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, toujours plus abondante. D’où notre intérêt pour l’information contenue dans le titre, quels mots et quelles structures sont utilisés pour la convoyer.

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. Nous voulons également étudier comment ces récurrences se répartissent entre les différentes disciplines scientifiques, des travaux antérieurs montrant l’existence de spécificités disciplinaires (Soler, 2011 pour l’anglais, Tanguy et Rebeyrolle, à paraître, pour le français). La principale différence méthodologie avec notre travail précédent est l’apport central de l’analyse syntaxique en dépendance (Schwischay, 2001) définie par Tesnière pour l’analyse des titres.

Nous nous intéressons en premier à 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 l’existence d’un vocabulaire académique transdisciplinaire des titres et le comparer 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 et à a quelle fréquence. Syntaxiquement, on peut s’interroger sur les récurrences d’emplois : position au sein du titre, type d’article ou présence de complément. Enfin, si des structures récurrentes émergeaient, nous pourrions essayer de montrer quelles informations elles convoient.

Dans ce document, nous donnons un rapide résumé de l’état de l’art avec de montrer les données et méthodes employées. Nous présentons ensuite les premiers résultats obtenus avant de proposer un échéancier pour la période finale. Une bibliographie et plusieurs annexes complètent ce document.

# 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 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 article.

Notre état de l’art a suivi trois grands fils directeurs. Le premier concerne les titres de travaux scientifiques. Il existe de nombreux travaux sur les titres. Certains articles didactiques donnent des conseils d’écriture. D’autres articles proposent une taxonomie des titres et de leur contenu. Des travaux testent l’incidence de la longueur du titre par rapport au nombre d’auteurs ou au nombre de citations. Le second fil aborde les noms au faible contenu sémantique. Partant du concept de noms généraux définis par Halliday et Hasan (1976), de nombreux travaux portent sur l’emploi de ces noms dans un contexte qui les précise. La définition de cet emploi est sujet à débat, de même que l’appellation employée pour désigner cet emploi.

# 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 :

* Le nombre d’auteurs
* Le type de document
* L’année de publication
* Le 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) 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 :

* Sa forme dans le titre
* Son lemme (égale à sa forme pour les signes de ponctuation)
* Sa catégorie morphosyntaxique/classe grammaticale (ou ‘PONCT’ pour les signes)
* Des informations complémentaires dépendantes de la catégorie, comme le genre pour les noms, le mode et le temps pour les verbes
* Son élément régisseur
* Son type de dépendance

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 racines se répartissent entre les deux segments :

XXX PB CALCUL NB ROOTS BY SEGMENTS XXX

Nb Parts/segments 🡺 nb roots 🡺 part-of-speech roots / position roots / links

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 est déterminer ainsi la structure syntagmatique du titre.

Nous constitutions ainsi un lexique des racines.

2) classifier les titres, dire que c’est essentiellement des GN ou GN : GN (partie ou segment ? Fusion des deux pour ne parler plus que de segment & racine :-) Ça devrait correspondre…

3) étudier le noyau/racine des GN

4) transdisciplinaire ?

5) quel type de nom 🡺 faible contenu sém

6) position, article, complément de nom ?

7) essayer de mettre en relation les racines entre elles et avec les « racines de niveau 2 » : mot d’un autre segment n’ayant comme gouverneur que la racine du segment précédent.

8) rajout d’exemple

Nous avons écarté ces titres pluri-paragraphes, considérant que la détection de plusieurs paragraphes s’apparente à une erreur. Nous avons ensuite

déterminé la nature des titres.

pour les titres restants

les titres possédant plusieurs parties, plus complexes et ne comptant que pour X % des titres.

# 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 le 23 juin à nos directeurs.

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

# Bibliographie

Anthony, L. (2001). Characteristic features of research article titles in computer science. *IEEE Transactions on Professional Communication, 44(3)*, 187-194.

Hallliday, M. A. K. et Hasan, R. (1976). *Cohesion in English*. London: Longman.

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.

Mabe, M. A. et Amin, M. (2002). Dr. Jekyll and Dr. Hyde: Author-reader asymmetries in scholarly publishing. *Aslib Proceedings, 54(3)*, 149-157.

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>

Soler, V. (2011). Comparative and contrastive observations on scientific titles written in English and Spanish. *English for Specific Purposes*, 30(2), 124-137.

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>

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.