# Récurrences lexico-sémantiques dans les titres

## I. Problématique de recherche

Un titre de document scientifique est un emplacement 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 est le premier contact entre le document et les lecteurs et, dans 92 % des cas, le dernier (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), littérature dont la production augmente constamment en doublant tous les 12 ans (Stix, cité dans Salager-Meyer et al. 2013). Cette discrimination par le titre porte notre intérêt sur l'information contenue dedans et quels mots et quelles structures sont utilisés pour la convoyer. Nous considérons le titre uniquement dans sa fonction informationnelle, considérant qu'elle est la plus importante, soutenu en cela par Haggan (2004) et Hartley (2005), et la plus facile à analyser, en laissant de côté la fonction d'attraction qui peut considérablement obscurcir son sens (Hartley, 2005).

Poursuivant nos travaux de première année, nous voulons mettre au jour, dans un grand corpus de titres de travaux scientifiques de langue française, des récurrences lexico-syntaxiques et la sémantique associée. La majorité de la littérature traitant des titres d'articles de journaux scientifiques, nous nous limiterons à ceux-ci et à ceux construits de manière similaire : chapitres d'ouvrages collectifs et communications dans des conférences. Nous voulons étudier comment ces récurrences se répartissent entre les différentes disciplines scientifiques, des travaux antérieurs montrant l'existence de spécificités disciplinaires dans l'écriture des titres (Soler, 2011 pour l'anglais, Tanguy et Rebeyrolle, à paraître, pour le français). La principale différence méthodologique avec notre travail précédent est l'apport central de l'analyse syntaxique en dépendance (Tesnière, dans Schwischay, 2001) des titres.

Nous voulons utiliser l'analyse syntaxique dépendentielle pour vérifier que la structure syntaxique des titres est en grande majorité un ou plusieurs syntagmes nominaux (Leech, 2000 ; Haggan, 2004 ; Soler, 2007 ; Cheng et al., 2012 ; Wang et Bai, 2007). Lexicalement, nous nous intéressons aux noms noyaux de ces syntagmes. En premier, pour savoir si ces noms appartiennent au vocabulaire académique transdisciplinaire (Hatier et al., 2016). En second, pour chercher un trait commun aux plus fréquents. Les premières observations montrent que ce trait pourrait être un faible contenu sémantique. La question se pose alors de savoir pourquoi, dans un espace où chaque mot est compté, ils y figurent si fréquemment. Syntaxiquement, on s'interroge sur leurs positions au sein des syntagmes, leurs déterminants ou la présence de compléments, et on cherche à détecter des structures lexico-syntaxiques récurrentes dont ils seraient les pivots. Enfin, sémantiquement, nous voulons essayer de montrer quelles informations ces structures convoient et comment elles structurent l'ensemble du titre.

## II. Résumé de l'état de l'art

Notre état de l'art a suivi deux grands fils directeurs. Le premier concerne les travaux sur les titres scientifiques, très nombreux en ce qui concerne l'anglais, alors qu'il en existe peu sur les titres français (Ho-Dac et al., 2001 ; Rebeyrolle et al., 2009). Certains articles (Aleixandre-Benavent et al., 2014) ou livres (Swales et Feak, 1994 ; Gustavii, 2008) ont une ambition didactique et prescriptive et donnent des conseils d'écriture en essayant de définir ce qu'est un « bon » titre et par extension, ses buts et son contenu (Grant, 2013). De ce contenu, d'autres articles en proposent une typologie comme la suite de travaux de Swales et Feak (1994), Anthony (2001) et Cheng et al. (2012). Une typologie peut également être basée sur la syntaxe, comme le travail de Jamali et Nikzad (2011). D'autres travaux testent l'incidence de la longueur du titre (Hartley, 2007 ; Jamali et Nikzad, 2011, Paiva et al., 2012), la présence d'humour (Sagi et Yechiam, 2008 ; Subotic et Mukherjee, 2014), d'une zone géographique précise (Jacques et Sebire, 2010) ou la présence d'un double point (Townsend, 1983) par rapport au nombre de citations et de téléchargements dans une perspective de performance, mais ils obtiennent des résultats non significatifs ou contradictoires (Merrill et Knipps, 2014). Notons toutefois un travail original mettant en relation le contenu avec le nombre de citations (Paiva et al., 2012), indiquant que les titres décrivant les résultats sont plus cités que ceux décrivant la méthode. D'autres travaux s'attachent à montrer les liens entre les caractéristiques des titres, comme 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). Si ce n'est la présence de délimiteurs (Anthony, 2001), dont le double point (Dillon, 1981, 1982 ; Townsend, 1983 ; Diers et Downs, 1994 ; Lewison et Hartley, 2005) ou la présence d'un point d'interrogation (Ball, 2009) dans des perspectives diachroniques, peu de travaux portent sur la structure

syntaxique des titres. Seuls émergent l'utilisation d'un délimiteur, souvent un double point, articulant deux segments sémantiques comme *sujet : point particulier*, et la forte tendance du titre à être un syntagme nominal (Leech, 2000 ; Haggan, 2004 ; Soler, 2007), 93 % des titres pour le corpus de Cheng et al. (2012), 99 % pour celui de Wang et Bai (2007), et non une phrase avec un verbe conjugué comme noyau, bien que cela varie selon la discipline (Soler, 2007). C'est justement cette différence des caractéristiques des titres entre les différentes disciplines que comparent plusieurs travaux (Haggan, 2004 ; Lewison et Hartley, 2005 ; Soler, 2007 ; Nagano, 2015).

Le second fil aborde les noms au faible contenu sémantique. Partant des *noms généraux* définis par Halliday et Hasan (1976), de nombreux travaux portent sur l'emploi de ces *« noms abstraits dont le sens complet peut seulement être spécifié en référence à son contexte »* (Flowerdew, 2006). 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). Une *construction spécificationnelle* (Legallois, 2008) met en relation le nom et le contenu spécifiant. La nature syntaxique du contenu est également sujette à débat. Les définitions opératoires employées par Schmid (2000) le limitent à une proposition introduite par *que*. Flowerdew et Forest (2015) y ajoutent *Nom en emploi + Préposition + Syntagme nominal*, ouvrant la possibilité de relier les noms des titres à cette classe.

Un troisième fil, mineur, est constitué de travaux sur l'analyse syntaxique en dépendance (Schwischay, 2001), les corpus (Cordi et David, 2008) et les grammaires de constructions et de patterns (François et Legallois, 2006).

# III. Origine et description des données

## III.1 Origine 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.

## III.2 Description des données

Chaque titre est fourni avec les informations supplémentaires suivantes relatives au document titré : nombre d'auteurs, type de document, année de publication et domaine scientifique. HAL permet d'attribuer plusieurs domaines à un document. Les domaines sont organisés en arbre, 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.

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 : forme dans le titre, lemme (pour les mots), classe grammaticale, informations complémentaires, élément régisseur, type de dépendance. 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.

## IV. Méthode employée et premiers résultats obtenus

Nous voulons, grâce à l'analyse syntaxique en dépendances, pousser plus loin l'analyse syntaxique des titres et les liens entre leurs différents segments tout en surveillant les éventuelles variations entre chaque discipline.

Dans un premier temps, nous avons analysé les titres selon trois axes : parties, segments et racines. Talismane considère, de façon erronée, certains titres comme ayant plusieurs paragraphes et son analyse de dépendance est effectuée paragraphe par paragraphe. Nous préférons l'appellation de parties à celle de paragraphes. Nous avons découpé les titres en segments en reprenant les délimiteurs d'Anthony (2001) : le double-point, le point-virgule et les différents points. Nous avons écarté des délimiteurs le tiret car il sert en français dans les mots composés. Nous avons étudié les deux dimensions ensemble dans le cadre d'une analyse par segmentation des données. Nous construisons une nouvelle caractéristique des titres, que nous nommons structure, notée *nombre de parties : nombre de segments*. En cumulant les quatre premières combinaisons, 1:1, 1:2, 2:2, 1:3, nous couvrons 94 % du corpus, dont 52 % pour la première et 29 % pour la seconde. Pour l'instant notre travail porte sur les combinaisons 1:1 et 1:2.

Nous étudions ensuite, pour les deux combinaisons sélectionnées, les racines : des mots uniquement régisseurs dans l'analyse dépendentielle. D'abord nous étudions leur nombre. Pour une structure 1:1, 88 % des titres ont une seule racine, 10 % en ont deux. Pour une structure 1:2, 65 % des titres ont une racine et 31 % en ont deux. Il est intéressant de voir comment les deux racines se répartissent entre les deux segments : 84 % des titres ont une racine dans chaque segment, 8 % en ont deux dans le premier, 7 %. 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 incluant une racine et vérifier que les titres sont avant tout des syntagmes nominaux.

Ce projet se heurte à deux problèmes. Le premier est celui des segments avec plusieurs racines. Un cas simple est deux racines reliées par une conjonction de coordination mais des cas, voir des erreurs d'analyses de Talismane, rendent ce problème non trivial. Le second est celui des segments sans racine. Une piste que nous avons observée est la présence dans ces segments d'un mot dépendant uniquement de la racine de l'autre segment. Si on considère le segment auquel ils appartiennent, ils sont uniquement régisseurs, on les appellera des racines *secondaires* pour les différencier des véritables racines, dites *primaires*. Nous pouvons ensuite étudier la catégorie morphosyntaxique de toutes les racines et caractériser le type de nos segments puis du titre et comment cette structure varie en fonction de la discipline.

Comme les titres sont avant tout constitués de syntagmes nominaux, on étudie après les racines nominales. En premier lieu, en constituant un lexique des racines et en étudiant leur transdisciplinarité. En deuxième lieu, en essayant de trouver des traits communs à ces noms : appartenance au vocabulaire académique (Hatier, 2016) ou à une classe de noms (Huyghe, 2015). Nos premières observations suggèrent qu'il s'agit de noms généraux (Adler et Moline, 2018) et nous voulons étudier si leur emploi rejoint celui des signalling nouns définis par Flowerdew et Forest (2015). Ensuite, nous étudions l'article utilisé pour les déterminer. 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 la majorité de ces groupes utilisent *of* ou *in*. Il serait intéressant de dresser un panorama de ces groupes prépositionnels compléments de nom. On peut surtout étudier la position de ces racines dans le 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.

## IV. Proposition d'échéancier sur la période finale

| Action Semaine ➔ | 20 – 26/5 | 27/5 – 2/6 | 3 – 9/6 | 10 – 16/6 | 17 – 23/6 | 24 – 30/6 | 1 – 7/7 |
|---|---|---|---|---|---|---|---|
| Lectures complémentaires | DG | DG | | | | | |
| Développement des outils | DG | DG | | | | | |
| Requêtes sur corpus | DG | DG | DG | | | | |
| Rédaction du mémoire | DG | DG | DG | DG | | | |
| 1ère lecture | | | | | Directeurs | | |
| Corrections du mémoire | | | | | | DG | |
| 2nd lecture | | | | | | | Directeurs |
| Soutenance finale | | | | | | | DG |