1   2              3   4   5        6     7       8   9        10
La  performativité de  l'  évidence :     analyse du  discours néolibéral
DET NC             P   DET NC       PONCT NC      P+D NC       ADJ
2   0              2   5   3        5     2       7   8        9
det _              dep det prep     ponct mod     dep prep     mod
```

Il y a 124 938 titres composés de deux segments, soit près de 37 % des titres récupérés initialement. Mais, 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 créé pour suppléer ces limitations. Si dans les exemples précédents, le titre n°1258625 (9) a bien deux segments avec une racine par segment, le titre n°360068 (10) a son second segment sans racine. Notre algorithme promeut *analyse* comme tête du segment car ce mot est uniquement régi par la tête de l'autre segment. Lorsqu'aucun mot du segment sans racine n'est régi par la tête de l'autre segment, nous regardons si le premier mot du segment sans tête est régi par un mot, n'importe lequel, du segment ayant une tête et si oui, nous prenons le mot régi comme tête du segment qui n'en avait pas.

Nous arrivons ainsi, sur les 56 851 titres ayant 2 segments mais une seule racine, à récupérer 46 798 titres soit 82 % d'entre eux. 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 presque 340 000 titres collectés sur HAL. Un corpus de titres ne sera toujours qu'un sous-ensemble de l'infinité des énoncés possibles pour la fonction de titre. Garder un corpus de travail d'une taille importante permet d'essayer de tendre vers la réalité de l'ensemble des énoncés produits, sans jamais y arriver, et renforcer la crédibilité des hypothèses que l'on teste sur lui ou que l'on émet à partir de son observation. 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.

B. Nature des têtes et nature des titres

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), *« 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).

Catégorie morphosyntaxique	Titres monosegmentaux	Titres bisegmentaux, segment 1	Titres bisegmentaux, segment 2
Noms communs	136 734 (80 %)	82 959 (75 %)	84 960 (77 %)
Noms propres	11 094 (6 %)	10 406 (9 %)	4 758 (4 %)
Noms	**147 828 (86 %)**	**93 365 (84 %)**	**89 718 (81 %)**
Verbes à l'indicatif	8 186 (5 %)	3 478 (3 %)	3 513 (3 %)
Verbes à l'infinitif	5 135 (3 %)	6 004 (5 %)	2 140 (2 %)
Verbes	**15 749 (9 %)**	**10 672 (10 %)**	**6 549 (6 %)**
Prépositions	**6 792 (4 %)**	**5 456 (5 %)**	**10 456 (9 %)**

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. Le tableau (3) présente les combinaisons les plus fréquentes couvrant 90 % des titres bisegmentaux :

Catégorie de la tête du premier segment	Catégorie de la tête du second segment	Nombre de titres et pourcentage
NC	NC	63 719 (56 %)
NPP	NC	7 974 (7 %)
NC	P	6 763 (6 %)
VINF	NC	4 643 (4 %)
NC	NPP	3 420 (3 %)
P	NC	3 283 (3 %)
NC	V	2 565 (2 %)

V	NC	2 534 (2 %)
NC	VINF	1 486 (1 %)
NC	P+D	1 295 (1 %)
NC	CC	1 168 (1 %)
P+D	NC	1 030 (1 %)

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

Le tableau (4) suivant agrège les différentes catégories nominales, verbales et prépositionnelles en trois catégories : Nom, Verbe et Préposition.

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 4 : Combinaisons agrégées les plus fréquentes de têtes dans les titres bisegmentaux

Pour les titres monosegmentaux, déterminer la nature du titre revient à prendre la nature de son unique segment. On obtient donc que 86 % de titres nominaux. Pour les titres bisegmentaux, on peut considérer qu'un titre est nominal si son premier segment l'est. On obtient alors 84 % de titres nominaux. Une autre solution 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) relèvent 99 % pour leur corpus.

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 majorité de nos titres et d'éliminer les cas les moins fréquents.

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 Mesures effectuées sur notre corpus de travail

Nous avons défini notre périmètre d'étude comme portant sur les 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 un segment ayant une tête nominale, l'autre ayant une tête verbale, nominale ou prépositionnelle. Nous obtenons un corpus de 250 998 titres, soit 74 % du matériau initial.

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, 49 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 comme nos corpus se répartit 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 %

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 %

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 publié 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 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 suivant 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 %	

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 sciences exactes globalement privilégient plus les titres monosegmentaux que les 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 11 titres, rendant ce calcul très sensible à l'ajout ou au retrait d'un titre dans l'ensemble considéré.

I.3 Conclusion intermédiaire

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 voulons à présent à vouloir étudier les têtes des segments des titres en fonction des différents domaines et étudier leur spécificité ou leur transdisciplinarité.

II. Caractérisation des têtes de segments

Dans cette partie, nous nous intéressons aux têtes de nos segments. Nous avons vu que nous avons une tête par segment et d'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. De plus, nous gardons à l'esprit que nous voulons étudier la variété disciplinaire : nous voulons savoir s'il est possible de mettre avant des têtes très représentatives d'une discipline. Au contraire, nous pouvons aussi circonscrire un ensemble de têtes très fréquentes dans de nombreux domaines, des têtes que nous appellerons transdisciplinaires. Ce sont ces dernières têtes, que nous voulons rapprocher des noms généraux sous-spécifiés dans la partie suivante.

Nous avons 27 domaines sur l'ensemble de notre corpus de travail et 354 168 occurrences de têtes, une tête pour chacun des 147 828 titres monosegmentaux, et deux têtes pour chacun des 103 170 titres bisegmentaux. Nous pouvons regrouper les différentes occurrences d'une même tête sous son lemme. Par la constitution de notre corpus, tous ces lemmes ne sont pas des noms : on autorise également des têtes non nominales dans les titres bisegmentaux si au moins une des têtes est nominale. Nous décidons néanmoins de ne considérer que les têtes nominales dans notre recensement. Nous comptons alors 123 227 lemmes différents. Nous notons $\mathbf{OCC^{TÊTE}}$ le nombre d'occurrences d'un lemme, noté **TÊTE**. Pour chaque **TÊTE**, nous avons :

- un nombre total d'occurrences dans le corpus, noté $\mathbf{NB_{CORPUS}(OCC^{TÊTE})}$,
- une série de nombres d'occurrences par domaine, noté $\mathbf{NB_{DOMAINE}(OCC^{TÊTE})}$,
- on note que la somme de la série des $\mathbf{NB_{DOMAINE}(OCC^{TÊTE})}$ pour une tête donnée et tous les domaines est égale à $\mathbf{NB_{CORPUS}(OCC^{TÊTE})}$,

Pour un domaine donné, on a également :

- un nombre total d'occurrences de têtes, noté $\mathbf{NB_{DOMAINE}(OCC)}$,
- on note que la somme de la série des $\mathbf{NB_{DOMAINE}(OCC^{TÊTE})}$ pour un domaine donné et toutes les têtes de ce domaine est égale à $\mathbf{NB_{DOMAINE}(OCC)}$,
- un nombre total de lemmes de têtes, noté $\mathbf{NB_{DOMAINE}(TÊTE)}$
- on note le nombre de domaines **NBDOM**

On peut prendre comme exemple le domaine de la physique. Il compte 2 471 occurrences de la tête *étude* qui compte 6 842 occurrences en tout dans le corpus. Le domaine physique compte 3 584 têtes différentes et 30 667 occurrences de têtes en tout. On a donc pour la tête *étude* les nombres suivants :

- $\mathbf{NB_{CORPUS}(OCC^{ÉTUDE})}$ = 6 842
- $\mathbf{NB_{PHYSIQUE}(OCC^{ÉTUDE})}$ = 2 471
- $\mathbf{NB_{PHYSIQUE}(OCC)}$ = 30 667
- $\mathbf{NB_{PHYSIQUE}(TÊTE)}$ = 3 584

Pour chaque tête, on peut donc établir deux séries statistiques : l'une est constituée des fréquences de la tête dans les différents domaines par rapport au nombre total de têtes dans le domaine. Pour l'obtenir il faut diviser pour chaque domaine $\mathbf{NB_{DOMAINE}(OCC^{TÊTE})}$ par $\mathbf{NB_{DOMAINE}(OCC)}$. Ce calcul

représente le poids de la tête dans un domaine particulier : plus il est haut, plus la tête est utilisée de façon importante dans ce domaine. On notera la variable **$FREQ_{DOMAINE}(OCC^{TÊTE})$** et la série pour une tête donnée **$FREQ(OCC^{TÊTE})$**. La série pour un domaine donné des fréquences des occurrences de ses têtes s'appellera **$FREQ_{DOMAINE}(OCC)$**.

L'autre série repose sur la répartition du nombre total d'occurrences dans le corpus, **$NB_{CORPUS}(OCC^{TÊTE})$**, dans les différents domaines, soit **$NB_{DOMAINE}(OCC^{TÊTE})$ / $NB_{CORPUS}(OCC^{TÊTE})$**. Nous appelons cette série **$DIST_{DOMAINE}(OCC^{TÊTE})$.** La moyenne de cette série est égale à **$NB_{CORPUS}(OCC^{TÊTE})$** divisé par le nombre de domaines, soit **NBDOM**, ce qui correspondrait à une répartition égalitaire des occurrences entre les différents domaines. Cette hypothèse est intuitivement rejetée, mais nous pouvons essayer de mesurer l'ampleur de la différence entre la répartition réelle et la répartition égalitaire.

Avec ces chiffres à notre disposition, nous pouvons essayer de résoudre deux questions : quelles sont les têtes représentatives d'une discipline et quelles sont les têtes transdisciplinaires, des têtes que l'on retrouve fréquemment dans de nombreuses disciplines ?

II.1 Têtes de segments représentatives

La question des têtes spécifiques à une discipline nous semble intéressant sur deux points. Le premier est de tester si nous pouvons réussir à faire émerger ces têtes et ainsi obtenir une liste de têtes spécifiques par domaine. Le second est de pouvoir la comparer à la liste de têtes transdisciplinaires que nous voulons également établir, pour si, entre têtes spécifiques et têtes transdisciplinaires, il s'agit d'une dichotomie ou d'un problème de seuil.

II.1.1 Définitions théorique et opératoire

Pour être véritablement spécifique à un domaine, une tête ne doit pas seulement y avoir beaucoup d'occurrences : ces occurrences doivent occuper une place importante dans le domaine considéré, la tête doit y être *fréquente*.

Nous commençons par sélectionner les têtes ayant une fréquence relative supérieure ou égale à 0,3 % dans le domaine donné, ce sera notre **seuil de fréquence**, par rapport au nombre de total de têtes dans ce domaine, soit **$FREQ_{DOMAINE}(OCC^{TÊTE})$ >= 0.003**. La valeur du seuil a été déterminée de façon arbitraire après une série d'essais empiriques. Il s'agit d'un premier filtre pour ne garder que les têtes dont le nombre d'occurrences dans le domaine les rend assez fréquente pour être considérées.

Sur certains domaines où il y a très peu de titres et beaucoup de têtes différentes, ce minimum peut ne jamais être atteint. Ainsi si pour un domaine A donné, **$MAX(FREQ_{DOMAINE\ A}(OCC))$ < seuil de fréquence**, aucune tête ne sera sélectionnée. Au contraire, si le filtre est trop bas, toutes les têtes d'un domaine donné pourront être sélectionnées selon la formule **$NB_{DOMAINE}(OCC)$ * seuil de fréquence < 1**, soit si **$NB_{DOMAINE}(OCC)$ < 1 / seuil de fréquence**.

Le seconde filtre que nous appliquons se base sur la différence entre l'hypothétique répartition égalitaire des occurrences d'une tête, soit **$NB_{CORPUS}(OCC^{ÉTUDE})$** / NBDOM, et la répartition réelle, **$NB_{DOMAINE}(OCC^{ÉTUDE})$** : nous prenons l'écart entre la répartition réelle et l'égalitaire que nous divisons par le nombre d'occurrence de la tête dans le corpus pour passer d'un nombre absolu à un pourcentage. On aura donc (**$NB_{DOMAINE}(OCC^{ÉTUDE})$ - $NB_{CORPUS}(OCC^{ÉTUDE})$** / NBDOM) / **$NB_{CORPUS}(OCC^{ÉTUDE})$** qui se simplifie en

$DIST_{DOMAINE}(OCC^{TÊTE})$ - 1 / NBDOM. Nous ne sélectionnons que les têtes pour lesquelles ce calcul dépasse ou égale notre **seuil de distribution**. Nous fixons celui-ci arbitrairement après une série d'essais empiriques à 0.025.

Notre filtre peut donc s'écrire avec le pseudo-code suivant :

```
          Filtre de sélection des têtes spécifiques à des domaines
pour chaque DOMAINE fait
      si NB_DOMAINE(OCC) * seuil de fréquence < 1 alors
          on ne peut pas calculer les têtes spécifiques pour ce domaine
      fin si
      Pour chaque TÊTE fait
          si FREQ_DOMAINE(OCC^TÊTE) >= seuil de fréquence alors
              si DIST_DOMAINE(OCC^TÊTE) - 1 / NBDOM >= seuil de distribution alors
                   sélectionne TÊTE
              fin si
          fin si
      fin pour
fin pour
```

II.1.2 Corrections de Talismane

Néanmoins, ce seuil demeure très faible dans l'absolu et rend notre filtre très sensible à un mauvais traitement d'un lemme par Talismane. Nous avons donc établi un dictionnaire des corrections pour essayer de corriger au maximum des erreurs de catégorisation et lemmatisation. Le tableau suivant liste certaines catégories d'erreurs. Pour savoir comment les corriger, nous avons regardé les différents titres concernés pour établir à chaque fois une règle ad-hoc :

Erreur	**Correction**	**Exemples**
Forme catégorisée comme nom propre avec un lemme inconnu car avec une majuscule.	Lemme ajouté, catégorie corrigée à nom commun.	Effet, Adolescence, Autoformation, Approche, Cohomologie
Forme non reconnue d'un nom commun car erreur d'orthographe	Lemme corrigé	Quantification, événement, indicateus (-r), Synthčse
Forme non reconnue d'un nom commun car caractère non compris	Lemme corrigé (en écrivant oeuvre)	œuvre
Forme non connue d'un nom commun	Lemme ajouté	démotorisation, maritimisation, Compactification,

		Ondelettes
Forme non reconnue d'un nom propre	Lemme corrigé	Paris, Freud
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
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
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)
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
Nom commun anglais non reconnu	Prise en considération de son lemme en français	Synthesis
Emploi d'un nom propre au pluriel	Lemme corrigé à la forme singulière	Venises

Une fois ces corrections effectuées sur notre corpus de travail, nous pouvons passer notre filtre dessus pour obtenir les têtes spécifiques à certaines disciplines, en les classant par leur fréquence dans le domaine.

II.1.3 Résultats et évaluations des résultats

Il est possible de jouer sur les seuils de fréquence et de distribution pour restreindre les têtes sélectionnées au détriment de la richesse du résultat. Plus les seuils seront haut, moins il y aura de têtes spécifiques détectées. Seule l'évaluation des résultats permet de juger de la pertinence des valeurs des seuils et l'efficacité de notre méthode.

Nous avons filtré l'ensemble des têtes du corpus de travail et sélectionné 356 têtes spécifiques. Nous avons ignoré les titres sans domaine unique associé et le domaine autres car il est difficile de trouver des spécifiés à des domaines fourre-tout. De plus le domaine autres n'avait pas assez de titres pour que notre algorithme marche : dans notre cas si $\mathbf{NB_{DOMAINE}(OCC)}$ doit être supérieur à 1 / 0.003, soit 333, ce qui disqualifie le domaine autre qui a seulement 79 titres. Cela nous laisse 25 domaines auxquels nous essayons d'associer des têtes caractéristiques qui sont pris en compte dans nos calculs, on aura **NBDOM** = 25 .

Nous proposons dans le tableau suivant un extrait, classé par domaine, des 356 têtes spécifiques en prenant les dix premières selon l'ordre de leur **$FREQ_{DOMAINE}(OCC^{TÊTE})$**. La liste complète est fournie en annexe. Les noms propres sont en gras. Pour chaque domaine, classés par ordre alphabétique, nous indiquons quatre nombres : combien de lemmes de têtes ont été sélectionnées pour le domaine, le nombre de lemmes de têtes différents dans ce domaine, **$NB_{DOMAINE}$(TÊTE)**, le nombre d'occurrences que représentent les lemmes sélectionnés en pourcentage par rapport au nombre de têtes dans ce domaine et cette dernière valeur, **$NB_{DOMAINE}$(OCC)** :

	Domaine	**Têtes associées**
01	Anthropologie **7** / 2 579 / **4 %** / 6 942	anthropologie, ethnographie, corps, mémoire, patrimoine, identité, objet
02	Archéologie et préhistoire **34** / 3 444 / **21 %** / 13 391	céramique, nécropole, sanctuaire, occupation, sépulture, site, dépôt, site, archéologie, décor, fouille
03	Architecture **15** / 1 624 / **11 %** / 4 629	ambiance, urbanisme, fortification, **Paris**, architecture, quartier, château, habitat, aménagement, ville
04	Art et histoire de l'art **19** / 3 376 / **11 %** / 8 685	vitrail, sculpture, artiste, peinture, cinéma, musique, collection, portrait, théâtre, décor
05	Chimie **19** / 788 / **19 %** / 2 710	ligand, hydrogénation, catalyse, catalyseur, membrane, oxydation, polymère, nanoparticule, réactivité, chimie
06	Droit **38** / 4 189 / **23 %** / 26 398	clause, obligation, juge, droit, assurance, chronique, contrat, responsabilité, commentaire, liberté
07	Économie et finance quantitative **4** / 273 / **2 %** / 489	aversion, **GRP**, complexification, tarification
08	Éducation **37** / 1 786 / **24 %** / 9 445	informatique, éducation, didactique, enseignant, pédagogie, accompagnement, orientation, école, formation, compétence
09	Géographie **3** / 604 / **3 %** / 1 191	démographie, écologie, migration
10	Gestion et management **50** / 3 546 / **36 %** / 25 955	management, déterminant, économie, entreprise, gouvernance, marché, innovation, crise, proposition, impact
11	Histoire **19** / 7 005 / **9 %** / 25 671	femme, mémoire, histoire, société, source, image, compte, ville, remarque, introduction
12	Informatique **49** / 3 281 / **40 %** / 16 241	ordonnancement, algorithme, segmentation, extraction, visualisation, planification, classification, plateforme, reconstruction, détection
13	Linguistique	verbe, grammaire, langue, corpus, nom, dictionnaire,

	35 / 3 435 / **20 %** / 15 512	français, expression, mot, acquisition
14	Littératures **25** / 5 142 / **12 %** / 14 278	roman, poétique, littérature, poésie, fiction, théâtre, lettre, écriture, voix, voyage
15	Mathématiques **21** / 888 / **15 %** / 2 745	cohomologie, package, théorème, régression, algèbre, géométrie, mathématique, assimilation, borne, approximation
16	Philosophie **16** / 2 800 / **9 %** / 7 856	philosophie, épistémologie, critique, idée, éthique, science, réception, concept, logique, vie
17	Physique **47** / 3 603 / **48 %** / 30 667	antenne, spectre, commande, réalisation, couplage, simulation, propriété, détermination, calcul, mesure
18	Planète et Univers **13** / 1 245 / **6 %** / 3 675	géologie, bassin, enregistrement, faune, gisement, légende, datation, variabilité, quantification, fonctionnement
19	Psychologie **8** / 943 / **4 %** / 2 663	autisme, psychanalyse, psychologie, clinique, croyance, différence, enfant, intervention
20	Science politiques **23** / 2 520 / **12 %** / 9 864	parti, élection, démocratie, mobilisation, **Europe**, État, sociologie, justice, politique, acteur
21	Sciences cognitives **4** / 1 164 / **2 %** / 3 141	catégorisation, psychologie, trouble, acquisition
22	Sciences de l'environnement **14** / 1 983 / **10 %** / 7 484	brève, bibliographie, agriculture, valorisation, indicateur, évaluation, changement, conséquence, impact, gestion
23	Sciences de l'information et de la communication **17** / 2 053 / **9 %** / 7 523	média, sémiotique, communication, bibliothèque, médiation, intelligence, appropriation, information, norme, dispositif
24	Sciences du Vivant **49** / 3 800 / **37 %** / 22 149	dosage, composition, influence, effet, intérêt, facteur, conséquence, qualité, variation, utilisation
25	Sociologie **40** / 5 268 / **24 %** / 32 398	géographie, sociologie, territoire, migration, ville, mobilité, paysage, espace, travail, dynamique

Le nombre d'occurrences que représentent les lemmes sélectionnés en pourcentage par rapport au nombre de têtes dans ce domaine permet d'observer la couverture des titres par notre sélection de têtes. Son étendu est de 48 % pour la physique à 2 % pour le domaine des sciences cognitives et le domaine d'économie et finance quantitative.

Il faut à présent évaluer ces résultats. Le premier contrôle que nous pouvons effectuer, bien que très subjectif et limité, et de parcourir nous-même ces têtes pour voir si certaines semblent ne pas correspondre au domaine associé. Ce premier contrôle montre que le filtre semble fonctionner : les mots

semblent effectivement soit des objets d'études des domaines, comme *céramique* et *nécropole* pour l'archéologie, soit des objets supports de l'activité scientifique comme *étude* ou *approche.*

Une méthode d'évaluation des résultats coûteuse en temps aurait été de soumettre l'ensemble des têtes du corpus à un panel de spécialistes de chaque domaine qui auraient ensuite catégorisé chaque tête comme étant propre à leur domaine, soit de façon binaire, soit sur une échelle. La difficulté matérielle de la tâche et sa part de subjectivité, que l'accord inter-annotateur peut néanmoins corriger, ne nous a pas fait considérer cette option, nous privant des mesures de bruit et de rappel.

On peut essayer de mesurer tout d'abord sa sélectivité : le pourcentage de têtes retenues pour chaque domaine, d'abord pour la condition sur le seuil de fréquence puis sur la condition sur le seuil de distribution. On peut calculer le taux de sélection de têtes, en parlant en lemmes et non en occurrences, pour chaque domaine. Pour la première condition, on a un taux de sélection qui va de près de 0,3 % à 29 %, pour une moyenne de 4 % et une médiane de 2 %, un écart-type de 0,057. Nous observons donc une très grande disparité. L'économie et finance quantitative a un taux de sélection de 29 %, suivi de la géographie avec 10 %, puis la chimie, les mathématiques et la psychologie à 7 %. Cela vient du rapport entre le nombre de lemmes de têtes différents par rapport au nombre total de têtes dans ce domaine. Certains domaines utilisent une grande variété de lemmes pour leurs têtes, dans d'autres les occurrences de têtes sont concentrées sur un plus petit nombre de lemmes. L'économie et finance quantitative utilise 273 lemmes de têtes différents pour 489 occurrences de têtes, soit un ratio de 56 %, la géographie en utilise 604 pour 1 191 occurrences, soit un ratio de 51 %. À l'opposé, la physique utilise 3 584 lemmes pour 30 667 occurrences de têtes en tout, soit un ratio de 12 %.

Pour la seconde condition sur $DIST_{DOMAINE}(OCC^{TÊTE})$, on calcule le taux de sélection par rapport à l'ensemble retourné par la première sélection. L'étendue est encore plus grande : de 5 % à 98 %. La moyenne est de 51 % et la médiane de 47 % néanmoins. L'écart-type lui confirme bien cette dispersion, il est de 0,31. Ainsi le domaine gestion et management retient 50 des lemmes de têtes sur les 51 retenus par la précédente condition. À l'inverse, l'économie et finance quantitative et la géographie n'en retiennent que 5 %. Cela s'explique notamment par le très faible nombre de titres, 860 pour la géographie, 346 pour l'économie et finance quantitative, et donc d'occurrences de têtes dans les deux domaines.

Pour l'ensemble des conditions, on oscille entre 2,41 % têtes retenues pour la chimie et 2,36 % pour les mathématiques, contre 0,27 % pour l'anthropologie et l'histoire. Notre filtre est donc très sélectif, et semble choisir des têtes adéquates. Mais on peut se poser la question du partage des têtes spécifiques entre les domaines.

Un autre contrôle possible est en effet de mesurer les collisions entre les domaines : une même tête, peut-elle se retrouver spécifique à plusieurs domaines ? Nous calculons donc pour les 355 têtes le nombre de domaines spécifiques auxquels elles sont associées.

Nombre de domaines associés	**Nombre de têtes**
6	4 1 %
5	11 3 %

4	19 5 %
3	37 10 %
2	55 15 %
1	230 64 %

Nous calculons pour cette série le maximum, 6 pour *état, analyse, évaluation* et *outil*, la moyenne, 1,70, la médiane 1,00 et l'écart-type 1,15 de la série obtenue ainsi. Nous constatons que 64 % des têtes sélectionnées ne sont associées qu'à un seul domaine, et 79 % à un ou deux domaines, ce qui est un bon résultat. Nos têtes spécifiques sont donc très peu partagées entre plusieurs domaines, ce qui amoindrirait leur spécificité. Ce partage n'est pas forcément antithétique de la spécificité. Les têtes étant spécifiques à trois domaines sont par exemple *ville,* partagée par les domaines histoire, architecture et sociologie ce qui nous apparaît comme logique, ou encore *architecture* entre architecture, informatique et art et histoire de l'art. Cela nous amène à une première limite de notre approche : la polysémie de certains termes fait que ce lemme *architecture* qui est une tête commune à ces trois domaines ne fait pas référence à la même. L'*architecture* en informatique peut désigner l'architecture des réseaux, des systèmes, des machines, des processeurs, c'est-à-dire leur agencement en vue d'accomplir leurs buts. Cette polysémie qui se cache sous un lemme unique plaide pour une plus grande spécificité réelle que celle déterminée par notre algorithme.

Une seconde limite qui découle de la première est qu'un sens peut être très spécifique à un domaine et un autre nom : synthèse est ainsi partagée entre physique, informatique et chimie. En ce qui concerne la chimie, *synthèse* désigne la création de façon artificielle d'un composé chimique. Mais *synthèse* désigne aussi un support du travail scientifique très générique : il est difficile de savoir à quel sens rattacher les emplois de ce lemme dans les deux autres domaines. Cela concerne surtout les d'objets supports de l'activité scientifique qui peuvent être partagées, ce qui se retrouvent néanmoins dans les nombreuses têtes spécifiques à plusieurs domaines. La tête *analyse* est ainsi présente dans tous les 25 domaines et apparaît comme une tête spécifique pour six domaines : gestion et management, physique, sciences de l'environnement, sciences du vivant, informatique, éducation et sociologie.

Cette remarque nous nous amène à aborder l'autre facette remarquable des têtes de segments : les têtes transdisciplinaires. Obtenir cette liste permettra également de contrôler encore notre liste de têtes spécifiques en croisant les deux.

II.2 Les têtes de segments transdisciplinaires

II.2.1 Définitions théorique et opératoire

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 avons calculé, pour chaque tête et pour chaque domaine, la fréquence des occurrences la tête par rapport au nombre total d'occurrences de têtes dans ce domaine : **FREQ(OCC$^{\text{TÊTE}}$)**. Nous avons donc pour chaque tête sa série de fréquences pour chaque domaine, **FREQ(OCC$^{\text{TÊTE}}$)**.

Pour trouver les têtes transdisciplinaires, nous nous méfions de la moyenne des fréquences de la tête dans les différents domaines par rapport au nombre total de têtes dans ce domaine : **MOYENNE(FREQ(OCC^{TÊTE}))**. Une moyenne peut en effet cacher des situations très disparates. Nous regardons donc la médiane de la série **FREQ(OCC^{TÊTE})** : plus elle sera élevée, plus la tête sera présente fréquemment dans de nombreux domaines. Nous établissons un seuil arbitraire de 0,001 (0,1 %), que nous nommons **seuil de médiane**, au-dessus duquel nous sélectionnons nos têtes transdisciplinaires.

```
            Filtre de sélection des têtes transdisciplinaires
pour chaque TÊTE fait
    si MEDIANE(FREQ(OCC^TÊTE)) > seuil de médiane alors
        sélectionne TÊTE
    fin si
fin pour
```

II.2.2 Résultats et évaluations du résultat

Sur les 123 227 lemmes de têtes de notre corpus de travail, cela en sélectionne 94 soit 0,08 %. Elles ont en tout 94 738 occurrences, soit près de 27 % des 354 168 occurrences de têtes que comptent notre corpus. Les occurrences de ce très petit nombre de têtes transdisciplinaires concentrent plus d'un quart de toutes les têtes.

Les 20 premières têtes des 94 classés par la médiane de la série **FREQ(OCC^{TÊTE})** 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 en annexe. Aucun nom propre ne figure dans cette liste ce qui est logique, il s'agit de noms communs abstraits.

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 présence d'une forme dans 15 domaines comme marque de sa transdisciplinarité. 15 domaines représentent 60 % des 25 domaines retenus pour nos calculs sur les 27 de notre corpus. Nos 94 têtes transdisciplinaires sont au minimum présentes dans 20 domaines, soit 80 % des 25 domaines. 35 têtes transdisciplinaires sont présentes dans les 25 domaines. Le nombre moyen de domaine où les 94 têtes sont présentes est 23,95 ce qui est extrêmement élevé sachant que le minimum est 20.

Un second contrôle est de le confronter à la liste des noms du lexique transdisciplinaire des écrits scientifiques (LTES) établie par Tutin (2007, 2008). Sur les 94 têtes transdisciplinaires, 74 sont présentes dans le LTES soit 79 %. Les 20 têtes qui ne figurent pas dans le LTES sont : *enjeu, histoire, dynamique, regard, impact, retour, essai, politique, enseignement, note, formation, science, remarque, émergence, point, conception, méthodologie, discours, défi, jeu*. Il nous semble paradoxal que certains lemmes ne figurent pas dans le LTES, surtout ceux sémantiquement liés directement à la science comme *méthodologie* ou *science*. Les autres peuvent avoir été considéré comme trop générique : il en effet difficile de délimiter ce qui est propre à la science, le lexique transdisciplinaire des écrits scientifiques étant considéré comme un sous-ensemble d'un lexique abstrait général (Tutin, 2007).

Un troisième contrôle est de mesurer le croisement entre la liste des têtes spécifiques et les têtes transdisciplinaires. Néanmoins, intuitivement, les deux ensembles ne sont pas forcément disjoints : une tête transdisciplinaire peut être très présente dans une discipline, au point d'en devenir représentative, tout en étant présente dans toutes. Sur nos 94 têtes transdisciplinaires, 89 sont présentes dans la liste des 356 têtes spécifiques, ce qui fait un recouvrement de 95 %. Les 5 têtes qui ne sont pas dans notre précédente classe sont : *an, cadre, défi, enquête* et *perception*. Nous pouvons donc dire que les têtes transdisciplinaires structurent tous les titres, comme l'armature d'un bâtiment dont les domaines seraient les différents étages. Les têtes transdisciplinaires sont tellement présentes qu'elles apparaissent comme spécifique à certains domaines, elles écrasent les spécificités disciplinaires sous leur nombre. Seul le recul apporté par l'étude global de tous les domaines permet de saisir l'image d'ensemble, le fait que l'armature soit bien commune à tous les titres.

II.2.3 Remarques sur les sous-corpus

Nous avons ensuite étudié les têtes transdisciplinaires sur 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.

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*. 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 (5) 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 5 : Nombre de têtes transdisciplinaires selon le corpus choisi

Un fait remarquable du sous-corpus de travail des seconds segments de titres bisegmentaux, c'est que certaines têtes transdisciplinaires sont surreprésentées spécifiquement dans ce corpus. Les occurrences des têtes *cas, exemple, étude, application* et *approche* représentent respectivement 4 %, 3 % et 2 % pour les trois dernières des 95 282 occurrences de têtes de ce sous-corpus. Cette très forte présence ne se rencontre pas dans l'ensemble du corpus et le corpus des premiers segments des titres bisegmentaux. Les occurrences de la tête *étude* du corpus de travail ne représente que 2 % du total des

occurrences de têtes, celles des têtes analyse et étude près de 1 % du corpus des premiers segments des titres bisegmentaux. Uniquement voit-on dans le sous-corpus des titres monosegmentaux poindre *étude* à 3 %. Il y a donc une concentration remarquable sur un petit nombre de têtes dans le sous-corpus des seconds segments de titres bisegmentaux.

On peut également étudier les titres bisegmentaux en prenant les deux têtes ensembles, formant ainsi des couples ordonnés de la forme (tête premier segment, tête second segment). Seuls cinq couples ont une médiane différente de 0 : *(de, exemple), (rôle, cas), (approche, cas), (apport, exemple) et (effet, cas)*. L'apparition de la préposition *de* s'explique car nous exigeons qu'une des têtes soient un nom, l'autre peut être un verbe ou une préposition. La préposition *de* étant la plus fréquente, il est logique qu'elle apparaisse dans les couples les plus fréquents. Cette préposition est utilisée dans des structures de la forme *de … vers …*, *de … à …* étudiées par Tanguy et Rebeyrolle (à paraître), mais comme il ne s'agit pas d'un nom nous l'écartons ici.

Un dernier aspect remarquable bien que plus anecdotique est l'existence de 409 titres dont le premier et le second segment ont le même lemme pour tête, pour achever un effet stylistique de répétition et introduire une comparaison ou un questionnement :

(11) Titre n°271743 : La **crise** ? Quelle **crise** ?

(12) Titre n°1735179 : **Crise** du logement ? Quelle **crise** ?

(13) Titre n°1522267 : **Ville** de jour. **Ville** de nuit

(14) Titre n°183060 : **Linux** embarqué. **Linux** Temps Réel

(15) Titre n°594563 : **Feu** l'arrêt Mercier ! **Feu** l'arrêt Mercier ?

(16) Titre n°740925 : **Corps** dansant. **Corps** glorieux

II.3 Conclusion sur les têtes spécifiques et transdisciplinaires

Après avoir regardé le corpus de travail dans son ensemble et séparément en sous-corpus, nous avons fait émerger d'un côté des têtes spécifiques à des disciplines et de l'autre des têtes transdisciplinaires. Nous avons sélectionné 356 têtes spécifiques pour les 25 domaines différents pris en compte par nos calculs. Nous avons néanmoins remarqué un fort recouvrement avec la seconde classe que nous avons fait émerger : celle des têtes transdisciplinaires. En effet, 84 % des têtes transdisciplinaires sont aussi des têtes spécifiques selon nos filtres.

Nous avons dans cette partie également identifié un petit nombre de têtes transdisciplinaires, 123 en tout si on reprend tous les lemmes identifiés dans les différents sous-corpus, 94 si on applique nos calculs au corpus de travail général. Ces têtes transdisciplinaires sont très fréquentes et donc utilisées dans de nombreux titres de notre corpus de travail et, pour 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. Du fait de leur caractère abstrait et de leur transdisciplinarité, on peut s'interroger sur l'importance de leur contenu sémantique. Que référence-t-on exactement lorsque l'on parle d'une *étude* ou d'un *cas*, d'un *outil* ou d'une *contribution* ? Nous devons à présent présenter un emploi nominal particulier, celui de nom général sous-spécifié, et en quoi cet emploi peut se rapprocher de notre classe de têtes transdisciplinaires.

III. Sous-spécification des têtes transdisciplinaires

III.1 Les noms généraux sous-spécifiés

III.1.1 Définition

De nombreux travaux se sont penchés sur les noms généraux sous-spécifiés (NGSS) en anglais et plus tardivement en français. Si les travaux s'accordent pour définir les NGSS comme un emploi particulier, les définitions théoriques et opératoires de cet emploi sont sujettes à débat, ainsi que la liste des noms pouvant être employé de la sorte, comme le reflète le foisonnement terminologique pour désigner cet emploi : *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).

Comme définition théorique, nous nous proposons de reprendre celle de Flowerdew (2006) pour sa concision et sa clarté (nous traduisons) : *« noms abstraits dont le sens complet peut seulement être spécifié en référence à son contexte »*. 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 mais cette fois-ci pour la conquérir*. Le sens complet de *défi* ne peut être appréhendé qu'en faisant référence au contexte, ici *marcher à nouveau sur la Lune mais cette fois-ci pour la conquérir*.

Pour compléter notre définition théorique, on rappellera également les trois fonctions clés de l'emploi sous-spécifié selon Schmid (2000) :

- Fonction textuelle : capacité de référence quasi-pronominale qui structure le texte.
- Fonction cognitive : création de concepts temporaires.
- Fonction sémantique : catégorisations de concepts, il s'agit d'une mise en perspective par le locuteur qu'il souhaite transmettre à l'interlocuteur.

Prenons deux exemples, (17) et (18), pour éclairer notre hypothèse, celle d'un rapprochement possible entre nos têtes transdisciplinaires et les NGSS :

(17) Titre n°862272 : Le **problème** de l'abandon de l'habitat dans la Corse médiévale

(18) Titre n°201595 : Le **problème** du Paléolithique final de Haute-Normandie

Problème est un terme listé comme pouvant être employé dans un emploi sous-spécifié notamment par Schmid (2000, p. 121) et selon Schmid (2018, p.118) de façon privilégiée et fréquente, ce que l'auteur appelle un « *prime shell noun »* comme *fait, idée, principe, problème, raison* et *chose*. Selon cet auteur, 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 (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 (17) et le Paléolithique final de Haute-Normandie pour (18). 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 référence cataphorique à son contenu spécificationnel dès le titre. Il pourra également en créer des anaphoriques en étant repris, non dans le texte car l'énoncé est trop court pour une reprise, mais dans le résumé ou le texte de la publication scientifique.

III.1.2 Les constructions spécificationnelles

Un NGSS s'insère au sein d'une construction spécificationnelle (CS) qui va relier le NGSS à un contenu spécifiant qui va le « remplir » ou le spécifier. Nous recensons ici les différentes constructions spécificationnelles traditionnelles (Legallois, 2008) de la littérature sur les NGSS, qui sont autant de définitions opératoires des NGSS (Schmid 2000) Nous commençons par les deux CS les plus fréquemment étudiées notamment par Schmid (2000) pour l'anglais et Legallois (2008) pour le français que Schmid appelle (2018, p.120) les *« four major patterns »* :

1. **NGSS** + [verbe être] + *proposition subordonnée complétive [attribut du sujet]* : "le **problème** est *que l'homme souhaite toujours plus*",
2. **NGSS** + [verbe être] + **de** + *proposition subordonnée infinitive* : "le **problème** est **de** *délimiter nos souhaits*".

Le crochet indique à chaque fois une optionnalité du verbe être. Nakamura (2017) ajoute également les trois constructions spécificationnelles suivantes :

3. **NGSS** + verbe être + *syntagme nominal* : "Notre **objectif** majeur est *la rédaction d'une proposition de loi*."
4. Nom + verbe avoir + pour + **NGSS** + **de** + *proposition subordonnée infinitive* : "Cet homme avait pour **ambition** *de devenir président*".
5. **NGSS** + de + *syntagme verbal à l'infinitif* : "L'**ambition** *de devenir président*". Pour le citer "il s'agit de la formation d'un syntagme nominal complexe, qui comporte à la fois la partie sous-spécifiée et la partie spécifiante".

Schmid (2018) indique que son étude n'a pris que les deux premières définitions pour des raisons techniques, mais il atteste dès son livre de 2000 l'existence de la troisième CS décrite par Nakamura, que Flowerdew et Forest (2015) évoquent également. La première CS de Nakamura est intéressante car c'est la seule dont le contenu spécificationnelle n'est pas une proposition mais un syntagme nominal or nos titres sont majoritairement averbaux. Dans nos exemples (17) et (18) il n'y avait aucune occurrence des constructions spécificationnelles décrites. À présent que nous avons rappelé la définition des NGSS et des CS qui les incluent, nous allons essayer de les chercher dans notre corpus.

III.2 Constructions spécificationnelles et schémas récurrents

III.2.1 Les constructions spécificationnelles dans notre corpus

Nous recherchons dans notre corpus les occurrences de ces constructions spécificationnelles. Pour cela nous, utilisons simplement une recherche sur un point saillant des CS. Pour les CS 2, 4 et 5, il s'agit de trouver des titres avec la préposition *de* suivi d'un verbe à l'infinitif. Pour la CS 1, de trouver des

titres avec un nom suivi éventuellement du verbe être conjugué suivi de *que*. Seule la CS 3 demande une recherche un peu plus large sur un nom suivi du verbe être conjugué suivi d'un nom. Nos résultats sont dans le tableau (6), une classe grammaticale est en majuscule (N pour nom, VINF pour verbe à l'infinitif), un lemme en gras et une forme en police standard, un élément optionnel entre crochets, le trait vertical indiquant un choix entre plusieurs éléments.

Schéma	CS correspondante	Nombre de titres
N [**être**] (que | qu')	CS 1	79
N de VINF	CS 2, 4, 5	134
N **être** N	CS 3	838

Tableau 6: Présence des constructions spécificationnelles dans notre corpus

Pour le premier schéma, N [**être**] (que | qu'), nous avons analysé manuellement les 79 titres sélectionnés. Seuls neuf titres mettent en œuvre des NGSS :

(19) Titre n°563566 : Des **grâces** *que Dieu m'a prodiguées* de Jalal al-Din al-Suyuti

(20) Titre n°775708 : Russell de la **logique** *qu'il cherchait* aux **logiques** *qu'il a trouvées*

(21) Titre n°1287097 : Des **bruits** *qu'on ne peut retenir*

(22) Titre n°631315 : Discours savants, discours militants : l'exemple de l'imbroglio occitaniste et les **leçons** d'épistémologie des sciences *que l'on peut en tirer...*

(23) Titre n°963042 : Le risque inondation et les installations industrielles. La **démarche** *que propose l'INERIS*

(24) Titre n°1402849 : Des **images** *que l' on mange*

(25) Titre n°760144 : Tout le **boulot** *qu' on a fait* ...

(26) Titre n°1487619 : Les **questions** juridiques *que posent les Smart Cities*

(27) Titre n°605192 : Deux ou trois **choses** *que je sais d' Émile Picard ...*

Dans les neuf titres, aucun n'utilise le verbe être conjugué qui est optionnel dans la construction spécificationnelle. Cela semble logique dans le type d'énoncé que sont les titres, largement averbaux. On remarque également la petitesse des contenus spécificationnelles : les NGSS peuvent être reliés à de vastes portions de texte mais qui n'existe pas dans les titres, des énoncés beaucoup plus courts. À noter que l'exemple (20) pose un problème car en définissant deux fois *logique*, l'énoncé rend peu clair ce que ce terme désignera ensuite.

Pour le deuxième schéma, nous avons analysé manuellement les 134 titres sélectionnés. Les exemples sont ici bien plus nombreux, nous n'en avons sélectionné qu'une partie de (22) à (40) :

(22) Titre n°1677702 : La métaphore dans l' Introduction à l'**art** *d'écrire de Kaiho Seiryô*

(23) Titre n°913895 : **Façons** *de parler d' Europe*

(24) Titre n°1463154 : Constitution économique et **liberté** *d'entreprendre en Italie*

(25) Titre n°1786205 : La doctrine publiciste et la **faculté** *d'empêcher*

(26) Titre n°640597 : **Pédagogie** *de l'entreprendre*

(27) Titre n°921514 : L'**injonction** *de faire*

(28) Titre n°139615 : Ce cher Montaigne ou le **plaisir** *de lire Montaigne*

(29) Titre n°1018878 : Bergson ou une nouvelle **pensée** *de l'apprendre*

(30) Titre n°279744 : Les **raisons** *de traduire*.

(31) Titre n°1531387 : L'Indianocéanie , un héritage partagé à travers l'**art** *de construire*

(32) Titre n°1531383 : L'indianocéanité, un héritage partagé à travers l'**art** *de construire*

(33) Titre n°112000 : L' évangélisation , une **manière** *de vivre*

(34) Titre n°655539 : Reid , Hume et les **raisons** *d'agir*

(35) Titre n°1062856 : L' **art** *de lire et de se construire*

(36) Titre n°423087 : Jansénisme et **joie** *de vivre*

(37) Titre n°423200 : Universalité et **art** *d'inventer chez Pascal*

(38) Titre n°1623696 : L' Union européenne et la **responsabilité** *de protéger*

(39) Titre n°124028 : Rome face à la **menace** *d'Alexandre le Grand*

(40) Titre n°1781367 Kafala et adoption : une réponse ministérielle tardive est **l'occasion** *de prendre la mesure du nouvel article 21-12, alinéa 3, 1°, du Code civil*

Pour ce deuxième schéma, N de VINF, on constate que sur les trois constructions spécificationnelles qu'il permet de rechercher, on ne retrouve que des occurrences de la CS **NGSS** + de + *syntagme verbal à l'infinitif*. Encore une fois, on remarque la concision du contenu spécificationnel et l'absence du verbe être conjugué optionnel.

Pour le troisième schéma, N **être** N, nous avons également sélectionné quelques exemples de titres sur les 838 récupérés.

(41) Titre n°1811865 : Cas où le **remariage** *est un antidote à une prestation compensatoire*

(42) Titre n°1451883 : Les **non-réponses** *sont -elles des réponses* ? Étude des valeurs manquantes dans un 360˚ feed – back

(43) Titre n°338371 : Neurosciences Computationelles : le **cerveau** *est-il un bon modèle de réseaux de neurones ?*

(44) Titre n°337617 : La haie et le bocage pavillonnaires : la **diversité** végétale *est-elle une utopie en zone urbaine ?*

(45) Titre n°170049 : Le **réseau** technique *est-il un impensé du XVIIIe siècle* : le cas de la poste aux chevaux .

(46) Titre n°339060 : Bouillon de cultures : la **culture** de l' [8]information *est-elle un concept international ?*

L'implémentation du schéma gagnerait à être affinée car de nombreux titres n'entre pas dans notre problématique et le nombre de 838 titres sélectionnés par le schéma est tromper. Ce que l'on remarque immédiatement dans les exemples (41) à (46), c'est la présence de la construction spécificationnelle N **être** N sous la forme d'une question : N **être** (il | elle) N ? Ici, la présence du verbe être est nécessaire, mais nous constatons toujours cette concision du contenu spécificationnel, nécessaire pour loger dans un titre.

Nous n'avons pas dénombré les véritables utilisations de NGSS pour les deux derniers schémas qui comptent 134 et 838 correspondances. Mais même en prenant toutes ces correspondances comme une utilisation de NGSS, ce qu'elles sont loin d'être, on obtient un total de 981 avec les neuf utilisations identifiées pour le premier schéma. À l'échelle de notre corpus de 250 998 titres, si on prend l'hypothèse qu'on ne rencontre qu'une utilisation de NGSS par titre, cela ne représente que 0.4 % des 250 998 titres de notre corpus de travail, soit une infime minorité, rendant ce phénomène très rare.

Nous n'avons trouvé que très peu de construction spécificationnelle classique dans notre corpus. Celles-ci se caractérisent par la non-utilisation du verbe être conjugué lorsqu'il est optionnel et une des contenus spécificationnels très réduits en nombre de mot, allant même jusqu'à un seul. Pourtant, l'existence de nos têtes transdisciplinaires, fréquentes, abstraites, au faible contenu sémantique, 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'insèreraient nos têtes transdisciplinaires et d'évaluer si ceux-ci pourraient jouer le rôle de construction spécificationnelle.

III.2.2 Schémas récurrents d'emploi des têtes transdisciplinaires

TODO

- Établir la forme syntaxico-lemmatique des schémas récurrents. Notamment :
  - Étudier la préférence des têtes transdisciplinaires pour ces segments
  - Étudier si ces schémas acceptent d'autres têtes que les têtes transdisciplinaires et la répartition entre les deux groupes (schéma + tête transdisciplinaire vs schéma + tête non étiquetée comme transdisciplinaire)
  - Étudier les schémas sur deux segments avec le double point (notamment NSS : contenu spécifiant)
- Étudier la sémantique associée à ces schémas
  - Nous essayerons d'aborder la sémantique des schémas, en faisant référence à ce que l'on trouve dans un titre (Grant, 2013 ; Paiva, 2012), mais en nous rapprochant également des typologies sémantiques que l'on plaque sur les titres bisegmentaux comme dans la suite de travaux de Swales et Feak (1994), Anthony (2001) et Cheng et al. (2012) qui partagent une orientation commune.

III.2.3 Transdisciplinarité des schémas

TODO

Dans cette partie nous étudions la répartition des schémas selon les disciplines.

On a montré qu'on ne trouve que très peu les constructions spécificationnelles classiques : elles comptent pour moins de 0,4 % de notre corpus, s'exemptant du verbe être conjugué s'il est optionnel et avec un contenu spécificationnel très réduit. Ces caractéristiques s'accordent bien aux caractéristiques des verbes le plus souvent averbaux et d'une taille réduite. La cinquième construction spécificationnelle que nous avons recensée de Nakamura (2017), **NGSS** + de + *syntagme verbal à l'infinitif*, a pour contenu spécificationnel un groupe prépositionnel ayant un groupe verbal avec un infinitif comme noyau. À la lumière des schémas récurrents où les têtes transdisciplinaires s'insèrent, nous soutenons que ces schémas remplissent le même rôle, mais en utilisant des noms plutôt que des infinitifs. Les contenus spécificationnels des titres seraient des syntagmes nominaux et la construction spécificationnelle prévalente un syntagme nominal complexe formé d'un premier syntagme contenant le NGSS suivi d'un groupe prépositionnel contenant le contenu spécificationnel. Nous allons donc essayer d'établir une liste de facteurs de rapprochement entre les têtes de segments transdisciplinaires et les NGSS pour soutenir cet éventuel rapprochement.

III.3 Rapprochements des NGSS et des têtes transdisciplinaires

III.3.1 Facteurs de rapprochement

TODO

Nous listons dans cette partie une liste de caractéristiques retenues pour rapprocher les têtes de segments des NGSS en les justifiant :

- Fréquence par rapport à l'ensemble des noms.
- S'il s'agit d'un nom abstrait oui ou non.
- Détermination : non définie, définie, pas de détermination.
- Nombre : pluriel ou singulier.
- Complémentation du nom, soit par un syntagme prépositionnel introduit par une autre préposition que *de,* par un groupe introduit par *de* et sans complémentation. Cheng et al. (2012) indique que 90 % des modificateurs des noms sont des groupes prépositionnels, ce qui est une caractéristique de l'écriture académique (Biber et al., 1999 ; Biber et Gray, 2010), et que la majorité de ces groupes utilisent *of* ou *in* comme préposition.
- Transdisciplinarité : Moyenne de la position du lemme dans le classement en fréquence dans les différents domaines. Plus elle sera haute, plus sa transdisciplinarité sera bonne. On fera attention de distinguer sur le sens : sémantique neutre ("modèle", "analyse"), sémantique en rapport avec un seul domaine ("architecture"), sémantique en rapport avec plusieurs domaines ("histoire", "ville"), sémantique mixte interprétable comme neutre mais aussi comme en rapport avec un seul domaine ("synthèse").
- Appartenance à la liste établie par Flowerdew et Forest (2015).
- Appartenance à la liste établie par Schmid (2000).
- Position de la racine dans leur segment. Roze et al. (2014) indique l'existence d'un schéma *Nom sous-spécifié : suite*, ce qui laisse à penser que la position des racines est importante, dans ce juste avant le signe de ponctuation segmentant.

- Position de leur segment dans le titre par rapport aux autres segments.
- Position de la racine dans le titre.

Exemple :

Lemme	Abstrait	Dét.	Compl.	Transdisciplinarité	NGSS fréquent ?	Pos.
cas	TODO					
problème						
objectif						

III.3.2 Règle de rapprochement

TODO

Essayer d'aboutir à une règle pour dire si une tête est un NGSS ou non NGSS et proposer une liste des têtes selon cette règle.

Citer des têtes dont on est sûr qu'ils sont des NSS.

Citer des têtes dont on est sûr qu'ils ne sont pas des NSS.

III.3.3 Résultats et évaluation des résultats

TODO

IV. 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.

IV.1 Têtes spécifiques aux domaines

La première liste obtenue, celle des têtes spécifiques à certains domaines soulèvent trois problèmes. Nous les abordons ici en mettant en avant de potentielles solutions.

IV.1.1 Définition des seuils

Le premier est l'arbitrage des seuils de fréquence et de distribution. Nous avons fixé ces seuils de façon empirique après plusieurs essais. On peut se poser la question d'une meilleure méthode pour les obtenir. Nous avons par exemple fait l'essai en abaissant le seuil de fréquence à 0.001 % et en gardant le seuil de distribution à 0.025 %. Nous obtenions 1 145 lemmes différents au lieu de 356. Une méthode serait de faire varier les valeurs des seuils de fréquence et de distribution et d'évaluer automatiquement le résultat produit. Celui-ci serait mesuré sur deux variables :

- Un indice de collisions des têtes spécifiques qui mesurerait le fait que des têtes soient partagées par plusieurs domaines. Plus il y en a, moins il est intéressant de les catégoriser comme spécifiques.
- Un indice de couverture des titres par les lemmes sélectionnés. Plus celui-ci est important, plus nous couvrons les domaines dont nous prétendons recueillir les objets d'études et supports scientifiques.

En itérant et en faisant varier les seuils, puis en calculant les indices comme un résultat d'efficacité de la combinaison choisie (seuil de fréquence, seuil de distribution), nous pourrions affiner nos valeurs et choisir le meilleur rapport collision / couverture.

De plus, nous nous demandons si des seuils différents devraient être appliqués pour la sélection des noms propres têtes de segment spécifiques à un domaine. Ceux-ci jouent sur des nombres beaucoup plus bas, néanmoins nombreux sont ceux spécifiques à un domaine particulier : les philosophies pour la philosophie, les mathématiciens pour les mathématiques. Néanmoins, dans notre essai avec un seuil de fréquence mis à 0.001 %, nous observions déjà les noms propres Spinoza, Kant, Nietzsche, Foucault, Bergson et Diderot apparaître pour la philosophie. La question se pose donc de savoir si des valeurs différentes pour les seuils permettront de les récupérer sans récupérer trop de noms communs trop peu spécifiques, ou s'il faut des seuils différents, propres aux noms propres.

IV.1.2 Présence des têtes transdisciplinaires dans les têtes spécifiques

Nous avons vu que sur nos 94 têtes transdisciplinaires, 89 sont présentes dans la liste des 356 têtes spécifiques soit 95 %. Si l'on prend les 123 têtes transdisciplinaires relevées par l'analyse des trois sous-corpus, le recouvrement baisse mais reste élevé, à hauteur de 84 %.

La caractéristique des têtes transdisciplinaires est d'être très fréquente dans tous les domaines. De ce fait, il est normal qu'elle apparaisse parfois comme spécifique à un domaine, lorsque leurs occurrences sont très représentées. Comment traiter ce recouvrement ? Une solution serait de retrancher de la liste des 356 têtes spécifiques les têtes transdisciplinaires. Cela entraîne une baisse drastique de la couverture des titres pour certains domaines.

IV.1.3 Topic modeling et catégorisation

Il nous est apparu après-coup que la distinction de têtes spécifiques à des domaines se rapprochait du topic modeling. La différence de notre approche est que nous travaillons sur des titres, des textes beaucoup plus courts que ceux habituellement utilisé pour cette fonction. Il serait intéressant de voir comment se comporte des outils classiques, comme Gensim (https://radimrehurek.com/gensim/) face à ce type de texte. En prenant les têtes du segment, on peut se poser la question de savoir si nous prenons les mots les plus importants du titre, surtout lorsqu'il s'agit d'une tête transdisciplinaire. Comparer les résultats d'un outil classique avec les nôtres donnerait un autre éclairage sur notre travail.

La question se pose également de l'utilisation de la méthode pondération TF-IDF (term frequency – inverse document frequency) là aussi détournée de son utilisation sur des documents pour être utilisée sur des titres. En détectant les termes importants dans un titre par rapport à l'ensemble du corpus, on pourrait peut-être retrouver la liste des têtes spécifiques.

La couverture des têtes spécifiques est assez faible selon le domaine considéré, surtout si l'on en retranche les têtes transdisciplinaires. 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.

IV.1.4 Utilisation des têtes spécifiques

TODO

IV.2 Têtes transdisciplinaires et NGSS

TODO

Sémantique distributive pour étudier les compléments nominaux des NSS.

Conclusion

La première étape de notre travail a été de revenir sur le travail effectué pour notre mémoire de M1 : l'identification de schémas récurrents après le double point dans les titres de publications scientifiques avait mis en avant une classe de noms communs abstraits, très fréquents et pluridisciplinaires. Nous sommes partis de cette découverte pour reformuler le problème et élargir son périmètre en une étude des têtes de segments des titres.

La deuxième étape a été de forger un périmètre de travail au sein du matériau initial, près de 340 000 titres tirés de HAL, qui nous ont été fournis par Tanguy et Rebeyrolle (à paraître) en utilisant la lemmatisation, la catégorisation morphosyntaxique et l'analyse en dépendances syntaxiques fournis par l'outil Talismane (Urieli, 2013). Nous avons opté pour garder les titres monosegmentaux ou bisegmentaux avec à chaque fois une tête par segment. 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, qui disposait lui d'une déjà tête. Nous avons pu conformer à notre règle « un segment une racine » près de 98 % des 56 851 titres auxquels il manquait une racine. Pour finir, nous avons constitué un corpus de travail de 250 998, gardant près de 74 % du matériau initial.

Après avoir délimité notre périmètre et donc notre corpus de travail et identifié toutes les têtes, nous nous sommes interrogés sur leur classe grammaticale. Il s'est avéré que l'extrême majorité des têtes étaient des noms conférant une nature nominale aux titres : 86 % dans le cas des titres monosegmentaux. Dans le cas des titres bisegmentaux, cette majorité est très claire si l'on ne considère que le premier segment, 84 %, beaucoup moins si l'on demande aux deux segments d'avoir un nom pour tête , 68 %.

Partant de cette constatation, nous avons voulu savoir s'il y avait des têtes nominales spécifiques à certains domaines et d'autres qui seraient transdisciplinaires. Nous avons construit à chaque fois un filtre pour sélectionner les deux types de têtes. Sur les 123 227 lemmes de têtes nominales, nous avons trouvé 356 têtes spécifiques et 123 têtes transdisciplinaires. Nous avons remarqué que sur les 123 têtes transdisciplinaires, 86 % appartiennent au lexique transdisciplinaire des écrits scientifiques relevé par Tutin (2008) et 84 % sont également des têtes spécifiques à certains domaines.

Nous avons ensuite essayé de rapprocher les têtes transdisciplinaires, dont la fréquence et la transdisciplinarité impliquent un faible contenu sémantique, des noms généraux sous-spécifiés qui se caractérisent par une très grande fréquence et un faible contenu sémantique également. Ce rapprochement se heurte à l'absence dans notre corpus des constructions spécificationnelles classiques dont la fonction est de lier le nom général sous-spécifié à un contenu spécificationnel présente dans son contexte et qui va le « remplir ».

Faute de construction spécificationnelle classique, nous avons donc étudié es 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 : des énoncés courts, essentiellement averbaux. Nous avons pu montrer que ces schémas récurrents jouent le même rôle que les constructions spécificationnelles classiques sur plusieurs exemples. En nous basant sur plusieurs facteurs de rapprochement, nous avons établi une règle pour détecter les emplois sous-spécifiés dans les titres.

Nous avons néanmoins quelques questions en suspens. Le recouvrement des ensembles de têtes spécifiques et transdisciplinaires pose la question d'une meilleure identification des spécifiques. Des évaluations sur les différentes valeurs des seuils de fréquence et de distribution, avec éventuellement des valeurs spécifiques pour les noms propres, pourraient améliorer notre détection. Sinon, il reste toujours la méthode de retrancher les têtes transdisciplinaires des têtes spécifiques. La question de l'emploi d'autres techniques issues du topic modeling ou de la recherche d'information se pose également pour concurrencer ou compléter nos calculs.

Annexes

A1. Références bibliographiques

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

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.

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.

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

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

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

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, 2016.

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).

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.

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.

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

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

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

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.

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.

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.

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.

A2. Liste des têtes

A2.1 Liste des têtes spécifiques aux domaines

Mathématiques

```
Filter1 :     64 heads /    888 (7.21 %)
Filter2 :     21 heads /    888 (2.36 %)
    lemma             POS     F / dom F / o  occ [ /   dom  /   occ
    estimation         NC   0.037158 0.189  102 [ /  2745  /   541 ] 4
    classification     NC   0.013115 0.143   36 [ /  2745  /   251 ] 2
    théorème           NC   0.009107 0.625   25 [ /  2745  /    40 ] 1
    algorithme         NC   0.008015 0.069   22 [ /  2745  /   320 ] 2
    géométrie          NC   0.008015 0.301   22 [ /  2745  /    73 ] 1
    mathématique       NC   0.007286 0.274   20 [ /  2745  /    73 ] 1
    test               NC   0.007286 0.134   20 [ /  2745  /   149 ] 1
    extension          NC   0.006922 0.094   19 [ /  2745  /   202 ] 2
    sélection          NC   0.006557 0.111   18 [ /  2745  /   162 ] 1
    approximation      NC   0.005100 0.212   14 [ /  2745  /    66 ] 1
    régression         NC   0.005100 0.389   14 [ /  2745  /    36 ] 1
    package            NC   0.004372 0.706   12 [ /  2745  /    17 ] 1
    statistique        NC   0.004007 0.175   11 [ /  2745  /    63 ] 1
    cohomologie        NC   0.004007 1.000   11 [ /  2745  /    11 ] 1
    résolution         NC   0.004007 0.076   11 [ /  2745  /   144 ] 2
    assimilation       NC   0.004007 0.250   11 [ /  2745  /    44 ] 1
    preuve             NC   0.003643 0.085   10 [ /  2745  /   117 ] 1
    existence          NC   0.003279 0.205    9 [ /  2745  /    44 ] 1
    généralisation     NC   0.003279 0.164    9 [ /  2745  /    55 ] 1
    borne              NC   0.003279 0.225    9 [ /  2745  /    40 ] 1
    algèbre            NC   0.003279 0.360    9 [ /  2745  /    25 ] 1
```

Sciences de l'information et de la communication

```
Filter1 :     52 heads /   2053 (2.53 %)
Filter2 :     17 heads /   2053 (0.83 %)
    lemma             POS     F / dom F / o  occ [ /   dom  /   occ
    pratique           NC   0.010900 0.082   82 [ /  7523  /  1005 ] 4
    communication      NC   0.010767 0.288   81 [ /  7523  /   281 ] 1
    usage              NC   0.008108 0.082   61 [ /  7523  /   741 ] 4
    dispositif         NC   0.007311 0.107   55 [ /  7523  /   514 ] 3
    média              NC   0.006912 0.565   52 [ /  7523  /    92 ] 1
    médiation          NC   0.006115 0.250   46 [ /  7523  /   184 ] 1
    bibliothèque       NC   0.005051 0.264   38 [ /  7523  /   144 ] 1
    image              NC   0.004918 0.071   37 [ /  7523  /   521 ] 4
    culture            NC   0.004652 0.090   35 [ /  7523  /   391 ] 1
    art                NC   0.004387 0.076   33 [ /  7523  /   436 ] 3
    discours           NC   0.004254 0.077   32 [ /  7523  /   417 ] 2
    intelligence       NC   0.003855 0.236   29 [ /  7523  /   123 ] 1
    norme              NC   0.003589 0.111   27 [ /  7523  /   243 ] 1
    information        NC   0.003323 0.134   25 [ /  7523  /   187 ] 1
    sémiotique         NC   0.003323 0.463   25 [ /  7523  /    54 ] 1
    appropriation      NC   0.003190 0.155   24 [ /  7523  /   155 ] 1
    mutation           NC   0.003057 0.092   23 [ /  7523  /   250 ] 1
```

Droit

```
Filter1 :     45 heads /   4189 (1.07 %)
Filter2 :     38 heads /   4189 (0.91 %)
    lemma             POS     F / dom F / o  occ [ /   dom  /   occ
    droit              NC   0.041594 0.822 1098 [ / 26398  /  1336 ] 1
    chronique          NC   0.014509 0.732  383 [ / 26398  /   523 ] 1
    responsabilité     NC   0.010721 0.658  283 [ / 26398  /   430 ] 2
```

```
    loi                NC   0.008258 0.544  218 [ / 26398  /   401 ] 1
    protection         NC   0.008220 0.627  217 [ / 26398  /   346 ] 1
    note               NC   0.007917 0.252  209 [ / 26398  /   830 ] 3
    réforme            NC   0.006402 0.412  169 [ / 26398  /   410 ] 2
    réflexion          NC   0.006364 0.123  168 [ / 26398  /  1364 ] 5
    commentaire        NC   0.006288 0.648  166 [ / 26398  /   256 ] 1
    principe           NC   0.005644 0.308  149 [ / 26398  /   483 ] 1
    évolution          NC   0.005417 0.088  143 [ / 26398  /  1631 ] 4
    contrat            NC   0.005341 0.659  141 [ / 26398  /   214 ] 1
    question           NC   0.005114 0.132  135 [ / 26398  /  1021 ] 5
    politique          NC   0.005076 0.126  134 [ / 26398  /  1065 ] 5
    obligation         NC   0.005076 0.893  134 [ / 26398  /   150 ] 1
    action             NC   0.004962 0.254  131 [ / 26398  /   515 ] 3
    juge               NC   0.004887 0.884  129 [ / 26398  /   146 ] 1
    rôle               NC   0.004584 0.080  121 [ / 26398  /  1519 ] 5
    contrôle           NC   0.004584 0.197  121 [ / 26398  /   615 ] 5
    notion             NC   0.004546 0.293  120 [ / 26398  /   410 ] 3
    aspect             NC   0.004432 0.160  117 [ / 26398  /   730 ] 3
    régime             NC   0.004053 0.450  107 [ / 26398  /   238 ] 1
    liberté            NC   0.004053 0.637  107 [ / 26398  /   168 ] 1
    retour             NC   0.004015 0.128  106 [ / 26398  /   827 ] 3
    cour               NC   0.004015 0.564  106 [ / 26398  /   188 ] 1
    condition          NC   0.003788 0.240  100 [ / 26398  /   416 ] 1
    rapport            NC   0.003750 0.215   99 [ / 26398  /   460 ] 2
    état               NC   0.003675 0.097   97 [ / 26398  /   996 ] 6
    procédure          NC   0.003675 0.588   97 [ / 26398  /   165 ] 1
    société            NC   0.003561 0.298   94 [ / 26398  /   315 ] 2
    assurance          NC   0.003447 0.784   91 [ / 26398  /   116 ] 1
    regard             NC   0.003409 0.103   90 [ / 26398  /   873 ] 3
    actualité          NC   0.003371 0.454   89 [ / 26398  /   196 ] 1
    union              NC   0.003371 0.464   89 [ / 26398  /   192 ] 1
    justice            NC   0.003106 0.347   82 [ / 26398  /   236 ] 2
    statut             NC   0.003106 0.280   82 [ / 26398  /   293 ] 1
    clause             NC   0.003106 0.932   82 [ / 26398  /    88 ] 1
    place              NC   0.003031 0.104   80 [ / 26398  /   770 ] 4
```

Linguistique

```
Filter1 :     52 heads /   3435 (1.51 %)
Filter2 :     35 heads /   3435 (1.02 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    cas                 NC   0.022112 0.081  343 [ / 15512  /  4218 ] 3
    exemple             NC   0.014183 0.075  220 [ / 15512  /  2927 ] 4
    langue              NC   0.011153 0.520  173 [ / 15512  /   333 ] 1
    construction        NC   0.008767 0.111  136 [ / 15512  /  1224 ] 5
    réflexion           NC   0.007671 0.087  119 [ / 15512  /  1364 ] 5
    discours            NC   0.007220 0.269  112 [ / 15512  /   417 ] 2
    représentation      NC   0.007027 0.108  109 [ / 15512  /  1013 ] 5
    rôle                NC   0.006962 0.071  108 [ / 15512  /  1519 ] 5
    aspect              NC   0.005802 0.123   90 [ / 15512  /   730 ] 3
    expression          NC   0.005802 0.353   90 [ / 15512  /   255 ] 1
    question            NC   0.005673 0.086   88 [ / 15512  /  1021 ] 5
    problème            NC   0.005544 0.140   86 [ / 15512  /   613 ] 3
    mot                 NC   0.005222 0.351   81 [ / 15512  /   231 ] 1
    nom                 NC   0.005157 0.500   80 [ / 15512  /   160 ] 1
    élément             NC   0.005093 0.094   79 [ / 15512  /   841 ] 5
    traitement          NC   0.004964 0.171   77 [ / 15512  /   449 ] 2
    remarque            NC   0.004771 0.129   74 [ / 15512  /   573 ] 4
    forme               NC   0.004706 0.107   73 [ / 15512  /   680 ] 3
    corpus              NC   0.004577 0.518   71 [ / 15512  /   137 ] 1
    présentation        NC   0.004513 0.121   70 [ / 15512  /   577 ] 1
    variation           NC   0.004384 0.142   68 [ / 15512  /   478 ] 2
```

```
    traduction          NC   0.004384 0.264   68 [ / 15512  /   258 ] 2
    place               NC   0.004319 0.087   67 [ / 15512  /   770 ] 4
    enseignement        NC   0.004319 0.107   67 [ / 15512  /   628 ] 3
    dictionnaire        NC   0.003804 0.492   59 [ / 15512  /   120 ] 1
    structure           NC   0.003610 0.097   56 [ / 15512  /   580 ] 3
    perspective         NC   0.003481 0.078   54 [ / 15512  /   692 ] 2
    notion              NC   0.003352 0.127   52 [ / 15512  /   410 ] 3
    grammaire           NC   0.003352 0.536   52 [ / 15512  /    97 ] 1
    verbe               NC   0.003352 0.765   52 [ / 15512  /    68 ] 1
    point               NC   0.003288 0.100   51 [ / 15512  /   511 ] 2
    interaction         NC   0.003223 0.106   50 [ / 15512  /   473 ] 2
    français            NC   0.003159 0.445   49 [ / 15512  /   110 ] 1
    acquisition         NC   0.003159 0.331   49 [ / 15512  /   148 ] 2
    théorie             NC   0.003030 0.075   47 [ / 15512  /   627 ] 4
```

Gestion et management

```
Filter1 :     51 heads /   3546 (1.44 %)
Filter2 :     50 heads /   3546 (1.41 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    cas                 NC   0.041495 0.255 1077 [ / 25955  /  4218 ] 3
    analyse             NC   0.024542 0.171  637 [ / 25955  /  3725 ] 6
    étude               NC   0.018917 0.078  491 [ / 25955  /  6306 ] 3
    approche            NC   0.016837 0.128  437 [ / 25955  /  3422 ] 4
    rôle                NC   0.012252 0.209  318 [ / 25955  /  1519 ] 5
    impact              NC   0.012098 0.259  314 [ / 25955  /  1213 ] 4
    enjeu               NC   0.011173 0.148  290 [ / 25955  /  1956 ] 4
    économie            NC   0.008553 0.446  222 [ / 25955  /   498 ] 2
    gestion             NC   0.008438 0.241  219 [ / 25955  /   907 ] 5
    apport              NC   0.008284 0.135  215 [ / 25955  /  1588 ] 5
    stratégie           NC   0.008207 0.241  213 [ / 25955  /   883 ] 3
    évaluation          NC   0.008091 0.139  210 [ / 25955  /  1510 ] 6
    modèle              NC   0.007898 0.116  205 [ / 25955  /  1771 ] 5
    effet               NC   0.007706 0.096  200 [ / 25955  /  2094 ] 3
    proposition         NC   0.007205 0.289  187 [ / 25955  /   648 ] 2
    évolution           NC   0.006974 0.111  181 [ / 25955  /  1631 ] 4
    management          NC   0.006935 0.714  180 [ / 25955  /   252 ] 1
    politique           NC   0.006704 0.163  174 [ / 25955  /  1065 ] 5
    application         NC   0.006280 0.071  163 [ / 25955  /  2304 ] 4
    déterminant         NC   0.005856 0.580  152 [ / 25955  /   262 ] 1
    dynamique           NC   0.005163 0.169  134 [ / 25955  /   793 ] 3
    influence           NC   0.005047 0.067  131 [ / 25955  /  1967 ] 3
    développement       NC   0.004970 0.134  129 [ / 25955  /   965 ] 4
    innovation          NC   0.004816 0.353  125 [ / 25955  /   354 ] 1
    gouvernance         NC   0.004816 0.392  125 [ / 25955  /   319 ] 2
    perspective         NC   0.004777 0.179  124 [ / 25955  /   692 ] 2
    pratique            NC   0.004700 0.121  122 [ / 25955  /  1005 ] 4
    théorie             NC   0.004700 0.195  122 [ / 25955  /   627 ] 4
    système             NC   0.004546 0.119  118 [ / 25955  /   995 ] 4
    réflexion           NC   0.004508 0.086  117 [ / 25955  /  1364 ] 5
    état                NC   0.004469 0.116  116 [ / 25955  /   996 ] 6
    contrôle            NC   0.004392 0.185  114 [ / 25955  /   615 ] 5
    crise               NC   0.004277 0.304  111 [ / 25955  /   365 ] 2
    relation            NC   0.004277 0.111  111 [ / 25955  /   997 ] 3
    contribution        NC   0.004123 0.095  107 [ / 25955  /  1131 ] 3
    outil               NC   0.004007 0.106  104 [ / 25955  /   977 ] 6
    entreprise          NC   0.003930 0.423  102 [ / 25955  /   241 ] 1
    enseignement        NC   0.003776 0.156   98 [ / 25955  /   628 ] 3
    comparaison         NC   0.003699 0.100   96 [ / 25955  /   960 ] 4
    mesure              NC   0.003660 0.086   95 [ / 25955  /  1104 ] 3
    introduction        NC   0.003545 0.078   92 [ / 25955  /  1181 ] 4
    intégration         NC   0.003468 0.182   90 [ / 25955  /   494 ] 2
```

```
    organisation       NC   0.003468 0.207   90 [ / 25955  /   434 ] 1
    responsabilité     NC   0.003390 0.205   88 [ / 25955  /   430 ] 2
    marché             NC   0.003352 0.367   87 [ / 25955  /   237 ] 1
    élément            NC   0.003313 0.102   86 [ / 25955  /   841 ] 5
    construction       NC   0.003236 0.069   84 [ / 25955  /  1224 ] 5
    risque             NC   0.003044 0.186   79 [ / 25955  /   425 ] 3
    processus          NC   0.003005 0.168   78 [ / 25955  /   463 ] 2
    bilan              NC   0.003005 0.122   78 [ / 25955  /   637 ] 3
```

Physique

```
Filter1 :     49 heads /   3603 (1.36 %)
Filter2 :     47 heads /   3603 (1.3 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    étude               NC   0.080575 0.392 2471 [ / 30667  /  6306 ] 3
    modélisation        NC   0.035478 0.526 1088 [ / 30667  /  2067 ] 3
    application         NC   0.030293 0.403  929 [ / 30667  /  2304 ] 4
    analyse             NC   0.020185 0.166  619 [ / 30667  /  3725 ] 6
    mesure              NC   0.020022 0.556  614 [ / 30667  /  1104 ] 3
    influence           NC   0.019826 0.309  608 [ / 30667  /  1967 ] 3
    méthode             NC   0.018554 0.366  569 [ / 30667  /  1554 ] 3
    caractérisation     NC   0.017511 0.538  537 [ / 30667  /   999 ] 2
    effet               NC   0.015195 0.223  466 [ / 30667  /  2094 ] 3
    approche            NC   0.014348 0.129  440 [ / 30667  /  3422 ] 4
    modèle              NC   0.012000 0.208  368 [ / 30667  /  1771 ] 5
    simulation          NC   0.011804 0.620  362 [ / 30667  /   584 ] 2
    détermination       NC   0.009815 0.586  301 [ / 30667  /   514 ] 2
    propriété           NC   0.009554 0.588  293 [ / 30667  /   498 ] 1
    conception          NC   0.008152 0.344  250 [ / 30667  /   726 ] 3
    évaluation          NC   0.008087 0.164  248 [ / 30667  /  1510 ] 6
    optimisation        NC   0.007891 0.505  242 [ / 30667  /   479 ] 2
    contribution        NC   0.007793 0.211  239 [ / 30667  /  1131 ] 3
    recherche           NC   0.007728 0.196  237 [ / 30667  /  1207 ] 4
    utilisation         NC   0.007109 0.201  218 [ / 30667  /  1082 ] 3
    comparaison         NC   0.006717 0.215  206 [ / 30667  /   960 ] 4
    comportement        NC   0.006293 0.444  193 [ / 30667  /   435 ] 2
    calcul              NC   0.006261 0.566  192 [ / 30667  /   339 ] 2
    identification      NC   0.005935 0.399  182 [ / 30667  /   456 ] 3
    développement       NC   0.005902 0.188  181 [ / 30667  /   965 ] 4
    diffusion           NC   0.005609 0.478  172 [ / 30667  /   360 ] 1
    spectre             NC   0.005380 0.859  165 [ / 30667  /   192 ] 1
    estimation          NC   0.005348 0.303  164 [ / 30667  /   541 ] 4
    structure           NC   0.005315 0.281  163 [ / 30667  /   580 ] 3
    théorie             NC   0.005120 0.250  157 [ / 30667  /   627 ] 4
    détection           NC   0.004598 0.321  141 [ / 30667  /   439 ] 2
    système             NC   0.004500 0.139  138 [ / 30667  /   995 ] 4
    réalisation         NC   0.004435 0.680  136 [ / 30667  /   200 ] 1
    expérience          NC   0.004370 0.161  134 [ / 30667  /   832 ] 3
    contrôle            NC   0.004337 0.216  133 [ / 30667  /   615 ] 5
    interaction         NC   0.004076 0.264  125 [ / 30667  /   473 ] 2
    remarque            NC   0.003978 0.213  122 [ / 30667  /   573 ] 4
    apport              NC   0.003946 0.076  121 [ / 30667  /  1588 ] 5
    synthèse            NC   0.003489 0.222  107 [ / 30667  /   483 ] 3
    antenne             NC   0.003391 0.912  104 [ / 30667  /   114 ] 1
    dispositif          NC   0.003359 0.200  103 [ / 30667  /   514 ] 3
    état                NC   0.003359 0.103  103 [ / 30667  /   996 ] 6
    impact              NC   0.003293 0.083  101 [ / 30667  /  1213 ] 4
    observation         NC   0.003293 0.226  101 [ / 30667  /   447 ] 3
    commande            NC   0.003261 0.685  100 [ / 30667  /   146 ] 1
    outil               NC   0.003228 0.101   99 [ / 30667  /   977 ] 6
    couplage            NC   0.003163 0.655   97 [ / 30667  /   148 ] 1
```

Anthropologie

Filter1 : 36 heads / 2579 (1.4 %)
Filter2 : 7 heads / 2579 (0.27 %)

lemma	POS	F / dom	F / o	occ	/	dom	/	occ	
anthropologie	NC	0.009795	0.472	68	/	6942	/	144	] 1
corps	NC	0.007203	0.134	50	/	6942	/	373	] 3
mémoire	NC	0.004754	0.093	33	/	6942	/	354	] 3
ethnographie	NC	0.004177	0.341	29	/	6942	/	85	] 1
identité	NC	0.004033	0.068	28	/	6942	/	412	] 2
patrimoine	NC	0.003025	0.090	21	/	6942	/	233	] 1
objet	NC	0.003025	0.065	21	/	6942	/	321	] 2

Histoire

Filter1 : 23 heads / 7005 (0.33 %)
Filter2 : 19 heads / 7005 (0.27 %)

lemma	POS	F / dom	F / o	occ	/	dom	/	occ	
exemple	NC	0.012855	0.113	330	/	25671	/	2927	] 4
histoire	NC	0.012699	0.257	326	/	25671	/	1269	] 4
réflexion	NC	0.006233	0.117	160	/	25671	/	1364	] 5
introduction	NC	0.005415	0.118	139	/	25671	/	1181	] 4
enjeu	NC	0.005064	0.066	130	/	25671	/	1956	] 4
image	NC	0.004441	0.219	114	/	25671	/	521	] 4
femme	NC	0.003934	0.298	101	/	25671	/	339	] 1
question	NC	0.003934	0.099	101	/	25671	/	1021	] 5
politique	NC	0.003895	0.094	100	/	25671	/	1065	] 5
mémoire	NC	0.003779	0.274	97	/	25671	/	354	] 3
regard	NC	0.003701	0.109	95	/	25671	/	873	] 3
ville	NC	0.003506	0.160	90	/	25671	/	562	] 3
compte	NC	0.003467	0.204	89	/	25671	/	436	] 2
construction	NC	0.003389	0.071	87	/	25671	/	1224	] 5
espace	NC	0.003350	0.109	86	/	25671	/	788	] 3
remarque	NC	0.003233	0.145	83	/	25671	/	573	] 4
représentation	NC	0.003116	0.079	80	/	25671	/	1013	] 5
source	NC	0.003116	0.241	80	/	25671	/	332	] 2
société	NC	0.003077	0.251	79	/	25671	/	315	] 2

Sciences de l'environnement

Filter1 : 51 heads / 1983 (2.57 %)
Filter2 : 14 heads / 1983 (0.71 %)

lemma	POS	F / dom	F / o	occ	/	dom	/	occ	
évaluation	NC	0.017504	0.087	131	/	7484	/	1510	] 6
impact	NC	0.012827	0.079	96	/	7484	/	1213	] 4
gestion	NC	0.009220	0.076	69	/	7484	/	907	] 5
outil	NC	0.008685	0.067	65	/	7484	/	977	] 6
bibliographie	NC	0.008418	0.474	63	/	7484	/	133	] 1
brève	NC	0.008418	0.984	63	/	7484	/	64	] 1
agriculture	NC	0.005612	0.205	42	/	7484	/	205	] 1
résultat	NC	0.005211	0.068	39	/	7484	/	577	] 3
changement	NC	0.004543	0.085	34	/	7484	/	402	] 1
risque	NC	0.003875	0.068	29	/	7484	/	425	] 3
valorisation	NC	0.003741	0.163	28	/	7484	/	172	] 1
conséquence	NC	0.003608	0.082	27	/	7484	/	330	] 2
méthodologie	NC	0.003474	0.076	26	/	7484	/	343	] 2
indicateur	NC	0.003207	0.136	24	/	7484	/	177	] 1

Philosophie

Filter1 : 35 heads / 2800 (1.25 %)
Filter2 : 16 heads / 2800 (0.57 %)

lemma	POS	F / dom	F / o	occ	/	dom	/	occ	
histoire	NC	0.011838	0.073	93	/	7856	/	1269	] 4
philosophie	NC	0.010438	0.547	82	/	7856	/	150	] 1

```
    question            NC   0.008783 0.068   69 [ /  7856  /  1021 ] 5
    science             NC   0.008656 0.173   68 [ /  7856  /   393 ] 1
    problème            NC   0.007001 0.090   55 [ /  7856  /   613 ] 3
    critique            NC   0.006746 0.238   53 [ /  7856  /   223 ] 1
    théorie             NC   0.006619 0.083   52 [ /  7856  /   627 ] 4
    concept             NC   0.005474 0.115   43 [ /  7856  /   375 ] 1
    idée                NC   0.003946 0.217   31 [ /  7856  /   143 ] 1
    logique             NC   0.003819 0.102   30 [ /  7856  /   293 ] 2
    notion              NC   0.003691 0.071   29 [ /  7856  /   410 ] 3
    corps               NC   0.003564 0.075   28 [ /  7856  /   373 ] 3
    éthique             NC   0.003310 0.197   26 [ /  7856  /   132 ] 1
    réception           NC   0.003310 0.115   26 [ /  7856  /   226 ] 2
    vie                 NC   0.003182 0.092   25 [ /  7856  /   271 ] 1
    épistémologie       NC   0.003182 0.338   25 [ /  7856  /    74 ] 1
```

Sciences du Vivant

```
Filter1 :     53 heads /   3800 (1.39 %)
Filter2 :     49 heads /   3800 (1.29 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    étude               NC   0.043614 0.153  966 [ / 22149  /  6306 ] 3
    influence           NC   0.031830 0.358  705 [ / 22149  /  1967 ] 3
    effet               NC   0.031830 0.337  705 [ / 22149  /  2094 ] 3
    analyse             NC   0.013590 0.081  301 [ / 22149  /  3725 ] 6
    utilisation         NC   0.012687 0.260  281 [ / 22149  /  1082 ] 3
    évolution           NC   0.012506 0.170  277 [ / 22149  /  1631 ] 4
    évaluation          NC   0.011784 0.173  261 [ / 22149  /  1510 ] 6
    impact              NC   0.010294 0.188  228 [ / 22149  /  1213 ] 4
    comparaison         NC   0.010113 0.233  224 [ / 22149  /   960 ] 4
    recherche           NC   0.009391 0.172  208 [ / 22149  /  1207 ] 4
    rôle                NC   0.009255 0.135  205 [ / 22149  /  1519 ] 5
    application         NC   0.008939 0.086  198 [ / 22149  /  2304 ] 4
    méthode             NC   0.008353 0.119  185 [ / 22149  /  1554 ] 3
    relation            NC   0.007495 0.166  166 [ / 22149  /   997 ] 3
    contribution        NC   0.006908 0.135  153 [ / 22149  /  1131 ] 3
    modélisation        NC   0.006682 0.072  148 [ / 22149  /  2067 ] 3
    intérêt             NC   0.006682 0.324  148 [ / 22149  /   457 ] 1
    apport              NC   0.006592 0.092  146 [ / 22149  /  1588 ] 5
    caractérisation     NC   0.006411 0.142  142 [ / 22149  /   999 ] 2
    résultat            NC   0.005869 0.225  130 [ / 22149  /   577 ] 3
    variation           NC   0.005689 0.264  126 [ / 22149  /   478 ] 2
    modèle              NC   0.005373 0.067  119 [ / 22149  /  1771 ] 5
    développement       NC   0.005328 0.122  118 [ / 22149  /   965 ] 4
    production          NC   0.004695 0.204  104 [ / 22149  /   510 ] 2
    aspect              NC   0.004515 0.137  100 [ / 22149  /   730 ] 3
    essai               NC   0.004470 0.145   99 [ / 22149  /   682 ] 1
    facteur             NC   0.004425 0.306   98 [ / 22149  /   320 ] 1
    conséquence         NC   0.004289 0.288   95 [ / 22149  /   330 ] 2
    mesure              NC   0.004063 0.082   90 [ / 22149  /  1104 ] 3
    état                NC   0.004018 0.089   89 [ / 22149  /   996 ] 6
    estimation          NC   0.004018 0.165   89 [ / 22149  /   541 ] 4
    outil               NC   0.003973 0.090   88 [ / 22149  /   977 ] 6
    système             NC   0.003792 0.084   84 [ / 22149  /   995 ] 4
    action              NC   0.003747 0.161   83 [ / 22149  /   515 ] 3
    détermination       NC   0.003657 0.158   81 [ / 22149  /   514 ] 2
    valeur              NC   0.003657 0.243   81 [ / 22149  /   334 ] 1
    stratégie           NC   0.003612 0.091   80 [ / 22149  /   883 ] 3
    dosage              NC   0.003612 0.842   80 [ / 22149  /    95 ] 1
    gestion             NC   0.003567 0.087   79 [ / 22149  /   907 ] 5
    contrôle            NC   0.003522 0.127   78 [ / 22149  /   615 ] 5
    qualité             NC   0.003431 0.268   76 [ / 22149  /   284 ] 1
    dynamique           NC   0.003386 0.095   75 [ / 22149  /   793 ] 3
```

```
    comportement       NC   0.003386 0.172   75 [ / 22149  /   435 ] 2
    identification     NC   0.003296 0.160   73 [ / 22149  /   456 ] 3
    activité           NC   0.003296 0.213   73 [ / 22149  /   342 ] 2
    bilan              NC   0.003296 0.115   73 [ / 22149  /   637 ] 3
    diversité          NC   0.003296 0.213   73 [ / 22149  /   343 ] 1
    observation        NC   0.003251 0.161   72 [ / 22149  /   447 ] 3
    composition        NC   0.003160 0.461   70 [ / 22149  /   152 ] 1
```

Architecture

```
Filter1 :     53 heads /   1624 (3.26 %)
Filter2 :     15 heads /   1624 (0.92 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    ville               NC   0.016418 0.135   76 [ /  4629  /   562 ] 3
    architecture        NC   0.014690 0.168   68 [ /  4629  /   405 ] 3
    projet              NC   0.013394 0.076   62 [ /  4629  /   820 ] 2
    espace              NC   0.012098 0.071   56 [ /  4629  /   788 ] 3
    ambiance            NC   0.010801 0.794   50 [ /  4629  /    63 ] 1
    mobilité            NC   0.008425 0.105   39 [ /  4629  /   372 ] 2
    urbanisme           NC   0.007777 0.554   36 [ /  4629  /    65 ] 1
    habitat             NC   0.005617 0.139   26 [ /  4629  /   187 ] 2
    aménagement         NC   0.004105 0.138   19 [ /  4629  /   138 ] 1
    Paris              NPP   0.003889 0.171   18 [ /  4629  /   105 ] 1
    château             NC   0.003889 0.151   18 [ /  4629  /   119 ] 2
    maison              NC   0.003672 0.116   17 [ /  4629  /   147 ] 2
    quartier            NC   0.003240 0.158   15 [ /  4629  /    95 ] 1
    imaginaire          NC   0.003024 0.100   14 [ /  4629  /   140 ] 1
    fortification       NC   0.003024 0.264   14 [ /  4629  /    53 ] 1
```

Informatique

```
Filter1 :     58 heads /   3281 (1.77 %)
Filter2 :     49 heads /   3281 (1.49 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    approche            NC   0.032880 0.156  534 [ / 16241  /  3422 ] 4
    application         NC   0.024937 0.176  405 [ / 16241  /  2304 ] 4
    modèle              NC   0.022905 0.210  372 [ / 16241  /  1771 ] 5
    analyse             NC   0.020442 0.089  332 [ / 16241  /  3725 ] 6
    modélisation        NC   0.020319 0.160  330 [ / 16241  /  2067 ] 3
    méthode             NC   0.015639 0.163  254 [ / 16241  /  1554 ] 3
    algorithme          NC   0.014039 0.713  228 [ / 16241  /   320 ] 2
    système             NC   0.013915 0.227  226 [ / 16241  /   995 ] 4
    extraction          NC   0.010344 0.654  168 [ / 16241  /   257 ] 1
    détection           NC   0.010283 0.380  167 [ / 16241  /   439 ] 2
    utilisation         NC   0.010036 0.151  163 [ / 16241  /  1082 ] 3
    outil               NC   0.009852 0.164  160 [ / 16241  /   977 ] 6
    apprentissage       NC   0.009051 0.352  147 [ / 16241  /   418 ] 2
    évaluation          NC   0.008866 0.095  144 [ / 16241  /  1510 ] 6
    gestion             NC   0.008497 0.152  138 [ / 16241  /   907 ] 5
    conception          NC   0.008312 0.186  135 [ / 16241  /   726 ] 3
    optimisation        NC   0.007697 0.261  125 [ / 16241  /   479 ] 2
    construction        NC   0.007327 0.097  119 [ / 16241  /  1224 ] 5
    architecture        NC   0.007081 0.284  115 [ / 16241  /   405 ] 3
    segmentation        NC   0.006958 0.665  113 [ / 16241  /   170 ] 1
    recherche           NC   0.006588 0.089  107 [ / 16241  /  1207 ] 4
    classification      NC   0.006404 0.414  104 [ / 16241  /   251 ] 2
    intégration         NC   0.005726 0.188   93 [ / 16241  /   494 ] 2
    génération          NC   0.005726 0.337   93 [ / 16241  /   276 ] 1
    estimation          NC   0.005665 0.170   92 [ / 16241  /   541 ] 4
    réseau              NC   0.005542 0.148   90 [ / 16241  /   609 ] 2
    représentation      NC   0.005357 0.086   87 [ / 16241  /  1013 ] 5
    simulation          NC   0.005295 0.147   86 [ / 16241  /   584 ] 2
    problème            NC   0.005295 0.140   86 [ / 16241  /   613 ] 3
```

```
    reconnaissance     NC   0.004926 0.265   80 [ / 16241  /   302 ] 1
    ordonnancement     NC   0.004926 0.889   80 [ / 16241  /    90 ] 1
    projet             NC   0.004803 0.095   78 [ / 16241  /   820 ] 2
    comparaison        NC   0.004741 0.080   77 [ / 16241  /   960 ] 4
    état               NC   0.004125 0.067   67 [ / 16241  /   996 ] 6
    proposition        NC   0.003941 0.099   64 [ / 16241  /   648 ] 2
    identification     NC   0.003694 0.132   60 [ / 16241  /   456 ] 3
    calcul             NC   0.003571 0.171   58 [ / 16241  /   339 ] 2
    visualisation      NC   0.003571 0.604   58 [ / 16241  /    96 ] 1
    traitement         NC   0.003510 0.127   57 [ / 16241  /   449 ] 2
    reconstruction     NC   0.003510 0.383   57 [ / 16241  /   149 ] 1
    planification      NC   0.003448 0.467   56 [ / 16241  /   120 ] 1
    contrôle           NC   0.003386 0.089   55 [ / 16241  /   615 ] 5
    adaptation         NC   0.003386 0.170   55 [ / 16241  /   323 ] 1
    extension          NC   0.003325 0.267   54 [ / 16241  /   202 ] 2
    résolution         NC   0.003263 0.368   53 [ / 16241  /   144 ] 2
    amélioration       NC   0.003202 0.236   52 [ / 16241  /   220 ] 1
    synthèse           NC   0.003140 0.106   51 [ / 16241  /   483 ] 3
    plateforme         NC   0.003140 0.389   51 [ / 16241  /   131 ] 1
    méthodologie       NC   0.003017 0.143   49 [ / 16241  /   343 ] 2
```

Éducation

```
Filter1 :     67 heads /   1786 (3.75 %)
Filter2 :     37 heads /   1786 (2.07 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    analyse             NC   0.028692 0.073  271 [ /  9445  /  3725 ] 6
    enjeu               NC   0.014082 0.068  133 [ /  9445  /  1956 ] 4
    éducation           NC   0.013340 0.452  126 [ /  9445  /   279 ] 1
    enseignement        NC   0.012070 0.182  114 [ /  9445  /   628 ] 3
    évaluation          NC   0.011752 0.074  111 [ /  9445  /  1510 ] 6
    pratique            NC   0.011435 0.107  108 [ /  9445  /  1005 ] 4
    formation           NC   0.011435 0.195  108 [ /  9445  /   555 ] 1
    usage               NC   0.007835 0.100   74 [ /  9445  /   741 ] 4
    outil               NC   0.007517 0.073   71 [ /  9445  /   977 ] 6
    expérience          NC   0.006670 0.076   63 [ /  9445  /   832 ] 3
    conception          NC   0.006353 0.083   60 [ /  9445  /   726 ] 3
    place               NC   0.006035 0.074   57 [ /  9445  /   770 ] 4
    apprentissage       NC   0.006035 0.136   57 [ /  9445  /   418 ] 2
    dispositif          NC   0.006035 0.111   57 [ /  9445  /   514 ] 3
    élément             NC   0.005823 0.065   55 [ /  9445  /   841 ] 5
    compétence          NC   0.005823 0.192   55 [ /  9445  /   286 ] 1
    travail             NC   0.005294 0.090   50 [ /  9445  /   558 ] 2
    point               NC   0.004553 0.084   43 [ /  9445  /   511 ] 2
    école               NC   0.004447 0.207   42 [ /  9445  /   203 ] 1
    démarche            NC   0.004341 0.148   41 [ /  9445  /   277 ] 1
    informatique        NC   0.004341 0.519   41 [ /  9445  /    79 ] 1
    rapport             NC   0.004235 0.087   40 [ /  9445  /   460 ] 2
    didactique          NC   0.004023 0.432   38 [ /  9445  /    88 ] 1
    activité            NC   0.003812 0.105   36 [ /  9445  /   342 ] 2
    accompagnement      NC   0.003812 0.333   36 [ /  9445  /   108 ] 1
    parcours            NC   0.003706 0.133   35 [ /  9445  /   263 ] 1
    situation           NC   0.003388 0.162   32 [ /  9445  /   197 ] 1
    transmission        NC   0.003388 0.126   32 [ /  9445  /   253 ] 1
    enseignant          NC   0.003282 0.431   31 [ /  9445  /    72 ] 1
    processus           NC   0.003282 0.067   31 [ /  9445  /   463 ] 2
    environnement       NC   0.003176 0.111   30 [ /  9445  /   271 ] 1
    jeu                 NC   0.003176 0.073   30 [ /  9445  /   410 ] 1
    pédagogie           NC   0.003176 0.417   30 [ /  9445  /    72 ] 1
    dimension           NC   0.003070 0.095   29 [ /  9445  /   306 ] 1
    technologie         NC   0.003070 0.123   29 [ /  9445  /   235 ] 1
    savoir              NC   0.003070 0.157   29 [ /  9445  /   185 ] 1
```

```
    orientation        NC   0.003070 0.212   29 [ /  9445  /   137 ] 1
```

Littératures

```
Filter1 :     34 heads /   5142 (0.66 %)
Filter2 :     25 heads /   5142 (0.49 %)
    lemma             POS    F / dom F / o  occ [ /   dom  /   occ
    littérature        NC   0.009945 0.645  142 [ / 14278  /   220 ] 1
    roman              NC   0.008475 0.791  121 [ / 14278  /   153 ] 1
    histoire           NC   0.007074 0.080  101 [ / 14278  /  1269 ] 4
    écriture           NC   0.006654 0.314   95 [ / 14278  /   303 ] 1
    lecture            NC   0.006303 0.170   90 [ / 14278  /   530 ] 1
    théâtre            NC   0.006163 0.425   88 [ / 14278  /   207 ] 2
    figure             NC   0.006093 0.192   87 [ / 14278  /   454 ] 2
    représentation     NC   0.005673 0.080   81 [ / 14278  /  1013 ] 5
    poésie             NC   0.005113 0.624   73 [ / 14278  /   117 ] 1
    traduction         NC   0.004903 0.271   70 [ / 14278  /   258 ] 2
    voyage             NC   0.004412 0.283   63 [ / 14278  /   223 ] 1
    récit              NC   0.004412 0.236   63 [ / 14278  /   267 ] 1
    fiction            NC   0.003852 0.440   55 [ / 14278  /   125 ] 1
    corps              NC   0.003852 0.147   55 [ / 14278  /   373 ] 3
    note               NC   0.003782 0.065   54 [ / 14278  /   830 ] 3
    poétique           NC   0.003782 0.667   54 [ / 14278  /    81 ] 1
    image              NC   0.003712 0.102   53 [ / 14278  /   521 ] 4
    notice             NC   0.003572 0.178   51 [ / 14278  /   287 ] 2
    art                NC   0.003502 0.115   50 [ / 14278  /   436 ] 3
    mémoire            NC   0.003292 0.133   47 [ / 14278  /   354 ] 3
    remarque           NC   0.003222 0.080   46 [ / 14278  /   573 ] 4
    forme              NC   0.003152 0.066   45 [ / 14278  /   680 ] 3
    lettre             NC   0.003152 0.315   45 [ / 14278  /   143 ] 1
    réception          NC   0.003082 0.195   44 [ / 14278  /   226 ] 2
    voix               NC   0.003012 0.287   43 [ / 14278  /   150 ] 1
```

Sciences cognitives

```
Filter1 :     54 heads /   1164 (4.64 %)
Filter2 :      4 heads /   1164 (0.34 %)
    lemma             POS    F / dom F / o  occ [ /   dom  /   occ
    acquisition        NC   0.005094 0.108   16 [ /  3141  /   148 ] 2
    catégorisation     NC   0.004776 0.221   15 [ /  3141  /    68 ] 1
    trouble            NC   0.003502 0.113   11 [ /  3141  /    97 ] 1
    psychologie        NC   0.003184 0.179   10 [ /  3141  /    56 ] 2
```

Sociologie

```
Filter1 :     43 heads /   5268 (0.82 %)
Filter2 :     40 heads /   5268 (0.76 %)
    lemma             POS    F / dom F / o  occ [ /   dom  /   occ
    cas                NC   0.019075 0.147  618 [ / 32398  /  4218 ] 3
    exemple            NC   0.016976 0.188  550 [ / 32398  /  2927 ] 4
    enjeu              NC   0.011760 0.195  381 [ / 32398  /  1956 ] 4
    approche           NC   0.011235 0.106  364 [ / 32398  /  3422 ] 4
    analyse            NC   0.010031 0.087  325 [ / 32398  /  3725 ] 6
    espace             NC   0.007346 0.302  238 [ / 32398  /   788 ] 3
    évolution          NC   0.007254 0.144  235 [ / 32398  /  1631 ] 4
    ville              NC   0.007161 0.413  232 [ / 32398  /   562 ] 3
    politique          NC   0.007130 0.217  231 [ / 32398  /  1065 ] 5
    construction       NC   0.006636 0.176  215 [ / 32398  /  1224 ] 5
    dynamique          NC   0.006544 0.267  212 [ / 32398  /   793 ] 3
    territoire         NC   0.005988 0.446  194 [ / 32398  /   435 ] 1
    pratique           NC   0.005988 0.193  194 [ / 32398  /  1005 ] 4
    introduction       NC   0.005803 0.159  188 [ / 32398  /  1181 ] 4
    géographie         NC   0.005556 0.662  180 [ / 32398  /   272 ] 1
    réflexion          NC   0.005371 0.128  174 [ / 32398  /  1364 ] 5
```

```
    sociologie          NC   0.005155 0.621  167 [ / 32398  /   269 ] 2
    usage               NC   0.005124 0.224  166 [ / 32398  /   741 ] 4
    regard              NC   0.005000 0.186  162 [ / 32398  /   873 ] 3
    travail             NC   0.004908 0.285  159 [ / 32398  /   558 ] 2
    relation            NC   0.004661 0.151  151 [ / 32398  /   997 ] 3
    apport              NC   0.004599 0.094  149 [ / 32398  /  1588 ] 5
    représentation      NC   0.004568 0.146  148 [ / 32398  /  1013 ] 5
    mobilité            NC   0.004506 0.392  146 [ / 32398  /   372 ] 2
    expérience          NC   0.004445 0.173  144 [ / 32398  /   832 ] 3
    rôle                NC   0.004414 0.094  143 [ / 32398  /  1519 ] 5
    histoire            NC   0.004352 0.111  141 [ / 32398  /  1269 ] 4
    question            NC   0.004321 0.137  140 [ / 32398  /  1021 ] 5
    place               NC   0.004074 0.171  132 [ / 32398  /   770 ] 4
    retour              NC   0.004074 0.160  132 [ / 32398  /   827 ] 3
    paysage             NC   0.004043 0.368  131 [ / 32398  /   356 ] 1
    forme               NC   0.004013 0.191  130 [ / 32398  /   680 ] 3
    modèle              NC   0.003982 0.073  129 [ / 32398  /  1771 ] 5
    développement       NC   0.003797 0.127  123 [ / 32398  /   965 ] 4
    réseau              NC   0.003642 0.194  118 [ / 32398  /   609 ] 2
    gestion             NC   0.003519 0.126  114 [ / 32398  /   907 ] 5
    élément             NC   0.003457 0.133  112 [ / 32398  /   841 ] 5
    migration           NC   0.003457 0.441  112 [ / 32398  /   254 ] 2
    risque              NC   0.003272 0.249  106 [ / 32398  /   425 ] 3
    stratégie           NC   0.003087 0.113  100 [ / 32398  /   883 ] 3
```

Géographie

```
Filter1 :     63 heads /    604 (10.43 %)
Filter2 :      3 heads /    604 (0.5 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    migration           NC   0.015113 0.071   18 [ /  1191  /   254 ] 2
    démographie         NC   0.007557 0.200    9 [ /  1191  /    45 ] 1
    écologie            NC   0.007557 0.085    9 [ /  1191  /   106 ] 1
```

Archéologie et Préhistoire

```
Filter1 :     43 heads /   3444 (1.25 %)
Filter2 :     34 heads /   3444 (0.99 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    exemple             NC   0.015383 0.070  206 [ / 13391  /  2927 ] 4
    site                NC   0.013218 0.596  177 [ / 13391  /   297 ] 1
    apport              NC   0.012994 0.110  174 [ / 13391  /  1588 ] 5
    céramique           NC   0.012546 0.889  168 [ / 13391  /   189 ] 1
    archéologie         NC   0.009559 0.590  128 [ / 13391  /   217 ] 1
    donnée              NC   0.008737 0.386  117 [ / 13391  /   303 ] 1
    bilan               NC   0.008065 0.170  108 [ / 13391  /   637 ] 3
    recherche           NC   0.007542 0.084  101 [ / 13391  /  1207 ] 4
    résultat            NC   0.007468 0.173  100 [ / 13391  /   577 ] 3
    production          NC   0.007244 0.190   97 [ / 13391  /   510 ] 2
    fouille             NC   0.006646 0.536   89 [ / 13391  /   166 ] 1
    habitat             NC   0.006497 0.465   87 [ / 13391  /   187 ] 2
    occupation          NC   0.006348 0.739   85 [ / 13391  /   115 ] 1
    état                NC   0.006124 0.082   82 [ / 13391  /   996 ] 6
    atelier             NC   0.006049 0.445   81 [ / 13391  /   182 ] 1
    dépôt               NC   0.005675 0.618   76 [ / 13391  /   123 ] 1
    élément             NC   0.005377 0.086   72 [ / 13391  /   841 ] 5
    décor               NC   0.005153 0.590   69 [ / 13391  /   117 ] 2
    sanctuaire          NC   0.005003 0.807   67 [ / 13391  /    83 ] 1
    note                NC   0.004481 0.072   60 [ / 13391  /   830 ] 3
    industrie           NC   0.004406 0.360   59 [ / 13391  /   164 ] 1
    nécropole           NC   0.004406 0.819   59 [ / 13391  /    72 ] 1
    sépulture           NC   0.004331 0.699   58 [ / 13391  /    83 ] 1
    maison              NC   0.003883 0.354   52 [ / 13391  /   147 ] 2
```

découverte NC 0.003809 0.425 51 [/ 13391 / 120] 1
église NC 0.003809 0.309 51 [/ 13391 / 165] 1
exploitation NC 0.003734 0.230 50 [/ 13391 / 217] 1
objet NC 0.003435 0.143 46 [/ 13391 / 321] 2
monnaie NC 0.003435 0.390 46 [/ 13391 / 118] 1
château NC 0.003435 0.387 46 [/ 13391 / 119] 2
ensemble NC 0.003360 0.429 45 [/ 13391 / 105] 1
campagne NC 0.003286 0.373 44 [/ 13391 / 118] 1
établissement NC 0.003136 0.400 42 [/ 13391 / 105] 1
inscription NC 0.003062 0.214 41 [/ 13391 / 192] 1

Chimie

Filter1 : 54 heads / 788 (6.85 %)
Filter2 : 19 heads / 788 (2.41 %)

lemma	POS	F / dom	F / o	occ	[/ dom /	occ]
synthèse	NC	0.045387	0.255	123	[/ 2710 /	483] 3
matériau	NC	0.015867	0.279	43	[/ 2710 /	154] 1
structure	NC	0.014391	0.067	39	[/ 2710 /	580] 3
catalyseur	NC	0.014022	0.844	38	[/ 2710 /	45] 1
catalyse	NC	0.013284	0.923	36	[/ 2710 /	39] 1
élaboration	NC	0.012915	0.137	35	[/ 2710 /	256] 1
préparation	NC	0.011070	0.303	30	[/ 2710 /	99] 1
oxydation	NC	0.010701	0.644	29	[/ 2710 /	45] 1
chimie	NC	0.008487	0.442	23	[/ 2710 /	52] 1
spectroscopie	NC	0.006642	0.198	18	[/ 2710 /	91] 1
nanoparticule	NC	0.006273	0.548	17	[/ 2710 /	31] 1
polymère	NC	0.005904	0.571	16	[/ 2710 /	28] 1
réduction	NC	0.004797	0.076	13	[/ 2710 /	172] 1
réactivité	NC	0.004428	0.500	12	[/ 2710 /	24] 1
membrane	NC	0.004059	0.647	11	[/ 2710 /	17] 1
ligand	NC	0.004059	1.000	11	[/ 2710 /	11] 1
réaction	NC	0.003690	0.088	10	[/ 2710 /	114] 1
hydrogénation	NC	0.003690	1.000	10	[/ 2710 /	10] 1
activation	NC	0.003321	0.273	9	[/ 2710 /	33] 1

Planète et Univers

Filter1 : 53 heads / 1245 (4.26 %)
Filter2 : 13 heads / 1245 (1.04 %)

lemma	POS	F / dom	F / o	occ	[/ dom /	occ]
observation	NC	0.008435	0.069	31	[/ 3675 /	447] 3
variabilité	NC	0.007619	0.151	28	[/ 3675 /	185] 1
implication	NC	0.005442	0.085	20	[/ 3675 /	234] 1
géologie	NC	0.004626	0.586	17	[/ 3675 /	29] 1
cartographie	NC	0.003810	0.067	14	[/ 3675 /	209] 1
quantification	NC	0.003810	0.139	14	[/ 3675 /	101] 1
enregistrement	NC	0.003810	0.259	14	[/ 3675 /	54] 1
gisement	NC	0.003537	0.203	13	[/ 3675 /	64] 1
bassin	NC	0.003265	0.300	12	[/ 3675 /	40] 1
datation	NC	0.003265	0.185	12	[/ 3675 /	65] 1
fonctionnement	NC	0.003265	0.135	12	[/ 3675 /	89] 1
faune	NC	0.003265	0.218	12	[/ 3675 /	55] 1
légende	NC	0.003265	0.188	12	[/ 3675 /	64] 1

Art et histoire de l'art

Filter1 : 33 heads / 3376 (0.98 %)
Filter2 : 19 heads / 3376 (0.56 %)

lemma	POS	F / dom	F / o	occ	[/ dom /	occ]
art	NC	0.012781	0.255	111	[/ 8685 /	436] 3
musique	NC	0.011054	0.393	96	[/ 8685 /	244] 1
image	NC	0.009672	0.161	84	[/ 8685 /	521] 4
vitrail	NC	0.009211	0.952	80	[/ 8685 /	84] 1

```
    notice              NC   0.007830 0.237   68 [ /  8685  /   287 ] 2
    théâtre             NC   0.006678 0.280   58 [ /  8685  /   207 ] 2
    architecture        NC   0.006218 0.133   54 [ /  8685  /   405 ] 3
    portrait            NC   0.006102 0.285   53 [ /  8685  /   186 ] 1
    peinture            NC   0.005987 0.456   52 [ /  8685  /   114 ] 1
    compte              NC   0.004260 0.085   37 [ /  8685  /   436 ] 2
    cinéma              NC   0.003915 0.395   34 [ /  8685  /    86 ] 1
    collection          NC   0.003685 0.286   32 [ /  8685  /   112 ] 1
    décor               NC   0.003569 0.265   31 [ /  8685  /   117 ] 2
    artiste             NC   0.003454 0.500   30 [ /  8685  /    60 ] 1
    figure              NC   0.003454 0.066   30 [ /  8685  /   454 ] 2
    livre               NC   0.003339 0.132   29 [ /  8685  /   220 ] 1
    sculpture           NC   0.003339 0.580   29 [ /  8685  /    50 ] 1
    oeuvre              NC   0.003224 0.206   28 [ /  8685  /   136 ] 1
    source              NC   0.003109 0.081   27 [ /  8685  /   332 ] 2
```

Psychologie

```
Filter1 :     66 heads /    943 (7.0 %)
Filter2 :      8 heads /    943 (0.85 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    psychologie         NC   0.007510 0.357   20 [ /  2663  /    56 ] 2
    enfant              NC   0.005633 0.098   15 [ /  2663  /   153 ] 1
    différence          NC   0.005257 0.113   14 [ /  2663  /   124 ] 1
    intervention        NC   0.004506 0.090   12 [ /  2663  /   133 ] 1
    psychanalyse        NC   0.004506 0.387   12 [ /  2663  /    31 ] 1
    autisme             NC   0.004131 0.423   11 [ /  2663  /    26 ] 1
    clinique            NC   0.003380 0.300    9 [ /  2663  /    30 ] 1
    croyance            NC   0.003004 0.170    8 [ /  2663  /    47 ] 1
```

Science politique

```
Filter1 :     49 heads /   2520 (1.94 %)
Filter2 :     23 heads /   2520 (0.91 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
    politique           NC   0.016322 0.151  161 [ /  9864  /  1065 ] 5
    introduction        NC   0.009732 0.081   96 [ /  9864  /  1181 ] 4
    retour              NC   0.009732 0.116   96 [ /  9864  /   827 ] 3
    élection            NC   0.008110 0.556   80 [ /  9864  /   144 ] 1
    usage               NC   0.005779 0.077   57 [ /  9864  /   741 ] 4
    sociologie          NC   0.005474 0.201   54 [ /  9864  /   269 ] 2
    démocratie          NC   0.005474 0.314   54 [ /  9864  /   172 ] 1
    Europe              NPP  0.005069 0.272   50 [ /  9864  /   184 ] 1
    mobilisation        NC   0.004765 0.273   47 [ /  9864  /   172 ] 1
    justice             NC   0.004663 0.195   46 [ /  9864  /   236 ] 2
    parti               NC   0.004663 0.561   46 [ /  9864  /    82 ] 1
    action              NC   0.004157 0.080   41 [ /  9864  /   515 ] 3
    réforme             NC   0.004055 0.098   40 [ /  9864  /   410 ] 2
    État                NC   0.003954 0.258   39 [ /  9864  /   151 ] 1
    gouvernance         NC   0.003751 0.116   37 [ /  9864  /   319 ] 2
    crise               NC   0.003650 0.099   36 [ /  9864  /   365 ] 2
    identité            NC   0.003650 0.087   36 [ /  9864  /   412 ] 2
    économie            NC   0.003345 0.066   33 [ /  9864  /   498 ] 2
    émergence           NC   0.003244 0.075   32 [ /  9864  /   428 ] 1
    logique             NC   0.003143 0.106   31 [ /  9864  /   293 ] 2
    régulation          NC   0.003143 0.101   31 [ /  9864  /   308 ] 1
    transformation      NC   0.003143 0.091   31 [ /  9864  /   340 ] 1
    acteur              NC   0.003041 0.123   30 [ /  9864  /   244 ] 1
```

Économie et finance quantitative

```
Filter1 :     79 heads /    273 (28.94 %)
Filter2 :      4 heads /    273 (1.47 %)
    lemma              POS    F / dom F / o  occ [ /   dom  /   occ
```

```
aversion            NC   0.004090 1.000    2 [ /   489  /     2 ] 1
tarification        NC   0.004090 0.105    2 [ /   489  /    19 ] 1
complexification    NC   0.004090 0.400    2 [ /   489  /     5 ] 1
GRP                 NPP  0.004090 1.000    2 [ /   489  /     2 ] 1
```

A2.2 Liste des têtes transdisciplinaires

Le tableau suivant présente nos 123 têtes transdisciplinaires. Est indiqué le lemme, la catégorie du discours, si le lemme appartient aux formes du lexique transdisciplinaire des écrits scientifiques (Tutin, 2008) (LTES) 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 123 têtes transdisciplinaires relevées, 86 appartiennent au LTES, soit 70 %.
- Sur les 123 têtes transdisciplinaires relevées, 110 sont également relevées par Flowerdew et Forest comme étant utilisées comme signalling nouns, soit 89 %.
- Sur les 123 têtes transdisciplinaires relevées, 103 sont également des têtes spécifiques, soit 84 %.

N°	Lemme	Tout le corpus	Titres monosegmentaux	1er segment des titres bisegmentaux	2e segment des titres bisegmentaux	Présence dans le LTES	Présence dans signalling nouns
1	activité	1		1		LTES	59
2	an	1			1	LTES	
3	analyse	1	1	1	1	LTES	178
4	application	1	1	1	1	LTES	44
5	apport	1	1	1	1	LTES	16
6	approche	1	1	1	1	LTES	246
7	aspect	1	1		1	LTES	78
8	bilan	1			1	LTES	83
9	cadre	1	1		1	LTES	31
10	cas	1			1	LTES	890
11	changement	1			1	LTES	209
12	comparaison	1	1		1	LTES	44
13	compte			1			18
14	concept	1			1	LTES	143
15	condition				1	LTES	248
16	conséquence	1	1		1	LTES	132
17	construction	1	1	1	1	LTES	2
18	contexte			1	1	LTES	73
19	contribution	1	1		1	LTES	16
20	contrôle		1			LTES	3
21	culture			1			
22	défi	1			1		26
23	définition				1	LTES	68
24	démarche				1	LTES	112
25	développement	1	1	1	1	LTES	39

26	dimension	1				LTES	8
27	discours	1			1		51
28	dispositif	1		1	1	LTES	7
29	donnée				1	LTES	44
30	dynamique	1	1	1	1		
31	économie			1			
32	effet	1	1	1	1	LTES	393
33	élément	1	1		1	LTES	33
34	émergence	1	1		1		
35	enjeu	1	1	1	1		
36	enquête	1			1		16
37	enseignement	1			1		
38	espace	1	1	1	1		
39	essai	1	1		1		41
40	état	1	1	1	1	LTES	10
41	étude	1	1	1	1	LTES	18
42	évaluation	1	1	1	1	LTES	10
43	évolution	1	1	1	1	LTES	39
44	exemple	1		1	1	LTES	421
45	expérience	1	1	1	1	LTES	3
46	figure	1	1	1	1		88
47	fonction		1			LTES	150
48	formation	1	1	1	1		
49	forme	1	1	1	1	LTES	88
50	gestion	1	1	1		LTES	
51	histoire	1	1	1	1		20
52	identité			1			3
53	illustration				1		33
54	image	1	1	1	1	LTES	5
55	impact	1	1	1	1		96
56	influence	1	1	1	1	LTES	44
57	intégration	1	1	1		LTES	
58	interaction	1	1	1	1	LTES	15
59	intérêt	1	1		1	LTES	29
60	introduction	1	1	1	1	LTES	70
61	jeu	1	1	1	1		
62	leçon				1		51
63	lecture	1	1		1	LTES	33
64	limite				1	LTES	10
65	mesure	1	1	1		LTES	46
66	méthode	1	1	1	1	LTES	280
67	méthodologie	1	1		1		13
68	mode				1	LTES	11

69	modèle	1	1	1	1	LTES	474
70	modélisation	1	1	1	1	LTES	3
71	mythe				1		2
72	note	1	1	1	1		13
73	notion		1			LTES	73
74	objet	1			1	LTES	3
75	organisation	1	1	1		LTES	13
76	outil	1	1	1	1	LTES	7
77	perception	1				LTES	85
78	paradoxe				1		11
79	parcours				1		36
80	perspective	1	1		1	LTES	36
81	piste				1		2
82	place	1	1	1	1	LTES	21
83	point	1	1		1		393
84	politique	1	1	1	1		2
85	pratique	1	1	1	1		73
86	présentation	1	1	1	1	LTES	11
87	principe	1			1	LTES	251
88	problématique				1		287
89	problème	1	1		1	LTES	619
90	processus	1	1	1		LTES	230
91	production	1	1	1		LTES	2
92	projet	1	1	1	1	LTES	37
93	proposition	1	1		1	LTES	46
94	question	1	1	1	1	LTES	313
95	rapport	1			1	LTES	10
96	réalité				1	LTES	23
97	recherche	1	1	1	1	LTES	2
98	réflexion	1	1	1	1	LTES	16
99	regard	1	1		1		5
100	relation	1	1	1	1	LTES	93
101	remarque	1	1		1		21
102	représentation	1	1	1	1	LTES	11
103	réseau	1	1	1		LTES	7
104	résultat	1			1	LTES	572
105	retour	1	1	1	1		29
106	revue				1		8
107	rôle	1	1	1	1	LTES	153
108	science	1					
109	source				1	LTES	10
110	stratégie	1	1	1	1	LTES	205
111	structure	1	1	1	1	LTES	13

112	synthèse				1	LTES	2
113	système	1	1	1	1	LTES	109
114	temps		1			LTES	184
115	théorie	1	1		1		494
116	traitement	1	1			LTES	300
117	transformation		1			LTES	2
118	travail	1	1	1		LTES	24
119	usage	1	1	1	1	LTES	73
120	utilisation	1	1	1	1		5
121	valeur		1			LTES	13
122	variation	1	1			LTES	15
123	voie				1	LTES	668
	123	**94**	**81**	**63**	**99**	**86**	**110**

A3. Étiquettes utilisées par Talismane et HAL

A3.1 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

A3.2 Code des 27 disciplines de HAL retenues

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 discipline associée

A4. Éléments techniques

A4.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 la discipline des titres. Le résultat est un dictionnaire où la clé est la discipline et la valeur le nombre de titre dans cette discipline.

```
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 racines.

```
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.

A4.B 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 racines 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-racine de la forme :
  - 1__ pour un titre ayant 1 segment et 1 racine,
  - 2__ pour un titre ayant 1 segment et 2 racines,
  - 1:0 pour un titre ayant 1 racine dans son premier segment et 0 dans son second,
  - 0:1 pour l'inverse,
  - 1:1 pour un titre ayant 1 racine 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 racine,
  - 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 racine 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 racine 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 racine 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 racine 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 racine 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 racine 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 repérer 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 racine*
            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 racine.*
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 racine 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 racine.*
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 racine. 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.

A5. Index des tableaux