# **Projet de recherche pour le Master 2 LITL, 2018-2019**

## Sommaire

Introduction 1

I. Cadre données et outils 1

II. Problématique et démarche de résolution 2

III. Chronologie des étapes 3

Conclusion 3

Bibliographie 4

## Introduction

Mon mémoire de recherche de Master 1 portait sur les titres de documents scientifiques possédant un unique double point, divisant le titre en deux segments. Dans ces titres, notre problématique portait sur le premier syntagme survenant immédiatement après le double point. Nous essayions de déterminer l’apport sémantique de ce second segment au premier, en étudiant les récurrences de syntagmes binominaux. Se limiter au second segment et ne rien connaître du premier segment était handicapant pour appréhender globalement l’objet titre, surtout si l’on admet que les mots les plus importants surviennent en premier.

Un titre est un espace sémantiquement condensé car il est court, 15,5 mots en moyenne dans notre corpus précédent. Cette condensation pourrait favoriser l’utilisation de constructions récurrentes, ayant un degré de figement plus ou moins prononcé.

Nous voulons cette année élargir notre cadre de travail, tout en continuant d’étudier les titres de documents scientifiques, et approfondir notre travail de recherche, en choisissant une problématique plus ciblée. Dans ce document, nous commençons par décrire le cadre, les données et les outils que nous nous proposons d’utiliser. Nous définissons ensuite une esquisse de problématique et une démarche pour la résoudre, avant de proposer une chronologie des différentes étapes. Nous concluons ce document par une courte conclusion et une brève bibliographie.

## I. Cadre, données et outils

Pour notre mémoire de recherche de Master 2, nous souhaitons élargir le cadre de notre recherche selon deux axes : ne plus nous limiter aux titres avec un unique double point et étudier l’ensemble du titre. Le double point pourra jouer encore un rôle important, mais notre recherche n’est plus centrée sur lui.

Nous gardons donc comme données de base des titres extraits de l’archive ouverte HAL écrits en français. M. Tanguy a procédé à la constitution d’un corpus de 500 000 titres, à partir de la méthode décrite dans notre mémoire de recherche de Master 1, sur lequel nous travaillerons. Nous filtrerons cette année les titres selon le type de document pour ne garder que les titres d’articles, de chapitres d’ouvrage et de communications lors de congrès car ils possèdent une cohérence dans la façon de les écrire.

Nous avons remarqué que la division en domaines scientifiques de HAL présente un désavantage : les sciences dures majeures, comme l’informatique ou les sciences du vivant, bénéficient d’une racine propre dans l’arbre des disciplines. Les sciences humaines sont, elles, toutes regroupées sous un niveau zéro appelé *Sciences de l'Homme et Société*. Si on divise les titres en disciplines selon le niveau 0, cela déséquilibre les analyse et les contraint dans leur granularité. Mme Rebeyrolle et M. Tanguy, dans un travail à paraître, ont construit une nouvelle classification que nous nous proposons d’adopter.

Pour les outils, nous souhaitons continuer à utiliser Talismane et Python. Notre premier travail nous a familiarisé avec Talismane et nous pensons pouvoir affiner notre utilisation de celui-ci en comprenant mieux son fonctionnement et en essayant de compenser certains ses défauts. Nous souhaitons mettre en œuvre, en plus de la catégorisation et de la lemmatisation des titres, l’analyse syntaxique en dépendances que cet outil propose sur l’ensemble du titre. Nous reprenons aussi le langage de script Python, pour lequel nous avons déjà écrit de nombreux scripts, notamment une interface entre Python et Talismane, qui nous serviront à bâtir les suivants.

Nous avons délimité notre cadre, défini les données sur lesquels nous travaillerons et établi nos outils de départ. Nous pouvons à présent esquisser notre problématique et notre démarche pour la résoudre.

## II. Problématique et démarche de résolution

Notre problématique prend pour fondation la liste des noms généraux sous-spécifiés (NGSS désormais) définis par Schmid (2000) pour l’anglais, qu’il appelle des *shell nouns*. Ces noms sont principalement caractérisés par un emploi dans lequel leur contenu sémantique propre est faible et c’est le contexte qui donne l’essentiel de leur contenu sémantique. En projetant des NGSS sur notre corpus, on pourra détecter les constructions les employant.

Par exemple, en prenant le NGSS *outil*, nous explorons HAL pour énumérer ses emplois. Pour ce document, nous avons sélectionné trois exemples :

1. Gaëtan Temperman, Christian Depover, Bruno De Lièvre. Le tableau de bord, un outil d'awareness asynchrone. Jun 2007, INRP, 2007.
2. Makram Chemangui, Christine Noël Lemaitre. LES CARTES CONCEPTUELLES, UN OUTIL FAVORISANT LA COOPERATION ENTRE LES CHERCHEURS ET LES PROFESSIONNELS DE L'AUDIT

*La place de la dimension européenne dans la Comptabilité Contrôle Audit*, May 2009, Strasbourg, France. pp.CD ROM, 2009

1. Karen Godary-Dejean. LPT : Little Parametric Tool, outil pour la validation d'une borne temporelle paramétrée*. CIFA: Conférence Internationale Francophone d'Automatique*, Sep 2008, Bucarest, Roumanie. 2008

Dans ces trois exemples, le mot *outil*, qui peut faire penser intuitivement à des outils de bricolage, a un contenu sémantique très faible. De ce fait, il a une capacité référentielle très grande, ou pour le dire autrement, il peut renvoyer à un grand nombre de référents.

On peut identifier une construction (ou « schéma », « patron ») que nous formalisons sous la forme suivante, le point d’interrogation indiquant une optionalité, les crochets indiquant un choix parmi les éléments entre, et la balise, encadrée par l’inférieur et le supérieur, indiquant un élément sémantique :

<OUTIL>, un? outil [favorisant pour de] <BUT>

<OUTIL> est le référent, celui que l’on désigne comme un outil. <BUT> est ce que l’emploi de l’outil permet de faire.

Après avoir identifié cette construction à partir d’un NGSS, nous pouvons nous demander si d’autres noms sont employés dans cette même construction. On peut s’attendre à retrouver d’autres NGSS, soit un déjà identifié par Schmid (2000), soit un que nous ajouterons à la liste. On pourra alors élargir la définition de la construction identifiée en remplaçant le NGSS *outil* par un ensemble de NGSS de sens éventuellement proches. Il s’agit d’établir des classes de NGSS utilisant les mêmes constructions. De même un NGSS utilisé dans une construction peut être également employé dans une autre construction. On peut se poser la question de savoir à quel point ces constructions sont semblables. Si elles le sont, on pourrait les fusionner, élargissant ainsi la formalisation de la construction. Sinon, on peut se demander si elles sont équivalentes sémantiquement pour établir des classes de constructions.

Nous pourrons ensuite comparer l’utilisation des NGSS et des constructions liés à ceux-ci dans les différentes disciplines scientifiques. On essayera d’établir si certains NGSS ou constructions sont spécifiques à une discipline ou suremployés dans certaines par rapport à d’autres. Au contraire, certains NGSS ou constructions pourraient appartenir à ensemble commun à toutes les disciplines.

Enfin, on pourra essayer, à partir des éléments sémantiques repérés, comme <OUTIL> et <BUT> dans notre exemple, de faire de l’extraction d’information (EI désormais) pour construire une base de données modélisant une discipline, notamment les problèmes qu’elle traite. Pour le choix de la discipline, nous pensons au traitement automatique des langues car c’est une des disciplines que nous connaissons le mieux (codée dans HAL sous *Traitement du texte et du document* et *Linguistique*).

## III. Chronologie des étapes

Dans cette partie, nous avançons une liste d’étapes pour accomplir notre démarche, sans nous avancer sur un calendrier, celui-ci devant être élaboré avec nos codirecteurs et pouvant encore être modifié dans les premiers mois de notre travail :

1. Récupération du corpus de 500 000 titres en français
2. Projection des NGSS sur le corpus
3. Identification de constructions récurrentes employant des NGSS
4. Identification de NGSS utilisant les mêmes constructions (regroupement de NGSS)
   1. Regarder les relations sémantiques entre les NGSS regroupés
5. Identification de construction utilisant les mêmes NGSS (regroupement de constructions)
   1. Regarder les relations sémantiques entre les constructions regroupées
6. Analyse de l’utilisation des constructions dans les différentes disciplines
   1. Différenciation des disciplines
   2. NGSS et constructions communes
7. Essai d’extraction d’information et de constitution d’une base de données

## Conclusion

L’analyse statistique des textes, notamment dans l’analyse sémantique distributionnelle, est actuellement très utilisée. Notre approche s’en éloigne pour revenir à une analyse fine de récurrences lexicales et syntaxiques. Tout en élargissant et approfondissant les thématiques de notre premier mémoire, nous réutiliserons tout ce que nous avons appris sur les titres, les données et les outils pour répondre à cette nouvelle problématique autour des NGSS et des constructions qui les emploient. Nous souhaitons les identifier, essayer de les regrouper et analyser leur répartition dans les différentes disciplines scientifiques. Enfin, nous aimerions essayer d’en extraire des informations sémantiques pour constituer une base de données.

## Bibliographie

Adler, S. (2018). Sémantique des noms généraux sous-spécifiés et construction du sens. *Langages, 210*(2), 71-86.

Aktas, R. N., & Cortes, V. (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes*, *7*(1), 3-14.

Aktas, R. N. (2005). *Functions of "shell nouns" as cohesive devices in academic writing: a comparative corpus-based study*. Ames: dissertation, Iowa State University.

Flowerdew, J., & Forest, R. W. (2015). *Signalling Nouns in Academic English*. Cambridge University Press.

Huyghe, R. (2015). Les typologies nominales : présentation. *Langue française, 185*, 5-27.

Legallois, D. & Tutin, A. (2013). Présentation : Vers une extension du domaine de la phraséologie. *Langages, 189*(1), 3-25.

Paquot, M. (2007). Towards a productively-oriented academic word list.

Schmid, H.-J. (2018). Shell nouns in English - a personal roundup. *Caplletra. Revista Internacional de Filologia, 2018*(64)*,* p. 109-128. https://www.raco.cat/index.php/Caplletra/article/view/335905 [Consulta: 03-10-18]

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin: Walter de Gruyter.

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.

Yu, H. (2016). *Shell Nouns as Cohesive Devices in Published Research Articles*. Hong-Kong: Doctoral dissertation, Hong Kong Polytechnic University, Department of English.