## 

Addendum dossier de recherche

Table des matières

[II.3 Schémas récurrents d’emploi des têtes transdisciplinaires 1](#_Toc18761177)

[II.3.1 Schémas récurrents 1](#_Toc18761178)

[II.3.2 Lexique des noms inclus dans nos schémas des têtes 3](#_Toc18761179)

[III. Discussion sur nos résultats, limites et perspectives 3](#_Toc18761180)

[Opérationnalisation des NSS 3](#_Toc18761181)

Ce document est un correctif au fichier 2019-09-**06**-Gouteux-Dossier-de-recherche.pdf. Il remplace le contenu de ce fichier à partir de la section II.3, Schémas récurrents d’emploi des têtes transdisciplinaires. Le fichier 2019-09-**07**-Gouteux-Dossier-de-recherche.pdf intègre ce correctif directement dans le fichier principal.

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

### II.3.1 Recherche de schémas d’emplois des têtes transdisciplinarité

La requête de la CS-VII, si on limite son premier nom à nos têtes transdisciplinaires, retourne 64 606 résultats, en en écartant 115 325. Nous comptons 224 108 occurrences de têtes transdisciplinaires, dans tous nos domaines ce qui fait que 29 % des têtes transdisciplinaires entrent dans une structure correspondant à notre schéma. Cela reste encore trop grand pour notre capacité d’analyse manuelle.

Du fait que les NSS soient une classe ouverte, et que les définitions varient d’un auteur à l’autre, aucune liste définitive n’est possible. Schmid en liste 670 (2000, p. 381), Flowerdew et Forest (2015), 845, et Tutin (2008, p. 3), 356. Schmid (2018, p. 118) souligne la convergence de sa liste avec celle de Flowerdew et Forest (2015) sur les NSS les plus fréquents malgré leurs différences de méthodes. Ces listes peuvent donc servir d’indice, mais en aucun cas de preuve, pour prendre en compte le potentiel d’emploi sous-spécifié de nos têtes transdisciplinaires. Sur les 94 têtes transdisciplinaires, 23 sont reconnues comme pouvant être un NSS par Legallois (2008, p. 3), soit seulement 24 %. Mais la définition opératoire que nous avons repose sur l’acceptation de certains schémas, seulement CS-I et CS-II pour Legallois, et l’utilisation d’un corpus, qui peut par sa taille et son contenu, ici les articles de l’année 1995 du quotidien *Libération* donner des résultats très différents. Cependant, sur les 94 têtes transdisciplinaires, 83 ont un lemme anglais correspondant à leur traduction qui apparaît dans la liste de Flowerdew et Forest (2015), soit 88 %. Son corpus est beaucoup plus proche de notre matériau, puisqu’il s’agit du Flowerdew Corpus of Academic English (Flowerdew et Forest, 2015, p. 68) composé de journaux académiques, de discours et de leçons. Cela nous amène à vouloir chercher les schémas récurrents des têtes transdisciplinaires dans nos titres.

#### B) Schémas récurrents d’emplois des têtes transdisciplinaires

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

La question se pose de distinguer les schémas récurrents des têtes transdisciplinaires des autres. Pour cela, nous reprenons directement une méthode formulée par Roze et al. (2014, p. 8), qui s’inspiraient de Quiniou et al. (2012) : la fouille de données séquentielles. Tout d’abord, nous construisons des séquences autour des noms. Chaque séquence est composée d’items qui sont pour les classes fermées le lemme de la forme et pour les classes ouvertes son étiquette morphosyntaxique, sauf pour les verbes *être* et *avoir* où nous gardons également leur lemme. Nous ajoutons les items INIT pour le début du titre et END pour sa fin.

Nous calculons toutes les séquences existences en utilisant une taille maximale en les répartissant en deux bases : d’un côté, les motifs dont le pivot est une tête transdisciplinaire et de l’autre ceux dont ce n’est pas le cas. Nous calculerons ensuite le taux de croissance de chaque motif spécifique aux têtes transdisciplinaires par rapport au motif correspondant dans l’autre base. S’il n’y a pas de motif correspondant, le taux de croissance est infini. Sinon il est égal au support de la séquence transdisciplinaire divisé par le support de la séquence non transdisciplinaire. Le support d’une séquence S est le nombre de séquences contenant S, c’est-à-dire les items de S dans le même ordre y compris de façon disjointe, dans une base donnée.

Les motifs émergent sont *« les motifs dont le support augmente de manière significative d’un ensemble de données à un autre »* (Roze et al., 2014, p. 8), ce qui se traduit par un taux de croissance supérieur à une valeur *p*. Nous fixons également la largeur de la fenêtre à un minimum de 2 items et un maximum de 5. Nos résultats ne permettent pas de distinguer des motifs émergents propres aux têtes transdisciplinaires.

### II.3.1 Schémas récurrents

Si le taux de croissance ne donne rien de probant, cela signifie que les têtes transdisciplinaires ne se distinguent pas syntaxiquement des têtes non transdisciplinaires. Néanmoins, nous pouvons utiliser le comptage des séquences pour faire émerger les plus fréquentes comprenant une tête transdisciplinaire sur les 1 604 847 recensées.

La figure (1) permet d'avoir un aperçu de ces séquences, une flèche indiquant la relation "est contenu dans" qui est transitive (pour trois séquences A, B et C, si A -> B -> C alors A -> C). Par souci de lisibilité, nous n'avons pas fait figurer les relations déductibles par transitivité.

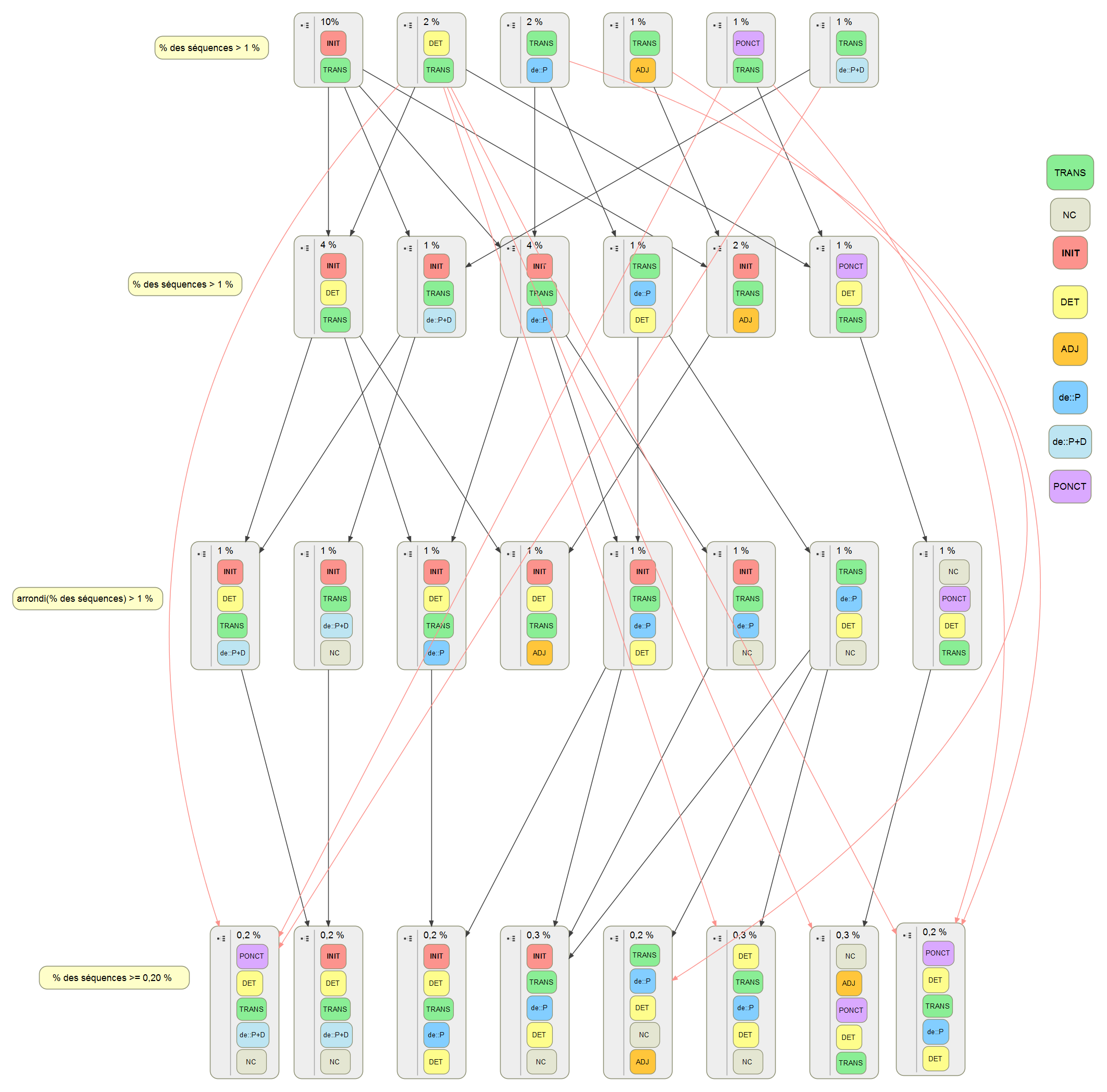


Figure 1 : Arbre des séquences les plus fréquentes des têtes transdisciplinaires

Nous avons étagé le diagramme selon le nombre d'éléments par séquence. Normalement, il n'y aucune relation sautant un niveau, néanmoins avec les filtres appliqués sur la fréquence de chaque séquence, certaines séquences pas assez fréquentes ne sont pas affichées. Nous avons donc fait figurer en rouge les relations qui sautent des niveaux de l'arbre.

La fouille de données séquentielles est une méthode qui peut être appliquée de façon neutre à n'importe quel type de données. Nos connaissances linguistiques nous permettent d'interpréter ce diagramme, notamment l'optionnalité de certains items, pour en tirer les deux schémas suivants qui sont les plus grands conteneurs de séquences :

**INIT[[1]](#footnote-1)** [DET] **TRANS** ((P [DET]) | P+D) **NC** [ADJ]

**NC** ADJ **PONCT** DET **TRANS** ((P [DET]) | P+D) **NC** [ADJ]

On remarque dans le premier la tête transdisciplinaire est exclusivement avant un nom alors que dans le second, la tête transdisciplinaire est au milieu de deux noms, l’un séparé par une ponctuation, l’autre une préposition. On remarque également la forte affinité des têtes transdisciplinaires pour ce qui pourrait être le début de segment, que cela soit le début du titre (INIT) ou une marque de ponctuation segmentatrice (PONCT), néanmoins il reste à vérifier si la marque est un segmentateur.

Les séquences de longueur cinq dont sont issus ces schémas ont un faible le taux de croissance : elles ne sont pas spécifiques aux têtes transdisciplinaires, ce qui s'explique car ce qui apparaît formellement comme une complémentation de nom est accessible à une grande quantité de nom, sinon tous, alors que la capacité de portage propositionnel (Huyghe, 2018) n'est propre qu'au NSS.

Pour déterminer s'il s'agit d'emploi sous-spécifié, il faut à présent s'interroger sur la sémantique du nom présent dans ces deux schémas. Pour cela, nous construisons un lexique de ces noms, en indiquant à chaque fois le nombre de fois où on le trouve après une tête transdisciplinaire étant directement après INIT (schéma 1), ou si on le trouve avant ou après une tête trans disciplinaire directement après une ponctuation (schéma 2). Remarquons que ce cas se rapproche de notre travail de première année, mais par un autre cheminement.

### II.3.2 Lexique des noms inclus dans nos schémas des têtes

### II.3.3 Nature de l’emploi

### II.3.4 Transdisciplinarité des schémas

Nous avons dans cette partie identifié un petit nombre de têtes transdisciplinaires, 123 en tout si on reprend tous les lemmes identifiés dans les différents sous-corpus, 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.

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

Sur les 250 998 titres, on compte 94 738 occurrences de têtes transdisciplinaires. Dans nos différents résultats, nous n’avons pas eu deux fois le même titre : il n’y a pas un titre avec deux occurrences d’une CS, si on fait l’exception de la CS-VII. On dénombre donc, par estimation, que 441 titres avec une construction spécificationnelle, soit un peu moins de 0,2 % des titres et 0,47 % des occurrences de têtes. Nous pouvons résumer les emplois sous-spécifiés des têtes transdisciplinaires dans les constructions spécificationnelles dans le tableau (11).

|  |  |
| --- | --- |
| **Schéma** | **Nombre de titres** |
| CS-I NSS + être + que | 3 |
| CS-II NSS + être + de + inf | 1 |
| C-III NSS + , + ce + être + que | 0 |
| C-IV NSS + , + ce être + de + inf | 0 |
| C-V NSS + que | 0 |
| CS-VI NSS + de + inf | 437 (estimation) |
| CS-VII NSS + de + DDAA | Impossible à déterminer |
| CS-VIII NSS + être + DDAA | 1 |
| CS-IX NSS + , + ce + être + DDDAA | 1 |
| **Total** | Sans compter la CS-VII, 443 (estimation) soit 0,2 % des titres de notre corpus et 0,47 % des occurrences de têtes. |

Tableau 11: Présence des constructions spécificationnelles classiques dans notre corpus

Nous n’avons trouvé que très peu de constructions spécificationnelles classiques dans notre corpus, nous avons décidé d’utiliser la fouille de données séquentielles pour mettre à jour des schémas d’utilisation récurrents des têtes transdisciplinaires. Nous en avons identifié plusieurs qui fonctionnent comme des constructions spécificationnelles.

# III. Discussion sur nos résultats, limites et perspectives

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

## III.1 Éléments de discussion

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

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

### Limitations des têtes spécifiques aux domaines

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

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

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

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

### Têtes transdisciplinaires

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

### Listes de NSS

Une grande difficulté a été de mettre la main sur des listes numériques des différentes acceptations des NSS. Elles peuvent servir seulement d’indices, mais précieux, car les NSS sont un emploi et non une classe lexicale a priori, bien qu’il existe des propriétés lexicales a priori de capacité pouvoir être employé comme NSS. Certains articles pointaient sur un site web n’était plus en ligne, d’autres ne prenaient même pas cette peine. Pour la linguistique de corpus, la mise à disposition pérenne des listes produites par la computation est parfois aussi importante que l’article. La capacité de stocker un article avec des pièces-jointes, parfois volumineuses, nous semble importante, notamment pour les archives ouvertes.

### Opérationnalisation des NSS

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

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

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

# 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 tête » près de 98 % des 56 851 titres auxquels il manquait une tête. Pour finir, nous avons constitué un corpus de travail de 250 998 titres, gardant près de 74 % du matériau initial.

Après avoir délimité notre périmètre et donc notre corpus de travail et identifié toutes les têtes, nous nous sommes d’abord interrogés sur le nombre de segments par titre en fonction du domaine. Il apparaît que les sciences humaines utilisent dans les mêmes proportions titres monosegmentaux et titres bisegmentaux tandis que les sciences exactes privilégient les titres monosegmentaux. Nous nous sommes interrogés sur 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 %. Nous pouvons donc conclure que les titres sont essentiellement des syntagmes nominaux.

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

Nous avons également recherché les têtes transdisciplinaires, fréquentes dans tous les domaines. Nous avons trouvé 94 têtes transdisciplinaires dans tout notre corpus de travail. Nous avons remarqué que sur les 123 têtes transdisciplinaires, 86 % appartiennent au lexique transdisciplinaire des écrits scientifiques relevé par Tutin (2008).

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

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

Faute de construction spécificationnelle classique, nous avons donc étudié les schémas récurrents dans lesquels s’insèrent nos têtes transdisciplinaires. Nous avons pu établir que ceux-ci sont très ramassés et averbaux ce qui est en accord avec les spécificités des titres. Nous avons pu montrer que ces schémas récurrents jouent le même rôle que les constructions spécificationnelles classiques sur plusieurs exemples.

Le problème est que le jugement à la fin repose moins sur des critères syntaxiques que des critères lexicaux et sémantiques. Pour le lexical, à défaut de liste établie des DDAA, la tâche est difficile, et pour la sémantique, nous n’avons pu faire autrement qu’en faisant appel au jugement de l’intuition ce qui est insuffisant dans une perspective d’automatisation. En nous fondant sur plusieurs facteurs de rapprochement et notre jugement, nous avons établi une liste des têtes transdisciplinaires fréquemment utilisées dans un emploi NSS.

1. Le schéma de couleur reprend celui de la figure (1). [↑](#footnote-ref-1)