Dossier de recherche

Damien Gouteux

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

Département des Sciences du Langage

M2 Linguistique, Informatique et Technologies du Langage

Université Toulouse II Jean-Jaurès

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 lecteur en restera là (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). Sa production augmente constamment en doublant tous les 12 ans (Stix, cité dans Salager-Meyer et al. 2013). Ce tri par la lecture du titre soulève la question de l'information qu'il contient. On peut s'interroger sur 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 français ont été moins étudiés, on peut citer les travaux de Ho-Dac et al. (2001), Rebeyrolle et al. (2009) et Tanguy et Rebeyrolle (à paraître). Nous prenons en compte 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 son sens (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 (";"). 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 obéit à la séquence décrite. Les éléments décrits par le schéma peuvent être alors individuellement associés à un mot ou un signe de ponctuation de la séquence textuelle correspondante, on dit que les mots et signes peuplent le schéma. Ainsi la séquence de mots "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 de mots "union parfaite ou mariage impossible" correspond au premier schéma mais pas au second. 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 (: NC P NC),
- un double point suivi d'un syntagme prépositionnel dont le nom est complémenté par un syntagme prépositionnel (: P NC P NC),
- un double point suivi d'un syntagme nominal constitué de deux noms coordonnés (: 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 issus du vocabulaire du domaine scientifique, des noms abstraits, dont les 11 plus fréquents étaient :

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

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 maintenir la cohérence du texte. Lexicalement, ces noms appartiennent aux listes de noms généraux fréquemment employés dans un emploi sous-spécifié (NSS) telles qu'elles ont été définies par Schmid (2000) et Flowerdew et Forest (2015). Un NSS 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 NSS est un emploi et non une nature

lexicale, même si certains ont une appétence pour cet emploi (Schmid, 2000). Un NSS 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 d'un rapprochement 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 deux CS les plus fréquemment étudiées (Legallois, 2008 ; Schmid, 2000) sont :

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

Cette définition opérationnelle s'accompagne d'une définition fonctionnelle en trois points présentés par Schmid (2000) :

- Fonction sémantique : mise en perspective de tranches d'information variables.
- Fonction cognitive : formation de concepts temporaires dans des concepts nominaux uniques.
- Fonction textuelle : structuration du texte à travers la représentation de segmentation de nature variable.

La dernière fonction, qui rapproche les NSS du fonctionnement des pronoms, est difficilement perceptible dans un énoncé aussi court qu'un titre. Néanmoins les deux premières peuvent s'appliquer : que l'on qualifie de *problème* ou de *question* un élément du titre est une mise en perspective que le locuteur impose à son lecteur. De même, est regroupé sous ce NSS tout ce qui vient après dans le titre, formant un concept temporaire. Néanmoins, on ne retrouve pas les CS définies plus haut dans les titres.

La classe de nom ayant émergé de notre premier travail se rapprocherait-elle de ces NSS ? Pour cela, nous voulons étudier un ensemble de caractéristiques qui permettraient de les rapprocher des NSS. Et en l'absence de constructions spécificationnelles dans les titres, nos noms s'intègrent-ils néanmoins dans des énoncés correspondant à des schémas d'utilisation très fréquents ? 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).

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 avant ou après le double point. Nous avions écarté 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 les noms que nous étudions immédiatement après le double point sont les noyaux (ou têtes) du syntagme après le double point. Nous pensons que notre nouvelle étude, au lieu de prendre le premier nom après le double point, peut élargir son périmètre en portant sur l'étude de toutes les têtes, ou noyaux, de syntagme de premier niveau d'un segment de titres.

Tout d'abord, pour ce nouveau travail, nous découpons nos titres en segments en reprenant et en amendant une liste de signes de ponctuation qui segmentent les titres en anglais établie par Anthony (2001). Ensuite, pour trouver les têtes de syntagmes, plutôt que de simplement parcourir le segment et prendre le premier nom rencontré comme nous le faisions en première année, nous avons décidé d'utiliser l'analyse syntaxique en dépendances (Tesnière, dans Schwischay, 2001) pour produire un arbre dont la racine sera 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 racines dont nous voulons étudier le rapprochement possible avec les noms sous-spécifiés. Nous voulons caractériser ces racines et les schémas récurrents dans lesquels elles s'insèrent, dans un corpus de titres de publications scientifiques.

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), Tanguy et Rebeyrolle, à paraître, pour le français). Nous ne manquerons pas de déterminer dans le cadre de notre problématique s'il existe des variations des racines et des schémas suivant les disciplines.

Notre étude se déroulera en quatre temps. Dans un premier temps, nous décrivons notre corpus de travail à l'aide de différentes mesures, en faisant référence aux nombreux travaux existants. Dans un deuxième temps, nous construisons la liste des têtes de segments et nous rappelons les apports de la littérature sur les noms sous-spécifiés. Partant des constructions spécificationnelles dans lesquelles ils s'inscrivent généralement, nous produisons une liste d'indices pour caractériser nos têtes de segments. Nous utilisons ces caractéristiques pour essayer de montrer en quoi ces têtes se rapprochent des emplois en noms sous-spécifiés en nous appuyant notamment sur leur forte fréquence et leur transdisciplinarité. Dans un troisième temps, nous discutons des schémas récurrents et fréquents dans lesquels s'insèrent les têtes de segments, qu'ils soient intrasegmentaux ou intersegmentaux. Nous voulons caractériser de façon détaillée ces schémas au niveau syntaxique et sémantique, que cela soit pour le contenu ou le fonctionnement discursif. Nous voulons également étudier ces schémas à la lumière des constructions spécificationnelles bien identifiées par nos prédécesseurs. Enfin, dans un quatrième temps, nous discutons nos résultats 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 des auteurs. Une archive ouverte présente l'avantage de centraliser l'accès aux travaux scientifiques, d'aider à leur diffusion et de les conserver 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) (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). 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.

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** auxquels est associée la publication dont nous déduisons un domaine principal selon la méthode de Tanguy et Rebeyrolle (à paraître).

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 ceux-ci et à ceux 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 de nombreuses disciplines scientifiques, tout comme « Science non linéaire » et « Économie et finance quantitative ». 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 fait 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 des fautes.

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 verbe des mots n'en étant pas. Pour mesurer la fiabilité de Talismane nous avons choisi un échantillon de 100 titres que nous avons vérifiés manuellement. Nous avons construit le tableau suivant qui résume notre étude :

Non reconnaissance	Erreur de catégorie	Erreur de régisseur	
TODO	TODO	TODO	

Une erreur de type non reconnaissances d'une forme entraîne forcément une erreur de catégorie, nous n'avons pas compté doublement ces erreurs. Nous obtenons un taux d'erreur de TODO pour 100 titres. Ce qui nous permet de confirmer que Talismane arrive à étiqueter morphosyntaxiquement et à analyser syntaxiquement en dépendances correctement des énoncés aussi courts que des titres.

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. Nous y ajoutons le point d'exclamation et les points de suspension. Nous avons donc les signes segmentant suivant :

Ponctuation forte	. ? ! …
Ponctuation faible	; :

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 effectuées 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. Pour une phrase constituée d'un segment unique, Talismane va produire dans TODO % des cas une racine également unique, nous aurons donc une phrase, un segment, une racine. Mais pour une phrase constituée de deux segments, nous écartons volontairement les cas avec plus de deux segments, Talismane peut produire une racine par segment ou une seule racine. Dans ce dernier cas, dans le segment ne contenant pas cette racine, que nous qualifions de "primaire", on constate l'existence d'un mot qui 1) est uniquement régisseur dans son propre segment 2) qui n'est régi que par un seul élément appartenant au segment contenant la racine primaire. Nous qualifions ce mot de racine "secondaire". Dans les deux cas, nous les regroupons sous l'appellation de têtes de segments.

TODO : schéma

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 un ensemble de 339 687 titres ayant les caractéristiques suivantes :

- identifiant
- année
- type de support
- 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 (mots uniquement)
- segments
    - Permet d'accéder aux différents segments du titre et notamment :
        - tête du syntagme
        - sa position dans le titre
- nombre de segments

Après avoir décrit nos données nous établissons le périmètre qui délimitera notre corpus de travail.

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

TODO

Dans un premier temps, on effectue le choix du périmètre de sélection des données. Ce périmètre est défini par rapport à la structure des titres. Nous prendrons les titres composés de 1 ou 2 segments, et nous le justifierons par le fait qu'il s'agit de la plus grande majorité des titres, qu'ils sont

plus facile à analyser, et moins sujet à une erreur d'analyse. On fera écho aux travaux didactiques sur l'écriture des titres (Aleixandre-Benavent et al., 2014 ; Swales et Feak, 1994 ; Gustavii, 2008) qui conseillent un format de titre à deux segments de la forme "*A : B*".

Un autre délimiteur, en plus du nombre de segments dans le titre, et le nombre de têtes de segments. Nous nous limiterons aux titres avec une tête, racine primaire ou secondaire, par segment, qui est le cas nominal. Avoir plus de racines peut relever d'une erreur d'analyse de Talismane. Il peut exister des segments sans racine primaire dans le cas d'une segmentation faible, mais ils auront alors une racine secondaire. Nous regroupons les deux types de racines sous l'appelation de têtes de segments.

Nous abordons dans un second temps la question de la nature des titres car on opéra également une sélection sur ce critère. 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 des deux modèles, déterminer le type de syntagme de chaque segment en étudiant la catégorie morphosyntaxique de sa tête.

Nous constatons que la grande majorité des têtes sont des noms : 86 % des têtes titres avec un segment sont des NC (80 %) ou des NPP (6 %). La grande majorité des segments sont donc des syntagmes nominaux et que les titres sont majoritairement constitués d'un ou plusieurs syntagmes nominaux, rejoignant ainsi les conclusions de nos prédécesseurs (Leech, 2000 ; Haggan, 2004 ; Soler, 2007 ; Cheng et al., 2012 ; Wang et Bai, 2007). 93 % des titres pour le corpus de Cheng et al. (2012) sont des syntagmes nominaux, 99 % pour celui de Wang et Bai (2007), et non pas une phrase avec un verbe conjugué comme noyau.

Nous pouvons ensuite étudier comment cette structure des titres varie en fonction de la discipline, Soler (2007) indiquant une préférence marquée de la biologie pour les phrases complètes. Néanmoins, dans notre corpus, tous les domaines ont avant tout des groupes nominaux comme titres (de 80 à 91 %).

On évoquera brièvement le nombre de titres écartés et le nombre de titres gardés en en montrant des exemples :

339 687	titres en tout
171 890	titres constitué d'un segment avec une seule racine.
53 621	titres constitués de deux segments avec une racine dans le premier segment.
30 554	titres constitués de deux segments avec une racine dans chacun.
3 230	titres constitués de deux avec une racine dans le second segment.
259 295	titres considérés dans mon corpus de travail (76 % du total)

On ne manquera pas de signaler si le corpus de travail est représentatif sur d'autres caractéristiques (années, domaines, nombre d'auteurs) de l'ensemble des données.

Une fois le périmètre définie sur la structure et la nature grammaticale des racines, nous avons constitué notre corpus de travail. Une fois établi notre corpus de travail, nous effectuons quelques mesures dessus, mises en rapport avec les mêmes mesures effectuées dans des travaux précédents avant de s'étudier plus avant les têtes de syntagmes.

I.2.3 Mesures effectuées sur notre corpus de travail

TODO

- nombre de titres
- nombre de titres par nombre de phrases, nombre de segment, organisation des racines
- nombre de titres par support
- nombre de titres par années
- nombre de titres par publication
- nombre de titres part domaine
- longueur moyenne des titres en mots

On évoquera les travaux de typologie syntaxique des titres comme celui de Jamali et Nikzad (2011), la présence d'un signe de ponctuation particulier que ce soit le double point (Dillon, 1981, 1982 ; Townsend, 1983 ; Diers et Downs, 1994 ; Lewison et Hartley, 2005) ou le point d'interrogation (Ball, 2009) et ceux sur la longueur du titre (Hartley, 2007 ; Jamali et Nikzad, 2011, Paiva et al., 2012). On évoquera les travaux montrants que les caractéristiques des titres ne sont pas indépendantes : la longueur du titre en nombre de mots et le nombre d'auteurs (Yitzhaki, 1994) ou la longueur du titre et la longueur de l'article (Yitzhaki, 2002). De plus, notre corpus n'est pas équilibré au niveau des domaines : il faut donc toujours procéder par fréquences relatives pour comparer une caractéristique des titres d'une discipline à l'autre.

On évoquera également un point que l'on ne traite pas : la recherche de "performance" en termes de citations et de téléchargements des titres par rapport à un point particulier (présence d'un signe de ponctuation) car les résultats sont non significatifs ou contradictoires (Merrill et Knipps, 2014).

Nous avons dans ce sous-chapitre établi le périmètre délimitant notre corpus de travail et mesuré ses contours. Nous avons décidé d'étudier le cas le plus nombreux : celui des segments nominaux. Cela nous amène à vouloir étudier les têtes de ces segments pour éventuellement les rapprocher des NSS. Dans un premier temps nous établissons la liste des lemmes des têtes et abordons les NSS. Partant de l'absence de constructions spécificationnelles dans les titres, nous allons essayer de caractériser le plus possible les têtes pour, dans un dernier temps, étudier ce qui les rapproche et les différencie des noms sous-spécifiés.

II. Caractérisation des têtes de segments

II.1 Têtes de segments et NSS

II.1.1 Lexique des têtes de segments les plus fréquentes

TODO

Établir un lexique des têtes les plus fréquentes sur les 15 692 lemmes recensés, proposer quelques remarques générales dessus notamment l'appartenance au vocabulaire académique transdisciplinaire (Hatier, 2016 ; Hatier et al., 2016) et le fait que cela soit des noms généraux définis par

Halliday et Hasan (1976) et étudiés notamment par Adler et Moline (2018). La question qui est en suspens est : sont-ils des NSS ?

On rappelle que les 10 premières : étude, analyse, modélisation, influence, effet, approche, méthode, modèle, évolution, évaluation.

II.1.2 Les noms sous-spécifiés

TODO

D'un autre côté, on résumera rapidement la littérature sur les noms sous-spécifiés. On rappellera la définition de Flowerdew (2006) : *« noms abstraits dont le sens complet peut seulement être spécifié en référence à son contexte »*. 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 un 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).

On rappelle également que les titres, pris isolément, sont des microdiscours où la capacité de référence quasi-pronominale de l'emploi sous-spécifié de noms n'est pas visible, l'espace étant trop court pour une reprise en anaphore ou cataphore comme dans l'exemple : « *Le problème des échanges commerciaux Chine-États-Unis. Un problème vital »*. Or, cette capacité de référence est une des trois fonctions clés de l'emploi sous-spécifié de noms selon Schmid (2000), avec la création de concepts temporaires et la catégorisation de ces concepts par le locuteur, qui elles sont bien présentes. Qualifier les échanges commerciaux dans l'exemple précédent de "problème" ou de "question" est une catégorisation qu'impose le locuteur sur la perception du contenu spécifiant. Un concept temporaire est bien formé mentalement dans la tête du lecteur, celui des échanges commerciaux entre la Chine et les États-Unis.

Enfin, on aborde le fait qu'il s'insère dans une construction spécificationnelle (CS).

II.2 Présence des constructions spécificationnelles dans notre corpus

Nous recensons ici les différentes *constructions spécificationnelles* (Legallois, 2008) traditionnelles de la littérature sur les NSS, qui sont autant de définitions opératoires des NSS (Schmid 2000), en essayant de les chercher dans notre corpus. Le lien avec les *grammaires de construction* sera également évoqué (François et Legallois, 2006). Nous commençons par les deux plus étudiées (Legallois, 2008 ; Schmid, 2000) :

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

Nakamura (2017) ajoute également les trois constructions spécificationnelles suivantes :

- **NSS** + [verbe être] + *syntagme nominal* : “Notre **objectif** majeur est *la rédaction d’une proposition de loi*.”
- Nom + [verbe avoir] + pour + **NSS** + **de** + *proposition subordonnée infinitive* : “Cet homme avait pour **ambition** *de devenir président*”.
- **NSS** + 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”.

TODO

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.

En recherchant dans notre corpus les occurrences de ces constructions spécificationnelles, on montre qu’on ne les trouve pas (TODO : calculer leur faible présence). C’est à partir de la dernière de Nakamura (2017), soutenu par Schmid (2018), que nous avons argument fort pour rapprocher nos noms des NSS : les titres sont des syntagmes nominaux complexes, comme l’est la cinquième construction spécificationnelle. Nous pourrions avoir un emploi de NSS sans verbe à l’infinitif dans les titres. Nous allons donc essayer d’établir des caractéristiques de têtes de segments pour soutenir cet éventuel rapprochement.

II.3 Caractéristiques retenues de rapprochement

TODO

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

- 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 Scmid (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	NSS fréquent ?	Pos.
cas	TODO					
problème						
objectif						

II.4 Points communs et différences avec les noms sous-spécifiés

II.4.1 Facteurs de rapprochement

TODO

Établir une liste de facteurs de rapprochement avec les noms sous-spécifiés et les justifier.

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

II.4.2 Facteurs de différenciation

TODO

Établir une liste de facteurs de différenciation avec les noms sous-spécifiés et les justifier.

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

II.4.3 Règle de rapprochement

TODO

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

III. Schémas récurrents et constructions spécificationnelles

III.1 Émergences de schémas fréquents récurrents

Nous étudions dans cette partie les schémas, dans lesquels s'insèrent nos têtes de segments, à l'intérieur d'un seul segment. Le but est d'essayer d'étudier plus en profondeur, au-delà de savoir qu'il s'agit d'un syntagme nominal complexe, l'emploi des têtes de segments et l'émergence de schémas fréquents récurrents. Nous commençons par étudier les schémas intrasegmentaires avant d'aborder les schémas s'étendant sur deux segments.

III.1.1 Schémas intrasegmentaires

TODO

Par exemple, nous voulons étudier ce qu'implique la différence entre les déterminants "le" et "un" comme dans "le cas de" et "un cas de", notamment sur la position du segment incluant ce schéma dans le titre et la relation sémantique qui s'établit avec le nom qui vient après. Y'a-t-il des préférences marquées pour la forme définie ou la forme indéfinie pour des noms donnés ?

Y'a-t-il des NSS substituables ? Si on prend "un cas de" et un "un problème de", peut-on faire des rapprochements entre les deux schémas suivant leur position dans le titre ou le nom qui suit ?

Quelle sémantique peut-on associer à ces schémas ?

III.1.2 Schémas sur deux segments

TODO

Nous étudions dans cette partie les schémas incluant une ou deux racines s'étendant sur plus d'un segment. On s'attardera notamment sur le cas de la complémentation par un segment de la forme "NSS : contenu spécifiant".

Nous essayerons là aussi 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 Transdisciplinarité des schémas et schémas non centrés sur têtes

III.2.1 Transdisciplinarité des schémas

TODO

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

III.2.2 Schémas non centrés sur une tête

TODO

Dans cette partie nous étudions si nous retrouvons nos schémas centrés non pas sur une tête de segment que nous aurions détectée mais sur d'autres noms. Cela pour étudier si les schémas sont propres aux têtes de segment ou non, éventuellement aux têtes de segment NSS.

III.3 Des schémas de constructions spécificationnelles ?

TODO

De la même manière que nous avons essayé de rapprocher têtes de segments et NSS, nous voulons voir si nous pouvons rapprocher les schémas récurrents dans lesquels les têtes s'insèrent et les constructions spécificationnelles. Nous pensons à un rapprochement plus lâche que celui entre têtes et NSS qui pourraient néanmoins relever des similitudes dans le fonctionnement sémantique dynamique. Nous pensons notamment aux travaux de Nakamura (2017) sur ce point.

Après avoir étudié les schémas récurrents où se trouvent les têtes de segments, notamment leur transdisciplinarité et leur éventuel rapprochement avec les constructions spécificationnelles, nous voulons discuter des limites de notre travail et des perspectives qu'il ouvre.

IV. Discussion sur les limites et les perspectives

IV.1 Limites de notre travail

TODO

Citer des têtes dont ne sait pas s'ils sont bien des NSS ou pas.

Les titres complexes avec plus de deux segments.

Limites techniques de l'outil : Talismane, à la manière d'une lunette astronomique, a rendu possible la perception de phénomènes dans notre corpus qui, en retour, demandes des hypothèses explicatives. Mais l'outil est faillible et nous devons en être conscient pour ne pas lui être inféodé (cas du mauvais élément régisseur dans une relation de dépendance par exemple).

IV.2 Perspectives

TODO

Sémantique distributive pour étudier les compléments nominaux des NSS.

Conclusion

TODO

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.

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.

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.

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. Étiquettes de catégorie de discours de Talismane

A2.1 Catégorie morphosyntaxique

TODO

A2.2 Relationde dépendance

TODO

A3. Éléments techniques

TODO

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.

A4. Index des tableaux

TODO

A5. Index des graphiques

TODO

A6. Index des notions mentionnées

TODO