Addendum dossier de recherche

Table des matières

Ce document remplace en la complétant la partie II.3 et la conclusion du fichier 2019-09-06-Gouteux-Dossier-de-recherche.pdf.

II.3 Schémas récurrents d'emploi des têtes transdisciplinaires

II.3.1 Recherche de schémas d'emplois des têtes transdisciplinarité

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

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

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

A) Impossibilité de trouver des schémas émergents

L'existence de nos têtes transdisciplinaires, fréquentes, abstraites, dotées d'un faible contenu sémantique, le fait que 83 % d'entre elles apparaissent dans la liste des signalling nouns, nous pousse à nous demander s'il n'existerait pas d'autres constructions spécificationnelles, propres aux titres. Nous allons à présent essayer de rechercher des schémas récurrents dans lesquels s'insèreraient 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 de mots autour des noms. Chaque séquence est composée d'items qui sont, pour les classes fermées, le lemme du mot, et, pour les classes ouvertes, son étiquette morphosyntaxique, sauf pour les verbes *être* et *avoir* où nous gardons également le lemme. Nous ajoutons les items INIT pour le début du titre et END pour sa fin.

Nous calculons toutes les séquences existantes en utilisant une taille minimale de deux éléments et une taille maximale de cinq éléments. Nous les répartissons 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 dans une base donnée est le nombre de séquences contenant S, c'est-à-dire qu'elles contiennent tous les items de S dans le même ordre, y compris de façon disjointe.

Les motifs émergeants sont *« les motifs dont le support augmente de manière significative d'un ensemble de données à un autre »* (Roze et al., 2014, p. 8), ce qui se traduit par un taux de croissance supérieur à une valeur *p* que nous fixons à un. Nos résultats ne permettent pas de distinguer des motifs émergents propres aux têtes transdisciplinaires par rapport aux autres têtes.

B) Deux schémas fréquents proches de la CS-VII

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 motifs les plus fréquents des têtes transdisciplinaires sur les 1 604 847 séquences 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. Nous avons étagé le diagramme selon le nombre d'éléments par séquence et nous avons filtré pour ne garder que les séquences les plus fréquentes à chaque niveau.

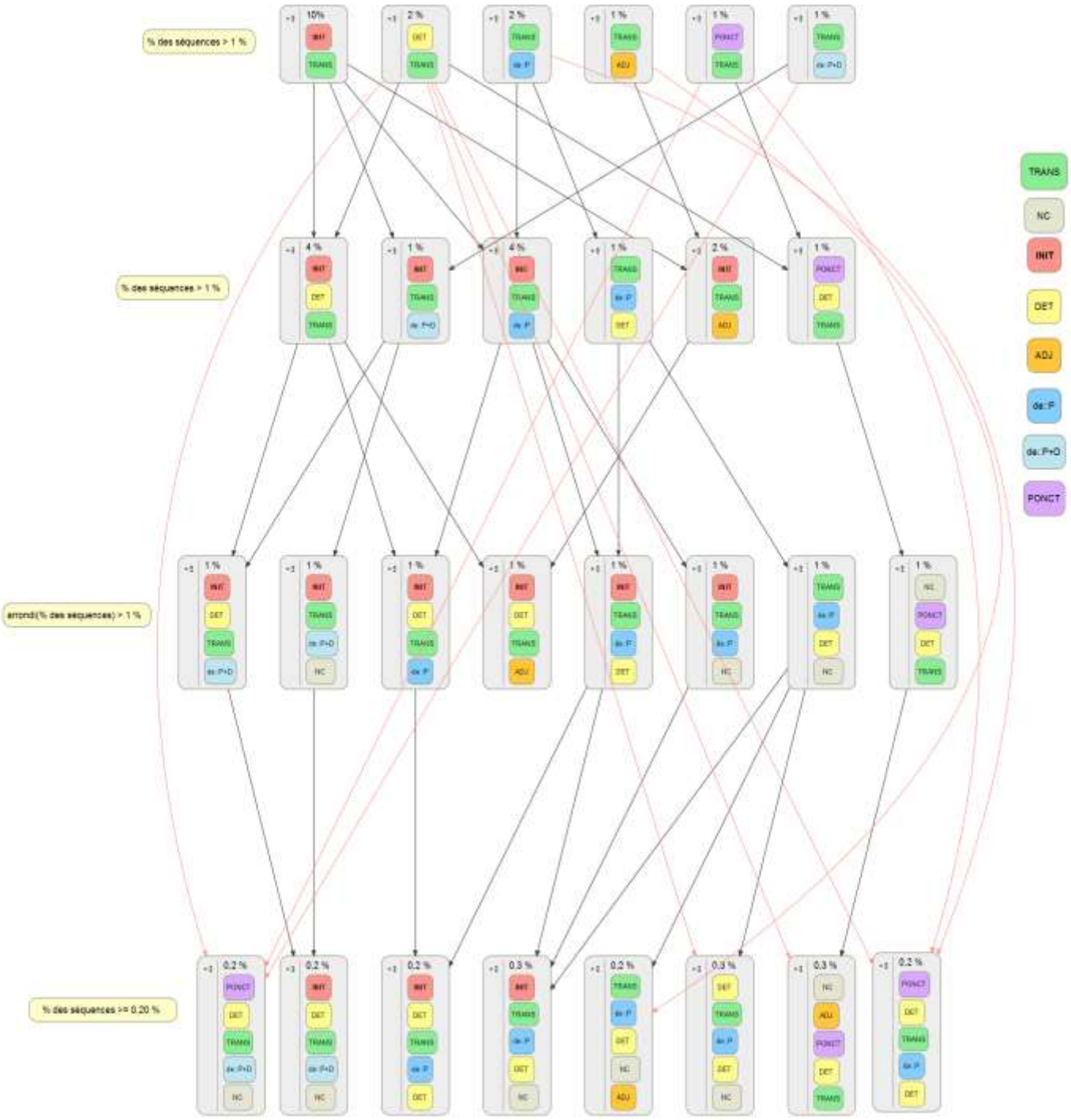

Figure 1 : Arbre des séquences les plus fréquentes des têtes transdisciplinaires

Normalement, il n'y aucune relation sautant un niveau, néanmoins avec les filtres appliqués sur les fréquences, certaines séquences existantes mais 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 :

Schéma 1 : INIT[1] [DET] TRANS ((P [DET]) | P+D) NC [ADJ]

Schéma 2 : PONCT DET TRANS ((P [DET]) | P+D) NC [ADJ]

Le fait le plus remarquable est que les deux schémas correspondent à la CS-VII, NSS de NC, en élargissant à toutes les prépositions d'un côté mais en contraignant sur l'emplacement dans le titre. Dans le premier la tête transdisciplinaire est le premier nom du titre, dans le second elle est le premier nom après une marque de ponctuation, potentiellement une marque de segmentation, ce qui ferait de la tête transdisciplinaire le premier nom d'un second segment. Remarquons que ce dernier cas se rapproche de notre travail de première année, mais par un autre cheminement.

Les motifs fréquents dont sont issus ces schémas ont un faible le taux de croissance par rapport aux motifs des têtes non transdisciplinaires : ils ne sont pas spécifiques aux têtes transdisciplinaires. Ce fait s'explique car ce qui apparaît formellement comme une complémentation de nom est accessible à une grande quantité de noms, sinon tous, alors que la capacité de portage propositionnel (Huyghe, 2018) n'est propre qu'aux NSS.

Pour déterminer s'il s'agit d'emplois sous-spécifiés, il faut à présent s'interroger sur la sémantique du nom commun présent dans ces deux schémas.

II.3.2 Lexique des noms et détermination de l'emploi

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 après une tête transdisciplinaire directement après une ponctuation (schéma 2). La détermination de l'emploi repose donc sur l'analyse du nom commun qui suit la tête transdisciplinaire. Nous avons vu que ce nom doit directement dénoter une action ou une activité ou amener une action ou une activité de façon implicite.

A) L'étude de problème

La recherche du schéma 1 sur notre corpus donne 33 849 couples (TRANS, NC) différents. On remarque que toutes les têtes transdisciplinaires apparaissent dans nos résultats. Pour analyser s'il pourrait s'agir d'un emploi NSS nous décidons de nous concentrer sur une tête transdisciplinaire *problème*. Dans nos résultats, il est la 55^{e} dans l'ordre de fréquence des têtes transdisciplinaires. Selon notre méthode de sélection, il est la 58^{e} tête transdisciplinaire, sur 94, en les classant par valeur de médiane. Il est également un *prime shell noun* de Schmid (2000, p. 85 ; 2018 ; p. 118). Dans la liste de Flowerdew et Forest (2015, p. 203), il est le 3^{e} par ordre de fréquence normalisée. Il y a donc une forte probabilité que *problème* puisse avoir des emplois en NSS dans notre corpus et plus particulièrement dans le schéma 1.

On coerce donc notre schéma 1 en prenant pour TRANS que les occurrences de *problème*. Mais, en prévision d'un faible nombre de résultats, nous décidons de recherche notre schéma comme une séquence : on autorise une correspondance disjointe des énoncés avec le schéma, cela permet de maximiser la correspondance avec des variantes que l'on n'aurait pas prévues dans la définition initiale : par exemple, avoir un adjectif pour la tête transdisciplinaire, ce qui est une des séquences qui avait été trouvées lors de notre fouille de données séquentielles. De plus, on lève dans ce cas la contrainte que TRANS soit une tête pour estimer s'il existe des énoncés qui correspondent à

[1] Le schéma de couleur reprend celui de la figure (1).

cette séquence où le lemme *problème* ne serait pas une tête de segment. TRANS peut donc correspondre à deux cas :

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

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

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

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

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

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

La paraphrase de ces exemples en CS-II, avec un infinitif, est immédiate : *Quelques problèmes d'analyser, le problème de regrouper, Problèmes de créer en multimédia*. *Analyse, regroupement, création* sont des déverbaux dénotant une action ou une activité.

Sur les 370 résultats, deux occurrences correspondent au schéma sans que *problème* ne soit une tête de segment : les exemples (37) et (38). Une seule de ces deux occurrences peut être reçue comme un emploi de *problème* en NSS, l'exemple (37).

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

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

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

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

Les 135 occurrences retenues utilisent toutes la préposition *de* même s'il existe également des occurrences non gardées qui utilisent une construction de la forme *le problème posé par X* comme dans l'exemple (39) et qui semblent correspondre à un emploi NSS dans 17 cas. L'exemple (39) peut en effet être paraphrasé en *Problèmes de prédire en persan*. Néanmoins, cet emploi est

propre à *problème*, on ne peut pas mettre un autre lemme à la place, nous ne le retenons dans notre tentative de trouver une détermination générale de l'emploi NSS.

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

Le résultat le plus marquant est la très forte présence de DDAA dans résultats dont le trait saillant est l'utilisation d'un suffixe. On catégorise leurs formations morphologiques dans le tableau (1).

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

Tableau 1: Suffixes associés au nom noyau du contenu spécifiant

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

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

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

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

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

Étudier *problème* nous amené à constater que :

- L'utilisation en tant que NSS semble liée au fait que le mot soit tête de segment, reste à déterminer l'éventuelle corrélation entre les deux faits.

- L'utilisation en tant que NSS peut être estimée en analysant la morphologie du nom après la préposition *de.*
- Il pourrait exister des constructions spécificationnelles propre à chaque nom, comme *le problème posé par*, comme Nakamura (2017, p. 7) signale *avoir pour/comme objectif de*.

B) Estimation globale à l'aide du schéma 1

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

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

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

B) Estimation globale à l'aide du schéma 2

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

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

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

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

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

Les deux utilisent des marques de ponctuation non reconnues comme segmentatrices par notre définition, la virgule pour (39) et le tiret pour (40). Nous pensons qu'il s'agit pourtant bien ici de titres bisegmentaux ayant pour tête de second segment *outil* et *étude*, qui sont tous les deux des

lemmes de têtes transdisciplinaires. La forte proportion d'emploi NSS de lemmes non-têtes provient donc d'une mauvaise segmentation initiale des titres.

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

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

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

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

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

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

On peut calculer la corrélation entre deux associations :

- Têtes transdisciplinaires et estimation d'emploi NSS : 0,35
- Têtes non transdisciplinaires et estimation d'emploi NSS : 0,18

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

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

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

- Tête et estimation d'emploi NSS : 0,30

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

E) Tests de corrélation entre emploi NSS et problème

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

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

Nous calculons la corrélation entre :

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

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

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

que d'ordre lexicale. Ce qui est renforcé par le fait que les NSS sont une classe fonctionnelle, un emploi potentiel d'un lemme, et non une classe lexicale, même si des propriétés sémantiques en facilitent l'utilisation en tant que NSS.

II.3.4 Transdisciplinarité des schémas

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

Domaine	Tête transdisciplinaire en emploi de NSS selon schéma 1	Tête transdisciplinaire en emploi de NSS selon schéma 2	Total tête transdisciplinaire en emploi de NSS
Anthropologie	121 22 % 17	73 9 % 13	194 11 % 20
Archéologie et Préhistoire	212 21 % 18	134 9 % 10	346 11 % 21
Architecture	125 26 % 11	56 9 % 12	181 14 % 15
Art et histoire de l'art	73 18 % 23	59 8 % 18	132 9 % 24
Chimie	116 20 % 20	20 7 % 23	136 14 % 14
Droit	521 19 % 21	124 7 % 21	645 13 % 18
Éducation	495 29 % 7	195 12 % 5	690 17 % 7
Gestion et management	1118 31 % 3	507 10 % 9	1625 16 % 10
Géographie	47 31 % 2	24 12 % 6	71 16 % 9
Histoire	253 19 % 22	199 8 % 19	452 9 % 23
Informatique	919 26 % 10	177 11 % 7	1096 19 % 2
Linguistique	484 26 % 9	272 12 % 4	756 14 % 13
Littératures	97 15 % 25	65 6 % 24	162 8 % 25
Mathématiques	128 20 % 19	20 9 % 14	148 15 % 12

Philosophie	124	18 %	24	53	7 %	22	177	10 %	22
Physique	2606	26 %	12	193	8 %	20	2799	21 %	1
Planète et Univers	165	24 %	15	42	9 %	15	207	15 %	11
Psychologie	139	30 %	5	51	12 %	3	190	18 %	4
Science politique	240	25 %	13	117	9 %	10	357	13 %	17
Sciences cognitives	158	32 %	1	67	12 %	2	225	18 %	3
Sciences de l'environnement	410	30 %	6	118	10 %	8	528	18 %	6
Sciences de l'information et de la communication	276	31 %	4	152	13 %	1	428	16 %	8
Sciences du Vivant	1414	25 %	14	228	8 %	16	1642	18 %	5
Sociologie	957	26 %	8	420	8 %	17	1377	13 %	16
Économie et finance quantitative	21	24 %	16	5	5 %	25	26	13 %	19

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

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

Nous avons ensuite rappelé le concept de NSS, un nom fréquent au faible contenu sémantique dont la particularité est d'être spécifié par son contexte à l'aide de plusieurs

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

Nous avons essayé de détecter les différentes occurrences de constructions spécificationnelles dans nos titres où une tête transdisciplinaire serait employée comme NSS. Nous nous sommes heurtés au problème que les définitions les plus contraignantes retournaient très peu de résultats, du fait qu'elles utilisent des verbes alors que les titres sont essentiellement averbaux. Nous pouvons résumer les emplois sous-spécifiés des têtes transdisciplinaires dans les constructions spécificationnelles dans le tableau (11).

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

Tableau 2: Présence des constructions spécificationnelles classiques dans notre corpus

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

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

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 et des lexiques scientifiques

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

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

Opérationnalisation des NSS

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

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

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

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. Les deux schémas détectés se rapprochent d'une construction spécificationnelle, la CS-VII, qui est la moins contraignante des CS étudiées. Ils ajoutent comme contrainte que le schéma doit se trouver en début de titre ou de segment. Nous avons ensuite voulu étudier si, outre la correspondance syntaxique entre les deux schémas récurrents détectés et la CS-VII, il y avait effectivement un emploi sous-spécifié des têtes transdisciplinaires.

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

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