 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). La lecture du titre seul, indépendamment du document titré, permet donc au chercheur de faire l’essentiel du tri parmi la littérature scientifique (Goodman et al., 2001), dont la production augmente constamment (Jacques et Sebire, 2010), en doublant tous les 12 ans (Stix, cité dans Salager-Meyer et al. 2013). Cette discrimination par le titre suscite notre intérêt pour l’information contenue dans le titre, son sens et quels mots et quelles structures sont utilisés pour la convoyer. Nous nous restreignons à une prise en compte du titre uniquement comme objet informationnel, considérant qu’il s’agit de là de son aspect le plus important, soutenu en cela par Haggan (2004) et Hartley (2005), et du plus facile à analyser, laissant de côté son aspect attractif 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 leurs sémantiques associées. La majorité de la littérature traitant des titres d’articles de journaux scientifiques, nous nous limiterons à ceux-ci et à ceux construit de manière similaire : chapitres d’ouvrages collectifs et communications dans des conférences. Nous voulons également 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 (Schwischay, 2001), définie par Tesnière dans les années 30, des titres.

Nous voulons en premier lieu utiliser l’analyse syntaxique dépendentielle pour interroger la nature syntaxique des titres. Leech (2000) et nos premières observations manuelles indiquent que les titres sont en majorité des syntagmes nominaux. Lexicalement, nous nous sommes intéressés aux noms noyaux de ces syntagmes. En premier, pour chercher si ces noms étaient transdisciplinaires et s’ils appartenaient bien au vocabulaire académique des textes (Hatier et al., 2016). En second, pour chercher à définir un trait commun aux noms les plus fréquents. Les premières observations laissent à penser 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 fréquemment. Syntaxiquement, on peut s’interroger sur leurs positions au sein du titre, comment ils sont déterminés ou la présence de complément et chercher à détecter des structures lexico-syntaxiques récurrentes. Enfin, sémantiquement, nous voulons essayer de montrer quelles informations ces structures convoient.

Dans ce document, nous donnons un rapide résumé de l’état de l’art avant de présenter les données et les méthodes employées. Nous présentons ensuite les premiers résultats obtenus avant de proposer un échéancier pour la période finale. Une annexe en trois parties complète ce document, contenant notre bibliographie, des informations sur notre corpus et nos outils.

# II. Résumé de l’état de l’art

Pour constituer notre état de l’art, nous avons suivi la méthodologie suivante : à partir d’un petit groupe de références initiales, nous regardons les articles cités dans chacune pour croiser les références données. Nous formons ainsi un réseau d’articles, chaque nœud étant un article étiqueté selon ses principales thématiques. Nous disposons ainsi d’une carte pour guider nos lectures, à l’aide des étiquettes et du degré de chaque nœud, informant de la connectivité de l’article avec les autres.

Notre état de l’art a suivi trois 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 lignée des trois 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 de la présence d’un double point (Townsend, 1983) par rapport nombre de citations et de téléchargements dans une perspective de performance mais avec des résultats faibles et mitigés (Merrill et Knipps, 2014). Notons un travail original mettant en relation le contenu avec le nombre de citation (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 et le nombre d’auteur (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 ayant un verbe conjugué comme noyau, bien que cela varie selon la discipline. C’est justement cette différence des caractéristiques des titres entre les différentes disciplines que se sont attelés à comparer plusieurs travaux (Haggan, 2004 ; Lewison et Hartley, 2005 ; Soler, 2007 ; Nagano, 2015). 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.

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, comme la liste des noms pouvant être employé de la sorte, ainsi que l’appellation 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), *noms généraux sous-spécifiés* (Legallois, 2008) et *noms porteurs* (Huygue, 2018). La définition opératoire employée par Schmid, limitée techniquement, est élargie par Flowerdew jusqu’à devenir *Nom en emploi + Préposition + Contenu spécifiant*, cette dernière définition étant déjà annoncée par Schmid (2018). Nous voulons savoir si nous pouvons nous rattacher aux définitions précédentes, soit en spécifiant une sous-classe, soit en spécifiant notre propre classe de noms.

Le troisième fil, mineur, est constitué de lecture sur les outils employés, notamment l’analyse syntaxique en dépendance (Schwischay, 2001), les corpus (Cordi et David, 2008) et les différents cadres théoriques comme les grammaires de constructions et de patterns (François et Legallois, 2006).

# III. Données et méthode employées

## 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 du catalogue des grands éditeurs scientifiques.

Nous avons utilisé 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 scientifiques.

## 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 | 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 avons également utilisée.

Les titres ont été analysé à l’aide du logiciel Talismane (Urieli et Tanguy, 2013 ; Urieli, 2013) qui fournit un découpage en différents éléments, nous regroupons par ce terme les mots et les signes de ponctuation, 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 sont des textes très travaillés qui ne nécessitent pas de prétraitement pour corriger les fautes d’orthographe ou une grammaire défaillante.

## III.3 Méthode employée

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 ensuite découpé les titres en segments en reprenant les délimiteurs d’Anthony (2001) : le double-point, le point-virgule et les différents points. Nous avons néanmoins écarté des délimiteurs le tiret car il sert en français dans les mots composés.

Nous avons ensuite étudié les deux dimensions ensemble pour rassembler les combinaisons les plus nombreuses dans le cadre d’une analyse par segmentation des données.

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre de parties | Nombre de segments | Nombre de titres | Pourcentage des titres |
| 1 | 1 | 175 915 | 52 % |
| 1 | 2 | 98 931 | 29 % |
| 2 | 2 | 26 077 | 8 % |
| 1 | 3 | 18 653 | 5 % |

En cumulant les quatre premières combinaisons, nous couvrons 94 % du corpus. Nous construisons une dimension secondaire du titre, combinant deux dimensions primaires, que nous nommons structure et notée nombre de parties : nombre de segments.

Nous avons mis en rapport la structure avec le domaine.

XXX MATRIX XXX

Nous avons ensuite les combinaisons de façon indépendante. Pour l’instant notre travail a porté sur les combinaisons les plus nombreuses notées 1:1 et 1:2, couvrant 81 % du corpus. Pour les titres de structure 1:2, une règle logique veut qu’une partie soit composée d’un ou plusieurs segments, mais un segment ne peut jamais s’étaler sur plusieurs parties, autrement dit le nombre de segments ne peut être inférieur au nombre de parties, mais cette condition n’est pas suffisante, et le dernier élément d’une partie suivie d’une autre partie doit être un délimitateur de segment. Nous écartons 462 titres enfreignant cette règle logique, une analyse manuelle sommaire montrant qu’ils ont des problèmes d’encodage.

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 :

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre de racines dans segments 1 | Nombre de racine dans segments 2 | Nombre de titres | Pourcentage des titres |
| 1 | 1 | 25 938 | 84 % |
| 2 | 0 | 2 454 | 8 % |
| 2 | 2 | 2 179 | 7 % |
| 1 | 3 | 173 | >1 % |

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 déterminer ainsi la structure syntagmatique du titre que de nombreux travaux soulignent comme nominale (XXXX).

Ce projet se heurte à deux problèmes. Le premier est celui des segments avec plusieurs racines. Le cas simple est deux racines reliées par une conjonction de coordination (et) 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 d’un 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 étudier la catégorie morpho-syntaxique de toutes les racines primaires et secondaires. Comme les titres sont avant tout constituer de syntaxe nominaux, on peut ensuite se concentrer sur ces noms.

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 ; Hatier et al., 2016) ou à une classe de noms (Huyghe, 2015). Nos premières observations laissent suggérer qu’il s’agit de noms généraux (Adler et Moline, 2018) et nous voulons étudier si leur type d’emploi ne rejoint celui des signalling nouns définis par Flowerdew notamment dans Flowerdew et Forest (2015). Ensuite, nous pouvons étudier 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 également étudier la position de ces racines dans le segment. Roze et al. (2014) indique l’existence d’un schéma *Nom sous spécifié : suite* pour *exemple*, *problème* et *question*, ce qui nous laisse à penser que la position où se trouve la racine dans le segment a toute son importance.

# IV. Premiers résultats obtenus

# V. Échéancier sur la période finale

Nous proposons ce diagramme de Gantt pour la période finale, avec pour objectif de remettre le document finalisé la semaine du 17 juin à nos directeurs pour une première lecture. N’ayant pas consulté nos directeurs, cet échéancier demeure hypothétique et soumis à révisions.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Semaine 🡺 | 20 – 26 mai | 27 mai – 2 juin | 3 – 9 juin | 10 – 16 juin | 17 – 23 juin | 24 – 30 juin | 1 – 7 juillet |
| Action |
| 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 |

# Bibliographie

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.

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.

Salager-Meyer, F. & Alcaraz Ariza, M. Á. (2013). Titles are" serious stuff": a historical study of academic titles. *Jahr, 4(7), 257-271.*

Sagi, I., & Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science, 34(5)*, 680-687.

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.

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.

# Annexe 1 : Information sur notre corpus

Notre corpus compte 339 687 titres.

Répartition par disciplines

Ordre Discipline Titres % du corpus % cumulés

1. 0.phys 35538 10.4620 % 10.46 %

2. 1.shs.socio 32228 9.4876 % 19.95 %

3. 1.shs.droit 30517 8.9839 % 28.93 %

4. 1.shs.hist 25764 7.5846 % 36.52 %

5. NONE 25115 7.3936 % 43.91 %

6. 0.sdv 24638 7.2531 % 51.16 %

7. 1.shs.gestion 23703 6.9779 % 58.14 %

8. 0.info 16932 4.9846 % 63.13 %

9. 1.shs.ling 16028 4.7185 % 67.85 %

10. 1.shs.litt 14938 4.3976 % 72.24 %

11. 1.shs.archeo 13745 4.0464 % 76.29 %

12. 1.shs.edu 9696 2.8544 % 79.14 %

13. 1.shs.scipo 9432 2.7767 % 81.92 %

14. 1.shs.art 8907 2.6221 % 84.54 %

15. 1.shs.phil 8450 2.4876 % 87.03 %

16. 0.sde 7621 2.2435 % 89.27 %

17. 1.shs.infocom 7333 2.1588 % 91.43 %

18. 1.shs.anthro 7095 2.0887 % 93.52 %

19. 1.shs.archi 4761 1.4016 % 94.92 %

20. 0.sdu 3842 1.1310 % 96.05 %

21. 0.scco 3141 0.9247 % 96.98 %

22. 0.math 3092 0.9102 % 97.89 %

23. 0.chim 2836 0.8349 % 98.72 %

24. 1.shs.psy 2654 0.7813 % 99.51 %

25. 1.shs.geo 1149 0.3383 % 99.84 %

26. 0.qfin 433 0.1275 % 99.97 %

27. 1.shs.autre 99 0.0291 % 100.00 %

On note qu’au niveau des domaines, le corpus n’est pas équilibré. Il faut donc toujours procéder par fréquences relatives pour comparer une caractéristique d’une discipline à l’autre.

# Annexe 2 : Exemples de requêtes sur notre corpus

Notre code en Python sert essentiellement à faire des statistiques sur notre corpus et le filtrer.

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.