

MÉMOIRE DE RECHERCHE

Département des Sciences du Langage

M1 Linguistique, Informatique et Technologies du Langage

STRUCTUCTURES LEXICO-SYNTAXIQUES RÉCURRENTES DANS LES TITRES SCIENTIFIQUES APRÈS LE DOUBLE POINT

**:**

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

# Remerciements

J’aimerais remercier mes deux codirecteurs de recherche, Mme Josette Rebeyrolle et M. Ludovic Tanguy qui m’ont accompagné par leurs conseils et leurs encouragements tout au long de ce travail d’une année.

Je tiens également à remercier Mme Cécile Fabre et Mme Lydia-Mai Ho-Dac qui, avec M. Tanguy, ont accueilli ma démarche de reprise d’études avec intérêt et bienveillance. Cette première année de master LITL m’a permis de combiner mes deux grands intérêts que sont la linguistique et l’informatique, en me faisant arpenter de nouveaux chemins. Le rythme est soutenu, mais pour rien au monde je ne regrette ce voyage.

Je voudrais saluer mes camarades de promotion qui ont voyagé avec moi. Peu importe les aléas de la route, lorsque l’on a de bons compagnons.

Et je remercie celle et ceux qui m’ont vu cheminer toute l’année au lieu de partager pleinement leurs vies. Qu’ils me pardonnent, je rentre à la maison.

Table des matières

[Remerciements 3](#_Toc519785357)

[Introduction 6](#_Toc519785358)

[I. Les titres scientifiques 7](#_Toc519785359)

[I.1 Problématiques étudiées 7](#_Toc519785360)

[I.2 Corpus utilisés 7](#_Toc519785361)

[I.3 Caractéristiques des titres 7](#_Toc519785362)

[I.4 Typologies des titres 7](#_Toc519785363)

[II. Corpus de travail 7](#_Toc519785364)

[II.1 Présentation de HAL et extraction des données 7](#_Toc519785365)

[II.2 Traitement des données et exemples 9](#_Toc519785366)

[A) Enrichissement des données 9](#_Toc519785367)

[B) Conversions 9](#_Toc519785368)

[C) Filtrage 11](#_Toc519785369)

[II.3 Constations et exemples 13](#_Toc519785370)

[II.3.1 Caractéristiques de nos corpus 13](#_Toc519785371)

[II.3.2 Connexions avec les autres études 15](#_Toc519785372)

[II.3.3 Quelques exemples de titres 18](#_Toc519785373)

[III. Syntagmes et patrons 19](#_Toc519785374)

[III.1 Le syntagme après le double point 19](#_Toc519785375)

[III.2 Définition des patrons 20](#_Toc519785376)

[III.3 Présentations de nos trois patrons 20](#_Toc519785377)

[IV. Études des trois patrons 21](#_Toc519785378)

[IV.1 Patron SN + P + SN 21](#_Toc519785379)

[V. Étude de structures particulières 21](#_Toc519785380)

[VI. Résultats et discussions 21](#_Toc519785381)

[VI.1 Résultats 21](#_Toc519785382)

[VI.2 Discussions 21](#_Toc519785383)

[Conclusion 22](#_Toc519785384)

[Bibliographie 23](#_Toc519785385)

[Annexes 24](#_Toc519785386)

[A1. Requêtes Apache Solr sur HAL 24](#_Toc519785387)

[A1.A Requêtes 24](#_Toc519785388)

[A1.B Résultats 24](#_Toc519785389)

[AX. Index des logiciels et technologies mentionnés 25](#_Toc519785390)

# Introduction

A

Aa thème, problématique, organisation

Structures dans les titres scientifiques

Introduction   
 Sujet : les titres scientifiques

Problématique : peut-on trouver des structures lexico-syntaxique récurrentes dans les titres scientifiques ? Comment peut-on les caractériser dans leur nature et leur usage ?

À l’aide de scripts écrits dans le langage de programmation Python[[1]](#footnote-1) et du logiciel de tableur Excel[[2]](#footnote-2), nous …

# I. Les titres scientifiques

## I.1 Problématiques étudiées

Les titres scientifiques ont déjà été étudiés avec différentes problématiques. Nous pouvons résumer ces problématiques en deux mouvements :

1. L’analyse de la constitution des titres, en en proposant éventuellement une typologie. Le point analysé peut être l’ensemble du titre ou un point particulier, comme sa longueur ou l’utilisation du double point.
2. La mise en rapport de cette analyse avec une ou plusieurs métadonnées du titre. Il peut s’agir du nombre d’auteurs, du domaine scientifique du texte titré, du nombre de téléchargement ou de citations du texte, ou encore l’adéquation avec des règles d’écriture.

## I.2 Corpus utilisés

## I.3 Caractéristiques des titres

## I.4 Typologies des titres

# II. Corpus de travail

Dans cette partie nous présentons notre corpus de travail et la méthode suivie pour l’obtenir. Nous commençons par présenter son origine et le travail d’extraction qui a été fait. Nous abordons après les traitements effectués sur les données brutes pour aboutir à notre corpus ainsi que quelques exemples de titres. Nous présentons ensuite l’outillage utilisé et les premières constatations effectuées sur notre corpus.

## II.1 Présentation de HAL et extraction des données

Nos titres sont issus de l’archive ouverte Hyper Article en Ligne[[3]](#footnote-3) (HAL) (Nivard, 2010). Elle compte, au 14 juillet 2018, 524 452 documents scientifiques et 1 563 014 notices. Chaque chercheur, quelle que soit sa discipline, ou documentaliste d’un centre de recherche, est libre de déposer un document sur HAL, s’il a l’accord de ses auteurs et de son éventuel éditeur. Ce document peut-être un texte, comme un article, un livre ou seulement un chapitre, une vidéo, un son, une image ou une carte. Pour les articles, contrairement à une publication dans une revue scientifique, il n’y a pas de contrôle par les pairs du contenu scientifique déposé. Seul un contrôle pour s’assurer du bon format du document et du respect des droits est effectué. En le déposant sur HAL, le document est rendu public et est partagé avec la communauté scientifique beaucoup plus rapidement que via une revue. Les deux options peuvent être complémentaires pour diffuser son travail. Un article déposé sur HAL sans être publié dans une revue à ce moment-là est appelé un preprint.

HAL est géré par le Centre pour la Communication Scientifique directe[[4]](#footnote-4) (CCSD), fondé en 2000 et rattaché au Centre National pour la Recherche Scientifique (CNRS). Il existe des sous-ensembles de HAL dédiés à une discipline spécifique, HAL-SHS et MédiHAL, ou pour un type de texte spécifique comme Thèses en ligne. Les avantages des archives ouvertes par rapport à un site d’une institution particulière ou le site web personnel d’un chercheur sont la centralisation de l’accès, la diffusion des connaissances et la conservation pérenne des documents. La création des archives ouvertes s’inscrit dans le mouvement pour un accès libre et gratuit aux connaissances scientifiques. La plus ancienne des archives ouvertes est arXiv[[5]](#footnote-5) , fondée en 1991 et limitée uniquement aux articles. Un dépôt d’un article dans HAL entraîne automatiquement la création d’une notice dans arXiv si elle entre dans les disciplines couvertes par cette dernière.

Une notice est créée sur HAL lors du dépôt du document et éventuellement dupliquée dans d’autres archives ouvertes. Une notice est un ensemble d’informations sur le document scientifique déposé, appelé métadonnées, comme son titre, sa date de dépôt, son type. La notice contient tout ce qui est nécessaire à notre travail. Pour notre travail, nous considérons que les métadonnées du document sont également celles de son titre.

Une archive ouverte A peut avoir la notice d’un texte scientifique hébergé sur une autre archive ouverte B, cette dernière aura alors à la fois la notice et l’intégralité du document. Dernier cas possible, il existe des documents qui ne sont pas hébergés par aucune archive ouverte mais simplement référencés par leurs notices. Il s’agit généralement de textes dont les droits appartiennent à des revues payantes. La création de telles notices se fait par le traitement automatisé des références bibliographiques des documents déposés.

Pour récupérer ces notices, il existe deux protocoles. Le premier est le protocole de moissonnage standardisé Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), version 2.0 . Ce protocole est standardisé, on peut donc accéder à d’autres archives ouvertes avec, dont arXiv, mais il est assez limité dans ses possibilités de sélection : de base, seule la possibilité de sélectionner des notices en fonction de la date est offerte. Les gestionnaires d’archives peuvent définir des sélecteurs supplémentaires mais ils sont encore très limités. Ce protocole est surtout conçu pour les robots s’assurant de la réplication des notices entre différentes archives ouvertes.

Le second protocole offert pour accéder à HAL est bien plus intéressant dans notre cas. Il permet de sélectionner finement nos données et repose sur Apache Solr (Smiley, Pugh, Parisa, & Mitchell, 2015), le moteur de recherche du projet Apache Lucene[[6]](#footnote-6). On peut sélectionner les notices à retourner en filtrant sur le contenu d’une des métadonnées, on peut définir le format des données en sortie parmi un large choix de standards (dont XML, CSV ou JSON). C’est ce second protocole et le format de sortie JSON, avec encodage des caractères en UTF-8, que nous utiliserons. JSON est un format simple de données structurées stockées dans un fichier texte dont les constituants sont les données atomiques (chaîne de caractères, nombres), les listes et les dictionnaires associant une clé à une valeur (Bray, 2017). Une requête Solr repose sur le protocole de transfert hypertexte (HTTP). Elle se présente donc sous la forme d’une adresse internet (URL) qui peut être testée de façon simple et rapide dans un navigateur internet pour visualiser son résultat.

Nous avons automatisé, à l’aide d’un script Python, la création et l’envoi de requêtes ainsi que la récupération et la sauvegarde des résultats retournés au format JSON. En une demi-heure, nous avons récupéré 304 600 titres ainsi que leurs métadonnées comme l’année de dernière modification de la notice, les domaines scientifiques associés au document, ses auteurs, son type et son identifiant unique. Nous présentons dans l’[annexe A1](#_A1._Requêtes_Apache) l’ensemble des requêtes utilisées.

## II.2 Traitement des données et exemples

Une fois les données brutes obtenues, nous avons effectués plusieurs traitements dessus qui se divisent en trois catégories : l’enrichissement des données, la conversion et le filtrage. Notons que l’avantage des titres est qu’ils résultent d’une production très travaillée par rapport à celle que d’un tweet, d’un SMS ou d’un texte sur un forums, soignée et relue. La présence de faute de grammaire ou d’orthographe est donc quasi-nulle, ce qui évite d’avoir à les considérer en ajoutant des traitements de tolérance ou de réparation dans les algorithmes.

### A) Enrichissement des données

Nous voulions enrichir nos titres en déterminant pour chaque mot son lemme et sa catégorie grammaticale à partir de la forme présente dans le titre. Dans un premier temps, nous avons utilisé le logiciel Stanford Core Natural Language Processing[[7]](#footnote-7). Celui-ci fournissait pour le français les catégories mais non les lemmes. Nous avons donc abandonné Stanford Core NLP pour passer à un logiciel développé à l’Université Jean-Jaurès, Talismane[[8]](#footnote-8) par Assaf Urieli (Urieli, 2013).

Avoir le lemme d’un mot permet de rassembler toutes ses formes fléchies sous une même étiquette et de compter le nombre d’occurrences de celle-ci ensuite. La catégorie permet de percevoir la structure syntaxique qui est derrière les formes.

À chaque fois, nous avons conçu un script Python qui envoyait le titre brut à Stanford Core ou Talismane et récupérait le résultat du traitement. Nous n’avons pas comparé les résultats des deux en ce qui concerne les catégories pour tenter d’améliorer la fiabilité des résultats, cette question s’éloignant trop de nos priorités. Une fois les catégories et les lemmes obtenues, nous procédions à l’enregistrement de notre corpus, par un traitement rattaché à ceux que nous qualifions de conversions.

## B) Conversions

Tout au long de notre travail, il nous a fallu récupérer des données dans un format donné et les sauvegarder dans un autre. Le premier traitement de conversion a été la transformation des données au format JSON récupérées de HAL vers un format XML propre à notre travail. Le second s’occupait de la conversion des résultats obtenus auprès de Talismane puis de la sauvegarde des catégories et des lemmes en enrichissant notre format XML.

Nous n’avons pas utilisé des standards reconnus comme CoNLL-U[[9]](#footnote-9), issu de la conférence du même nom, ou TEI P5 de la communauté Text Encoding Initiative pour deux raisons. La première c’est que le contenu textuel d’un titre est très court mais nous en avons énormément. TEI P5 nous semble plus adapté pour encoder de véritables textes et CoNLL-U n’utilise pas XML mais un format texte utilisant les lignes et les tabulations pour traduire la structure des données. Nous souhaitions maîtriser notre format, sachant que celui-ci utilise XML, sa conversion vers un autre format basé sur XML ne poserait pas de problème.

Exemple de données récupérées auprès de HAL au format JSON ( { } indiquant un dictionnaire qui associe une clé à une valeur, [ ] indiquant une liste de valeurs et " " une chaîne de caractères) :

**{**

"docid": 1712921,

"domain\_s": **[** "0.shs", "1.shs.phil" **]**,

"title\_s": **[** "La logique de l'action de Michael Quante",

"Michael Quante on Logic and Action" **]**,

"authFullName\_s": **[** "Alain Patrick Olivier" **]**,

"language\_s": **[**"fr"**]**,

"docType\_s": "ART",

"modifiedDateY\_i": 2018

**}**

HAL nous donne une notice de document avec ses métadonnées. Idéalement, il y a une 1 notice pour 1 document qui possède 1 titre. Cette vision sera remise en cause plus fortement dans la partie suivante mais nous pouvons déjà voir que le champ titre correspond à une liste et que cette liste contient pour cette notice deux éléments. Le premier titre est en français et le second est sa traduction en anglais.

Malheureusement, il y a une certaine hétérogénéité dans le remplissage des différents champs mis à disposition par HAL. Pour certaines notices, le champ titre ne possède qu’un seul élément, mais il s’agit d’une concaténation du titre français avec le titre anglais, avec entre un marqueur qui n’est pas standardisé, certains utilisant « / » ou « [ » et d’autres « Titres en anglais : ». Cela nous amène à la nécessité de filtrage des données qui sera exposée dans le chapitre suivant.

Une fois transformé en XML et enrichi des catégories et lemmes, ce même titre se présente ainsi, les balises ouvrantes < > et fermantes </ > structurant les données :

<notice>

<id>1712921</id>

<type>ART</type>

<date>2018</date>

<text>La logique de l'action de Michael Quante</text>

<words>

<word><form>La</form><lemma>la</lemma><pos>DET</pos></word>

<word><form>logique</form><lemma>logique</lemma><pos>NC</pos></word>

<word><form>de</form><lemma>de</lemma><pos>P</pos></word>

<word><form>l'</form><lemma>le</lemma><pos>DET</pos></word>

<word><form>action</form><lemma>action</lemma><pos>NC</pos></word>

<word><form>de</form><lemma>de</lemma><pos>P</pos></word>

<word><form>Michael</form><lemma>Michael</lemma><pos>NPP</pos></word>

<word><form>Quante</form><lemma>\_</lemma><pos>NPP</pos></word>

</words>

<authors>

<author>Alain Patrick Olivier</author>

</authors>

<domains>

<domain>0.shs</domain>

<domain>1.shs.phil</domain>

</domains>

</notice>

Un titre a donc un identifiant correspondant à celui partagé par la notice et le document, un type correspondant à celui du document et une date. Cette date est issue de la dernière date de modification du dépôt du document. Nous avons pris cette date car c’était la seule systématiquement remplie sur tous les types de document de HAL de façon cohérente et qui indique la date de création du document scientifique.

Le titre en lui-même est présent sous une forme complète et une forme décomposée en mots, avec pour chaque mot la forme fléchie présente dans le texte, son lemme et son étiquette POS. Lorsque Talismane n’arrive pas à déterminer le lemme d’un mot, il indique ‘\_’ pour son lemme. La principale difficulté de cette étape est la gestion des caractères spéciaux dans les textes qui ont une signification spécifique pour XML comme « < », « > » ou « & ». Pour éviter cela, ils sont échappés, c’est-à-dire remplacés par un caractère neutre, lorsqu’ils sont sauvegardés sur le disque. Le caractère spécial est correctement restitué lorsque nous rechargeons en mémoire le corpus.

Enfin, nous avons la liste des auteurs et des disciplines scientifiques, appelées domaines dans HAL, auxquelles se rattache le document. On notera que les disciplines sont organisées en arbre, le chiffre avant son nom indiquant son niveau dans celui-ci, et qu’un même article peut être étiqueté sous plusieurs disciplines. Le titre présenté dans cette partie est néanmoins le fruit d’une autre catégorie de traitement : le filtrage effectué sur les données.

## C) Filtrage

Tout au long de notre travail, il a été nécessaire d’appliquer certains filtres à nos données. Tout d’abord, nous avons remarqué des incohérences dans nos données. Par exemple, certains documents avec plusieurs notices, nous avions donc des titres en double. D’autres, nous l’avons vu, concaténaient un titre français et un titre anglais, ou même étaient en anglais. Beaucoup avaient deux titres, un en français et sa traduction en anglais. Un autre cas, bien plus rare, était la présence d’un titre dans une autre langue.

Nous avons donc appliqué plusieurs filtres à nos données : en regardant le champ langue de la notice retournée par HAL, nous avons éliminé toute celle qui avaient plus d’un langage ou qui avait un langage qui n’était pas le français. Notre raisonnement était simple : si un document scientifique est en français, comme indiqué dans sa notice, son titre sera en français. Nous avons également gardé qu’un seul titre par notice s’il y en avait plusieurs, le premier, qui, d’après nos constations visuelles sur toutes les notices présentant cette caractéristique, était toujours celui en français.

En construisant un premier lexique des formes utilisés dans les titres nous avons pourtant remarqué une forte fréquence de « on », « and », « a », « in », « the », « und ». Les formes  « and », « in » et « the » appartiennent indiscutablement à l’anglais, tandis que « on » et « a » peuvent t’appartenir à l’anglais ou au français. Enfin « und » appartient lui à l’allemand. Pour mieux filtrer nos titres, utilisé un programme de détection automatique des langues écrits en Python appelé langdetect[[10]](#footnote-10) en gardant que les titres qu’il estimait être en français. Nous avons ainsi supprimé 12 205 titres.

Nous avons aussi supprimé certains titres car leurs notices nous semblaient incohérentes : 33 n’avaient pas d’auteurs, 6448 n’avaient pas de domaines associés, 1 n’avait de type de document et 7096 étaient des doublons. D’autres notices, au nombre de 11, avaient un titre vide et ont également été supprimées. À la fin, nous avions un corpus général de **278 806** titres, ce qui reste un nombre assez conséquent pour étudier un phénomène linguistique particulier dans celui-ci.

Enfin, nous avons appliqué la restriction découlant de notre problématique : nous voulions étudier les structures lexico-syntaxiques après un double point, il nous fallait donc extraire un sous-corpus, spécialisé pour notre problématique. Nous avons choisi de ne prendre que les titres contenant qu’un et un seul double point, considérant que les titres ayant plusieurs doubles points relevaient de cas très particuliers et complexes qui dépassent le cadre de notre travail. Un rapide décompte nous donne :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Nombre de « : »** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Nombre de titres | 190 123 | **86 095** | 2268 | 258 | 40 | 13 | 4 | 2 | 1 | 2 |

Tableau 1 : nombre de doubles points dans les titres

Il y a donc 86095 titres avec un et un seul double point. Ce tableau à l’avantage de nous montrer que des cas très particuliers existent, comme deux titres avec 9 doubles points dedans, mais qu’ils sont également très rares. En les écartant, il s’agit de nettoyer nos données car nous ne nous intéressons pas à ces cas à la marge. C’est le même souci qui nous amène à considérer le nombre de mots après l’unique double point de ces 86 095 titres pour obtenir le graphisme suivant :

On voit que 99% ont entre 0 et 29 mots après le double point. Nous écartons donc tous les titres en ayant plus (466 titres) ainsi que ceux en ayant 0 (98 titres). Une explication possible pour ces derniers était que le double point annonçait un sous-titre mais que nous ne l’avons pas récupéré de HAL. Avec le recul, nous aurions pu le concaténer mais encore une fois, il s’agit d’un traitement supplémentaire pour récupérer seulement 98 titres. Nous obtenons finalement notre corpus de travail de **85 531** titres, soit 31% des 278 806 titres du corpus général.

## II.3 Constations et exemples

Notons que, comme le souligne Cori et David (Cori & David, 2008), l’étape de sélection des données pour constituer un corpus peut comporter une part de subjectivité. Dans notre cas nous nous en exemptons car nous n’avons pas opéré pas de choix dans les titres que nous retournent HAL autres que ceux visant à assurer la cohérence de nos données. Pour ces choix, nous avons suivi la précaution méthodologique préconisée par ces deux auteurs, de faire *« un inventaire soigneux de toutes les décisions prises en amont »* que nous avons exposées dans ce document. Nous pouvons à présent observer notre corpus de travail sous plusieurs angles.

### II.3.1 Caractéristiques de nos corpus

#### Types de documents

Nous nous intéressons tout d’abord au type des documents scientifiques titrés. Nous indiquons seulement les 7 types ayant le plus de titres ce qui représentent 93% du corpus de travail.

|  |  |  |
| --- | --- | --- |
| **Type** | **Nombre** | **%** |
| Article | 25 648 | 30% |
| Communication | 19 966 | 23% |
| Chapitre | 12 007 | 14% |
| Thèse | 10 632 | 12% |
| Mémoire | 6 716 | 8% |
| Ouvrage | 2 281 | 3% |
| Autre publication | 2 147 | 3% |

On remarque que ces 7 premiers types de document sont des documents textes. Les vidéo et les sons ne représentent que 0,32% des titres récupérées.

#### Années des documents

Nous indiquons ici seulement les 7 années ayant le plus de titres ce qui représentent 99% du corpus de travail.

|  |  |  |
| --- | --- | --- |
| **Année** | **Nombre** | **%** |
| 2018 | 54 627 | 64% |
| 2017 | 21 658 | 25% |
| 2016 | 3 970 | 5% |
| 2015 | 1 996 | 2% |
| 2014 | 1 156 | 1% |
| 2013 | 817 | 1% |
| 2012 | 255 | 0% |

Sans surprise, les années les plus récentes sont les plus fournies. La 8e est 1988 avec 75 titres. Autant la diffusion sur HAL de nouveaux articles sert directement les chercheurs, et la pratique se généralise dans le monde de la recherche en France, autant la mise en ligne d’anciens articles d’eux-mêmes ou d’autres auteurs est une tâche longue et moins gratifiante. Ce déséquilibre ne nous permet pas d’étudier sur une large fenêtre en diachronie l’évolution des phénomènes autour du double point.

#### Longueurs des titres

La moyenne de la longueur en mots des titres est de 17,86 mots, pour des titres qui vont de 3 à 101 mots. La médiane est 16 mots, le premier quartile est 12, le dernier est 21.

#### Nombre d’auteurs

Nous regardons à présent le nombre d’auteurs par document scientifique.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Nombre**  **d’auteurs** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| Nombre de  titres (%) | 59 182  69% | 12 035  14% | 6 015  7% | 3 310  4% | 1 765  2% | 1 065  1% | 689  1% |

Si le nombre d’auteurs dans notre de corpus va de 1 à 147, 98% des titres ont néanmoins entre 1 et 7 auteurs et 69% ont un seul auteur, pour une moyenne de 1,8 auteurs par titre.

#### Domaines et nombre de domaines

Nos documents se répartissent en 13 domaines de premier niveau. Certains titres sont apparentés à plusieurs domaines.

|  |  |  |  |
| --- | --- | --- | --- |
| **Domaine** | **Nombre** | **Domaine** | **Nombre** |
| Sciences de l’Homme et Société | 80 441 | Sciences cognitives | 1 354 |
| Sciences du Vivant | 15 614 | Chimie | 1 051 |
| Informatique | 5 244 | Mathématiques | 877 |
| Sciences de l’ingénieur | 4 469 | Économie et finance quantitative | 283 |
| Sciences de l’environnement | 3 414 | Statistiques | 228 |
| Planète et Univers | 2 157 | Science non linéaire | 22 |
| Physique | 2 013 |  |  |

Si on fait une dichotomie entre les titres référençant les Sciences de l’Homme et Société et ceux ne le faisant pas, on prend mieux en compte le poids très important de ceux-ci dans notre corpus : 61 252 contre 24 279, soit 72% et 28% respectivement. Si l’on regarde les autres domaines, on voit que cette dichotomie reprend celle entre sciences « dures » et « soft ».

#### Lexique

TODO PPP

### II.3.2 Connexions avec les autres études

#### Taille des corpus

Notre corpus général se rapproche de celui utilisé par Lewison et Hartley (2005) qui comportait 349 700 titres, tous les autres corpus des articles ne dépassant pas les 2200 titres. Notre corpus spécialisé reste toujours bien au-dessus de ce seuil avec 84 293 titres.

#### Lien entre phrase complète et biologie

Haggan (Haggan, 2004) avait montré que l’utilisation d’une phrase complète pour titre était une caractéristique des titres en biologie. Nous avons voulu savoir ce qu’il en était sur notre corpus.

TODO

#### Longueur du titre et nombre d’auteurs

Lewison et Hartley (Lewison & Hartley, 2005) ont montré que plus il y a d’auteurs, plus le titre aura tendance à être long jusqu’à un plateau de 11 mots à partir de 4 auteurs.

TODO

#### Titres segmentés ou composés

On peut également évaluer le nombre et le type des partitions internes dans les titres. Le caractère segmentant peut-être un double point mais aussi un point, un point d’interrogation, un point d’exclamation ou d’autres signes de ponctuation repérés par Haggan (2004).

TODO

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type (traductions[[11]](#footnote-11))** | **Nb dans corpus** | **% dans corpus** | **Nb dans HAL** | **% dans HAL** |
| Article (Article dans une revue) | 45 355 | 30.94 % | 681 988 | 49.65 % |
| Report (Rapport) | 3 558 | 2.43 % | 25 144 | 1.83 % |
| Other (Autre publication) | 7 803 | 5.32 % | 31 462 | 2.29 % |
| ConferenceObject (Communication dans un congrès) | 35 860 | 24.46 % | 393 978 | 28.68 % |
| Preprint (Pré-publication, document de travail) | 2 878 | 1.96 % | 37 462 | 2.73 % |
| BookPart (Chapitre d’ouvrage) | 23 747 | 16.2 % | 105 565 | 7.69 % |
| Lecture (Cours) | 291 | 0.2 % | 1 102 | 0.08 % |
| Book (Livre) | 10 831 | 7.39 % | 24 705 | 1.8 % |
| Patent (Brevet) | 323 | 0.22 % | 3 179 | 0.23 % |
| DoctoralThesis (Thèse) | 15 954 | 10.88 % | 68 927 | 5.02 % |
| MasterThesis (Mémoire d’étudiant) | 2 | <0.1 % | 26 | <0.1 % |
| **Total** | 146 603 | 100 % | 1 373 538 | 100 % |

*Tableau 2 : Répartition des titres par type*

Nous avons choisi de mettre également les nombres pour l’ensemble de l’archive ouverte HAL. Cela dans un but d’essayer de mesurer la représentativité de notre corpus par rapport à l’ensemble des données. Un bon corpus se doit d’être représentatif ce qui est globalement le cas. Les livres sont néanmoins surreprésentés (~7% contre ~2%) dans notre corpus, les chapitres d’ouvrage (~16% contre ~8%) ainsi que les thèses (~11% contre ~5%). Les articles sont par contre sous représentés (~31% contre ~50%). Martin (2002) pose comme notion fondamentale qu’« en raison de sa finitude, le corpus ne réalise donc qu'une part infime de ce qui est réalisable. (…) Et en toute rigueur, une grammaire construite à partir d'un corpus ne vaut que pour le corpus qui l'a produite. » Ainsi les disparités que nous constations avec l’ensemble des données de HAL nous mettent en garde contre toutes généralisations hâtives des conclusions que nous pourrions découvrir sur notre corpus. De plus, si nous devions associer une caractéristique quelconque des titres avec une caractéristique extrinsèque particulière, comme le type de texte par exemple, nous construirions des corpus monotypes pour vérifier nos hypothèses.

Caractères spéciaux : Nous avons compté la présence de points d’interrogation, d’exclamation,

- Si la plage des années est assez grande (de 2018 à 1921), la surreprésentation de 2018 (63%) ne permettra pas d'études diachroniques représentatives.

- Si on réduit notre corpus aux articles, communications, chapitres d'ouvrage, thèse et mémoire, ont couvre 261 985 titres, soit 87% de notre corpus total. À voir si on le réduit à cela.

- Nous avons une moyenne pondérée de 87 caractères par titre.

- Nous avons une moyenne pondérée de 13 mots par titre.

- 86% des notices concernent la SHS

- 31% des titres comportent un « : »

- noms les plus fréquens : étude, analyse, France

Une caractéristique essentielle commune aux articles étudiés dans l’état de l’art est qu’ils ne portaient que sur des titres d’articles scientifiques en anglais, sauf un (Rebeyrolle, Jacques, & Péry-Woodley, 2009) qui portait sur des articles de la presse généraliste en français. Notre travail se distinguera donc déjà par cet aspect puisque notre matériel de départ est constitué de titres scientifiques en français. De plus, ils s’agissaient de titres d’articles de revues alors qu’HAL héberge d’autres types de travaux scientifiques.

Retrouve-t-on le fait que plus un titre a d’auteurs, plus son titre est long ?

Dimensions du corpus (tailles, …)

II.3.2 Longueurs des titres et présences des caractères segmentant

Une seconde tâche est de compter les mots, encore faudra-t-il décider si on les compte tous : Jamali et Nikzad (2011) ne comptent par exemple que les substantifs.

Nous avons déjà donné quelques coups de sonde dans notre corpus sur certaines caractéristiques dont nous présentons les résultats dans le tableau 7. Une remarque importante est de pas confondre le point final, d’interrogation ou d’exclamation à l’intérieur d’un titre, qui le partitionne, de celui pouvant éventuellement le terminer.

|  |  |  |
| --- | --- | --- |
| **Caractéristique** | **Valeur** | |
| Nombre de titres | 146 603 | |
| Estimation de la longueur moyenne | 13 mots / titre | |
| **Nombre de titres avec au moins un…** |  | Dont en dernière position |
| Double point | 53 169 (36,27%) | 0 (0%) |
| Point | 26 121 (17.82 %) | 7 230 (27.68%) |
| Point d’interrogation | 11 949 (8,15%) | 7 064 (59 .12%) |
| Guillemets : «, ", » | « 6902  » 6939[[12]](#footnote-12)  " 5813 |  |
| Point-virgule | 1 256 (0.86 %) | 5 (0.4%) |
| Point d’exclamation | 671 (0,46%) | 264 (39.34%) |

*Tableau 3 : Présence de quelques caractères segmentant dans notre corpus*

Il faut lire ce tableau ainsi : il y a 11 949 titres avec au moins un point d’interrogation. Parmi ceux-ci, il y en a 7 064, environ 60%, qui l’utilise comme symbole terminal. Les plus intéressants pour l’étude des partitions sont bien sûr les 40% restant qui utilisent le point d’interrogation à l’intérieur du titre, comme caractère segmentant.

On voit que la structure qui nous intéresse, utilisant un double point, est très bien représentée dans ce corpus : plus d’un tiers des titres de celui-ci en comporte un. De plus, il n’est jamais utilisé comme symbole terminal, ce qui semble logique. Nous regarderons lexicalement les mots situés immédiatement à droite du deux points, en faisant tout d’abord une approche lexicométrique.

II.3.3 Le lemme d’après

Nous avons voulu observer quelles formes et quels lemmes nous trouvions à proximité du double point sur sa droite. Nous avons amélioré notre script pour en obtenir une première vision, avec une fenêtre de 3 formes à droite. Bien sûr, les articles sont parmi les plus présents, ainsi que des coordinations et des prépositions. Si on les élimine pour ne garder que les 15 substantifs les plus employés, on obtient le tableau 8. Nous avons additionné les différentes formes d’un même lemme :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Forme** | **Compte** | **Forme** | **Compte** | **Forme** | **Compte** |
| Cas (m) | 1758 | analyse (f) | 838 | état (m) | 283 |
| étude (f) | 1210 | enjeu (m) | 669 | effet (m) | 269 |
| application (f) | 1087 | apport (m) | 404 | réflexion (f) | 256 |
| exemple (m) | 1019 | histoire (f) | 345 | modèle (m) | 255 |
| approche (f) | 1006 | rôle (m) | 321 | outil (m) | 251 |

*Tableau 4 : lemmes les plus présents après un double point*

Nous avons donc effectué une cooccurrence non pas entre deux lemmes mais entre une marque de ponctuation et un lemme. Nous avons effectué la lemmatisation à l’aide de règles simples, les suppressions des -s et des -x finaux, et sélectionné les substantifs à la main. Nous envisageons éventuellement dans une itération ultérieure d’utiliser des ressources supplémentaires comme un lexique morphologique et syntaxique tel que le Lexique des Formes Fléchies du Français (Sagot, 2010). Le LEFFF a été développé au sein de l’Institut National de Recherche en Informatique et en Automatique (INRIA) puis du Laboratoire Bordelais de Recherche en Informatique (LARBI) depuis 2003 par Lionel Clément et Benoît Sagot. En accès libre et gratuit, il donne, pour de nombreuses formes, son lemme de rattachement et la catégorie syntaxique de celui-ci. De plus, nous restons pour le moment en surface, sans chercher à comprendre syntaxiquement le syntagme nominal à droite du double point en construisant un arbre de constituants ou un graphe de dépendances. Pour cela, nous devrions faire appel à des outils plus puissants.

À ce tableau, nous devons ajouter le cas de *nouvelle* qui compte 348 occurrences sous les formes *nouvelle* et *nouvelles*. Notre algorithme ne distinguant par les catégories syntaxiques, il nous est difficile de savoir s’il s’agit du lemme nouvelle : nom commun ou du lemme nouveau, nouvel, nouvelle : adjectif. Pour comprendre cela, nous avons cherché à visualiser ses cooccurrences. Après les avoir observées, nous pouvons affirmer que l’extrême majorité des 348 occurrences concernent l’adjectif.

Sémantiquement, les mots retournés sont très intéressants : on pourrait les qualifier intuitivement d’objets de la recherche scientifique. Nous colorons ceux se rattachant à ce trait en vert. Les 4 derniers en orange, *enjeu, histoire, rôle* et *état* peuvent se rattacher à certaines disciplines ou à notre premier trait selon le contexte. Nous proposons à présent quelques exemples de titres.

### II.3.3 Quelques exemples de titres

Nous avons ici sélectionné quelques exemples de titre tirés de notre corpus correspondant à des réalisations de la structure lexico-syntaxique qui nous intéresse :

* **Loches. Première campagne de fouilles sur la forteresse : Rapport de sondages archéologiques** [Ligne 24082, Identifiant hal-01269678, Archéologie]

Ce titre est un exemple quasi-prototypique de la structure qui nous intéresse. Son seul écart par rapport au prototype est d’avoir trois partitions au lieu de deux. Le double point est sinon immédiatement suivi d’un seul syntagme nominal dont le noyau est le mot *rapport* qui indique la nature du document. On note également qu’il y a une autre partition au début, qui précise un lieu géographique.

* **Le promontoire du "Malpas" à Soyons (Ardèche) : un habitat fortifié à l'âge du Fer sur les rives du Rhône** [Ligne 95974, Identifiant halshs-01359782, Archéologie]

Ce titre utilise encore une fois la structure qui nous intéresse. Il comporte deux partitions. Cette fois-ci, le noyau du syntagme nominal à la droite du double point n’apporte pas une précision sur la nature du document mais sur son sujet. Notons également l’utilisation des guillemets pour encadrer une indication géographique. Leur utilisation est désapprouvée par Aleixandre-Benavent et al. (2014).

* **Qualité et consommateurs : nouvelles approches de la segmentation** [Ligne XXX, Identifiant hal-01197932, Économie]

Ce titre est un exemple prototypique de notre patron. On remarque une utilisation de l’adjectif *nouveau, nouvel, nouvelle* qui qualifie le noyau du syntagme nominal, le lemme *approche* avec un complément.

* **Dynamique des structures : méthodes approchées, cinématiques ; Analyse Modale ; Recalage de Modèle** [Ligne 32, Identifiant cel-00611705]

Ce titre est intéressant car il fait survenir un nouvel élément : l’utilisation de points-virgules. Ce titre est composé de 4 partitions, ce qui très important. Il est intéressant car on peut estimer qu’il y a une hiérarchie des symboles de partition. Intuitivement, le point-virgule semble établir la partition la plus forte, le double point crée une partition moyenne et enfin la virgule qui ponctue une énumération et non une partition.



La méthode présentée pour obtenir notre corpus dans cette partie est reproductible et permet d’obtenir d’en obtenir de nouveaux à partir de HAL. Le corpus que nous avons utilisé présente l’avantage d’être de grande taille et de présenter une grande variété de titres. Cette taille permet d’étudier un phénomène linguistique particulier, comme l’utilisation du double point, sur un nombre important de titres. Une fois notre corpus de travail prêt, nous pouvons tenter de répondre à notre problématique en le scrutant.

# III. Syntagmes et patrons

## III.1 Rappel sur les syntagmes

Le syntagme est selon (Maingueneau, Chiss, & Filliolet, 2007, p. 35), dans le cadre de la linguistique structurale, « un groupe d’éléments linguistiques [consécutifs] formant une unité [syntaxique] dans une organisation hiérarchisée », centré sur un élément principal. La catégorie de cet élément principal donne le type du syntagme. Concrètement, un syntagme est représenté dans notre travail par une suite de catégories comme « DET NC P DET NC » pour « la maison de la forêt ». On remarque que ce syntagme est lui-même organisé en sous-syntagmes : la linéarité de la suite ne doit pas faire oublier l’organisation hiérarchique des constituants du syntagme. (Maingueneau, Chiss, & Filliolet, 2007, p. 119) indique qu’un consensus existe pour les représenter en utilisant des graphes arborescents, comme pour notre exemple (en bleu les syntagmes, en orangé les éléments centraux) :



Figure 1 : arbre d'analyse

Pour étudier notre corpus, nous avons inventorié toutes les suites existantes de catégories après le double point. Elles sont au nombre de 45 098. Nous les avons observés sous deux angles : leurs constitutions et leurs nombres d’occurrences.

## III.2 Le premier syntagme après le double point

Pour réduire la portée de notre étude aux dimensions de cet exercice, nous nous limitons à étudier le syntagme qui suit immédiatement le double point dans un titre.

En parcourant notre inventaire, nous avons eu trois intuitions. La première est que le syntagme nominal est le plus représenté. La deuxième est que ce syntagme nominal inclus souvent un syntagme prépositionnel, un complément du nom, qui contient lui aussi un nom. La troisième est que ce syntagme complexe est représenté par une grande variété de suites :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DET | NC | P | DET | NC |  |
| DET | NC | P | NC |  |  |
| DET | NC | P+D | NC |  |  |
| NC | P | NC |  |  |  |
| NC | P | DET | NC |  |  |
| DET | NC | P | DET | NC | ADJ |

Tableau 5: exemples de suites de catégories correspondant à un syntagme nominal après le double point

Pour capturer toute ces variantes, nous avons besoin d’un autre outil conceptuel : le patron.

## III.3 Définition des patrons

Nous définissons le patron comme un ensemble de possibilités de catégories consécutives. Certaines catégories peuvent ne pas apparaître, d’autres peuvent être répétées, enfin certaines doivent être choisies entre plusieurs alternatives. Par exemple, dans un syntagme nominal, un déterminant initial est optionnel ainsi que des adjectifs qualifiants le nom. Le seul élément qui sera obligatoirement présent est l’élément central du syntagme, le nom.

Pour représenter nos patrons, nous utilisons un langage spécifique très simple qui tient en 5 principes :

1. A B signifie une catégorie A suivi d’une catégorie B
2. [ A B ] signifie un choix : soit la catégorie A, soit la catégorie B
3. A? signifie l’optionalité : la catégorie A peut apparaître mais ce n’est pas obligé
4. ( A B ) permet de grouper les catégories
   1. à l’intérieur d’un choix : [ (A B) C ] signifie soit A suivi de B, soit C
   2. ou pour signifier qu’une suite entière est optionnelle ( A B )? .
5. La répétition est représentée par le fait de répéter plusieurs fois la même suite d’éléments : A A . Notre langage ne permet pas de représenter la répétition infinie d’une catégorie.

Nous pouvons donc écrire un patron à l’aide de ce langage qui correspond aux différentes écritures possibles d’un syntagme. Par exemple, le patron qui correspond à la fois à notre exemple dans la Figure 1 : arbre d'analyse et ceux dans le Tableau 5: exemples de suites de catégories correspondant à un syntagme nominal après le double point peut être écrit ainsi :

DET? NC [ (P DET?) P+D ] NC ADJ?

Ce patron correspond aux exemples proposés précédemment, mais il correspond également à d’autres constructions possibles comme « DET NC P+D NC ADJ ». Éventuellement, ce patron pourrait correspondre à des suites qui ne sont pas représentées dans notre corpus. Ce n’est pas un problème car notre but n’est pas d’avoir un patron générant toutes les suites de notre corpus, mais de capturer toutes celles qui s’y trouvent et y correspondent.

## III.4 Construction des trois patrons

Nous avons dû arbitrer entre complexité et faisabilité : plutôt que de reconstruire l’ensemble de l’arbre syntaxique après le double point, nous nous contentons de regarder le premier syntagme et dans celui de se limiter arbitrairement dans son analyse.

Les trois types de syntagmes que nous avons décidé d’étudier sont : le syntagme nominal, le syntagme prépositionnel et le syntagme verbal. À chaque fois, nous avons limité la complexité de ceux-ci en interrompant notre exploration. Pour le syntagme nominal par exemple, nous demandons la présence d’un sous syntagme prépositionnel mais nous ne regardons pas s’il y en a plus d’un ou ce qui suit.

Pour obtenir nos trois patrons correspondants, nous avons choisi une méthode itérative en se basant sur l’observation de l’inventaire, la connaissance des règles de syntaxe et un script Python. Celui effectuait deux opérations :

1. Le comptage automatique de la couverture du patron. La couverture de patron est le nombre de suites ou de titres auxquels il correspond, par rapport au nombre total de suites inventoriées (45 098) et de titres dans notre corpus (85 531) respectivement. Le plus important est la couverture des titres, car certaines suites sont très peu utilisées.
2. La séparation en deux des suites : d’un côté celles qui s’accordaient avec notre patron, de l’autre, celles qui ne s’accordaient pas. En regardant attentivement ces dernières, on peut décider alors d’augmenter la variabilité de notre patron pour qu’il couvre plus de suites.

Nous sommes partis à chaque fois d’un patron minimaliste, par exemple NC P NC, avant de rajouter les différents éléments optionnels puis les choix et les répétitions possibles pour obtenir trois patrons ayant une couverture maximale sans dénaturer la nature du syntagme capturé par le patron. De ce fait, les trois patrons sont mutuellement exclusifs : une suite ne peut s’accorder qu’avec un seul d’entre eux ou aucun.

Dans la partie suivante, nous présentons les 3 patrons ainsi que l’étude de différents aspects.

# IV. Études des trois patrons

## IV.1 Patron n°1 : syntagme nominal

### IV.1.A Fiche d’identité

**Description** Il s’agit d’un syntagme nominal incluant un syntagme prépositionnel.

**Patron minimal** NC P NC

**Version étendue**

DET? ADJ? [NC NPP] [NC NPP]? ADJ? [(P DET?) P+D] ADJ? [NC NPP] [NC NPP]? ADJ?

## IV.2 Patron n°2 : syntagme prépositionnel

### IV.2.A Fiche d’identité

Description : il s’agit d’un syntagme prépositionnel incluant un syntagme nominal qui lui-même inclut un syntagme

|  |  |  |  |
| --- | --- | --- | --- |
| **Patron minimal** | **Version étendue** | **Couv. suites** | **Couv. Titres** |
| NC P NC | DET? ADJ? [NC NPP] [NC NPP]? ADJ? [(P DET?) P+D] ADJ? [NC NPP] [NC NPP]? ADJ? |  |  |
| P NC P NC |  |  |  |
|  |  |  |  |

PPP

## IV.3 Patron n°3 : syntagme verbal

### IV.2.A Fiche d’identité

### IV.2.B Exemples

### IV.2.C Statistiques

Couverture (du corpus, du titre)

Fréquences

Répartition par domaine

Les trois patrons minimaux que nous avons choisis sont dans le tableau ci-dessous ainsi que leurs versions étendues et leurs couvertures des suites et des titres.

Syntagme nominal, prépositionnel et verbal

Pas d’analyse syntaxique complète du titre

Les travaux de Quiniou et al. (2012) nous mettent sur la piste de chercher des motifs émergents de syntagmes nominaux. Intuitivement, on peut penser qu’il se caractérise d’un article (optionnel), d’une suite d’adjectifs, éventuellement modifiés par un adverbe, du substantif noyau et d’adjectifs postposés, éventuellement complémentés, ce que ne sont jamais les adjectifs postposés.

Le premier point fait que nous ne saurons pas si le syntagme nominal est le sujet d’une phrase complète ou pas.

La poursuite de notre travail doit également permettre d’approfondir complètement notre connaissance du patron que nous avons choisi et ses réalisations. Notre chaîne de traitement doit faire émerger les variations du patron « … : SN … ». Pour cela, nous devons détecter à droite le syntagme nominal et ses membres. Éviter de faire toute l’analyse syntaxique du titre serait un grand avantage. Nous savons le caractère délimitant à gauche le syntagme nominal : un double point. Il nous faut à présent réfléchir à comment le délimiter sur la droite. Mais il serait plus intelligent de s’arrêter dès le noyau trouvé. Des sous-patrons pourraient alors émerger.

# V. Étude de structures particulières

Dans la partie précédente, nous avons étudié uniquement les catégories et nos patrons ne portaient uniquement sur celles-ci. À présent, nous voulons étudier les lemmes qui se cachent derrière les catégories.

PPP

# VI. Résultats et discussions

## VI.1 Résultats

## VI.2 Discussions

Parler du fait qu’on ne choisit pas l’émergence auto

Dire que notre langage ne gère pas l’infinité et ce que cela fait

VI.1.1 Défauts remarqués

Talismane : ADJ pris comme NC : il cherche des phrases, spécificités des titres

Il est rare qu’un titre forme une phrase verbale. De plus, un titre est souvent très segmenté par un double point, des virgules voir même des points et ces partitions sont autant de phrases incomplètes. TreeTagger (Schmid, 2013) et la plupart des logiciels de lemmatisation et de catégorisation grammaticale sont parfois perplexes pour analyser de telles phrases car ils ont été plutôt pensés pour le contenu des textes et non leur titre. Leech (2000) propose la notion de C-units, des petites unités grammaticales indépendantes très présentes à l’oral, pour aborder ces constructions. Dans notre cas, l’outillage de notre mémoire se fera au cours du second semestre et nous espérons ne pas avoir à reprogrammer un tel outil adapté aux spécificités des titres.

Outil propre : « étude de cas » vu comme 3 suites N1 P N2 différentes,

Heureusement, 2 sont très marginales.

4.1.2 Ouverture sémantique

*Sur le plan sémantique, que peut-on dire à ce stade ?*

5.1 Perspectives

5.1.1 Couverture totale du corpus

5.1.2 Couverture totale des titres

5.1.3 Émergence automatique de patrons

# Conclusion

Perspectives

# Bibliographie

Bray, T. (2017). *The JavaScript Object Notation (JSON) Data Interchange Format*. Retrieved from IETF Tools: https://tools.ietf.org/html/rfc8259

Cori, M., & David, S. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages, (3)*, 111-129.

Haggan, M. (2004). Research paper titles in literature, linguistics and science: dimensions of attraction. *Journal of Pragmatics, 2.36*, 293-317.

Lewison, G., & Hartley, J. (2005). What's in a title? Numbers of words and the presence of colons. *Scientometrics, 63 (2)*, 341-356.

Maingueneau, D., Chiss, J.-L., & Filliolet, J. (2007). *Introduction à la linguistique française.* Hachette Éducation.

Nivard, J. (2010). *Les Archives ouvertes de l’EHESS*. Retrieved from La Lettre de l'École des hautes études en sciences sociales, n°34: http://lettre.ehess.fr/index.php?5883

Rebeyrolle, J., Jacques, M.-P., & Péry-Woodley, M.-P. (2009). Titres et intertitres dans l'organisation du discours. *Journal of French Language Studies, 19*, 269-290.

Smiley, D., Pugh, E., Parisa, K., & Mitchell, M. (2015). *Apache Solr enterprise search server.* Birmingham: Packt Publishing Ltd.

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.

# Annexes

## A1. Requêtes Apache Solr sur HAL

### A1.A Requêtes

Nous proposons d’explorer un exemple de requête avec la plate-forme Apache Solr (Requête 1) que nous formatons afin de le rendre plus visible. Nous demandons les informations supplémentaires de la discipline (domain\_s), des auteurs (authFullName\_s), du type de document (docType\_s), la date de modification (modifiedDateY\_i) et bien sûr le titre (title\_s). Nous classons en fonction de la date de modification, de la plus récente à la plus vieille. Par défaut, 30 résultats sont retournés, le maximum autorisé étant de 10 000, nous en demandons ici 1000.

https://api.archives-ouvertes.fr/search/ ?  
 wt = json &   
 fl = docid, domain\_s, authFullName\_s,   
 docType\_s, title\_s, modifiedDateY\_i &   
 indent = true &  
 sort = modifiedDateY\_i desc &  
 rows = 1000

*Requête 2 : Un exemple de requête avec l’API Apache Solr formaté pour plus de lisibilité*

Un système de cache permet de récupérer bien plus de résultats que la limite de 10 000 par requête. En triant sur l’identifiant numérique des notices, le champs docid , une clé unique car aucune notice ne possède la même valeur, on peut demander à Solr un nombre de résultats, disons X, et la création d’un cache. La première réponse de X résultats comporte à la fin un identifiant. En relançant la même requête et en remplaçant la demande de création d’un cache par l’identifiant fourni, on obtient les X résultats *suivants*. Les deux requêtes suivantes (Requêtes 2a et 2b), la première de création, la seconde de poursuite de la demande avec l’identifiant fourni par la première requête, AoFVmLIG , illustrent cette puissante fonctionnalité.

https://api.archives-ouvertes.fr/search/?wt=json&fl=docid,domain\_s,authFullName\_s,docType\_s,title\_s,modifiedDateY\_i&indent=true&sort=docid%20desc&rows=1000&cursorMark=\*

https://api.archives-ouvertes.fr/search/?wt=json&fl=docid,domain\_s,authFullName\_s,docType\_s,title\_s,modifiedDateY\_i&indent=true&sort=docid%20desc&rows=1000&cursorMark=AoFVmLIG

*Requêtes 3a et 2b : création et consultation d’un cache*

Nous pouvons à présent nous pencher sur les résultats de ces requêtes qui vont constituer notre corpus.

### A1.B Résultats

Notons que, comme le souligne Cori et David (2008), l’étape de sélection des données pour constituer un corpus peut comporter une part de subjectivité. Dans notre cas nous nous en exemptons car nous n’opérons pas de choix dans les titres que nous retournent HAL. Si nous devions préparer plus avant nos données, en particulier en écartant certains titres pour des raisons d’invalidité technique ou de doublon, nous respecterions la précaution méthodologique que préconisent ces deux auteurs, faire *« un inventaire soigneux de toutes les décisions prises en amont »* L’encadré Résultat 3 présente un élément tiré d’un exemple de résultat de requête Apache Solr.

{

"docid" : 1675646,

"domain\_s" : [

"0.scco", "1.scco.ling", "0.scco", "1.scco.psyc"],

"title\_s" : [

"Récits d’enfants et d’adolescents - Développements typiques, atypiques, dysfonctionnements"],

"authFullName\_s" : [

"Christiane Préneron", "Claire Martinot"],

"docType\_s" : "DOUV",

"modifiedDateY\_i" : 2018

},

*Résultat 4 : exemple d’un élément de résultats au format JSON*

Un premier jeu de données, retravaillé à partir de requêtes à HAL pour être mis dans un format tabulaire, nous a été fourni par M. Tanguy le 20 octobre 2017. Il s’agit de notre premier corpus de travail. Il comporte 146 603 titres, accompagnées d’informations supplémentaires que nous détaillons dans le paragraphe suivant. Sa taille le rapproche bien plus de celui utilisé par Lewison et Hartley (2005) que de tous les autres, même si ce dernier reste plus de deux fois plus volumineux. C’est ce premier jeu qui nous a servi pour nos réflexions et forger notre problématique.

Les titres nous parviennent accompagnés des autres métadonnées de la notice. Pour simplifier, nous considérerons que les métadonnées de la notice sont également des métadonnées du titre. Elles ne sont pas incluses dans le contenu du titre lui-même, on parlera de caractéristiques extrinsèques, mais elles sont très intéressantes pour nous aider à mieux comprendre le titre. Elles sont : l’identifiant numérique de la notice, le type du texte titré, son année d’enregistrement sur HAL, indispensable pour des études en diachronie, le nombre d’auteurs, son champ et ses disciplines. On notera que les disciplines peuvent être hiérarchisées en arbre et qu’un même article peut être étiqueté sous plusieurs disciplines. Le champ étant une « discipline racine » ou « super discipline ».

Ces métadonnées pourront ensuite être mises en relation avec une caractérise intrinsèque du titre. Des auteurs avaient déjà mis en avant certaines régularités. Haggan (2004) avait par exemple montré que l’utilisation d’une phrase complète pour titre était une caractéristique des titres en biologie. Lewison et Hartley (2005) ont montré que plus il y a d’auteurs, plus le titre aura tendance à être long jusqu’à un plateau de 11 mots à partir de 4 auteurs. En guise d’exemple, Résultat 4 présente une ligne de notre corpus avec le titre et ses métadonnées.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Titre** | **Type** | **Année** | **Nauteurs** | **Champ** | **Disciplines** |
| artxibo-01200715 | Deux dichotomies de la langue basque | other | 2014 | 1 | SHS | SHS.LANGUE.SOCIO |

*Résultat 5 : une ligne de notre premier corpus de travail*

Après cette présentation de nos données de travail, nous prouvons entrer dans le vif de celui-ci en dévoilant nos premières analyses de ces données.

## AX. Index des logiciels et technologies mentionnés

Apache Lucene 8

Apache Solr 8

CoNLL-U 9

CSV 8

Excel 6

HTTP 8

JSON 8

langdetect 12

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) 8

protocole de transfert hypertexte *HTTP*

Python 6

Stanford Core Natural Language Processing 9

Talismane 9

XML 8

Table des requêtes et résultats

Table des tableaux

Ensemble de fichier

Corpus XML 450 Mo 40 Mo (zippé)

Spécification de mon format ?

|  |  |  |
| --- | --- | --- |
| **Nom attribut** | **Type Attribut** | **Signification** |
| title | Chaîne de caractères | Le titre en lui-même |
| words | Liste de chaîne de caractères | Liste des mots |
| pos\_tags | Liste de chaîne de caractères fixées par Stanford Core NLP | Liste des étiquettes de catégorie du discours |
| authors | Liste de chaîne de caractères | Liste des auteurs |
| docid | Entier | Identifiant de la notice |
| kind | Chaîne de caractères fixées par HAL | Type du document référencé |
| date | Chaîne de caractères | Année de dernière modification |

*Les attributs de la structure de données « Title » en mémoire*

A.1 Le dépôt

Tout est le code Python est public et stocké sur la plateforme GitHub à cette adresse :

https://github.com/Xitog/tal

Il est composé de 3 fichiers qui correspondent chacun à une étape :

- corpus1.py : récupération des données brutes de HAL (correspond au chapitre I.1)

- corpus2.py : traitement des données brutes pour obtenir le corpus de travail (correspond aux chapitres I.2 à I.6).

- corpus3.py : traitement du corpus de travail (correspond au chapitre II).

corpus3.py utilise certains éléments de corpus2.py.

Le code n’a pas été contrôlé par un outil de mesure de qualité automatique comme PyLint.

https://www.pylint.org/

Ce dépôt contient également le corpus de travail au format zip :

https://github.com/Xitog/tal/raw/master/output\_dump\_repo/dump.zip

Ainsi que le lexique obtenu à partir des données brutes :

https://github.com/Xitog/tal/blob/master/output\_lexicon/lexicon.xlsx

Et les différents documents de présentation et liaison.

A.2 Les bibliothèques

Nous utilisons la bibliothèque openpyxl pour la gestion du format Excel XLSX. Celui comporte l’avantage par rapport au format Excel XLS de pouvoir dépasser plus de 65 535 lignes par onglet.

https://openpyxl.readthedocs.io/en/stable/

Elle est plus puissante que la première que j’avais utilisée, l’ensemble XLRD et XLWT qui gère le format XLS. Mais la documentation d’openpyxl est néanmoins plus complexe.

Pour le chargement progressif du XML, j’utilise la bibliothèque standard, plus particulièrement :

xml.etree.ElementTree

Pour la connexion avec Stanford Core NLP, j’utilise cette bibliothèque, qui permet depuis Python de choisir la langue cible ainsi que la quantité de mémoire :

https://github.com/Lynten/stanford-corenlp

Le site officiel de Stanford Core NLP est : https://stanfordnlp.github.io/CoreNLP/

La bibliothèque langdetect a également été utilisé pour détecter la langue des titres suspectés de n’être pas écrits en français.

1. <https://www.python.org/> [↑](#footnote-ref-1)
2. <https://products.office.com/fr-fr/excel> [↑](#footnote-ref-2)
3. <https://hal.archives-ouvertes.fr/> [↑](#footnote-ref-3)
4. <https://www.ccsd.cnrs.fr/> [↑](#footnote-ref-4)
5. <http://arxiv.org/> [↑](#footnote-ref-5)
6. <https://lucene.apache.org/> [↑](#footnote-ref-6)
7. <https://stanfordnlp.github.io/CoreNLP/> [↑](#footnote-ref-7)
8. <http://redac.univ-tlse2.fr/applications/talismane/talismane.html> [↑](#footnote-ref-8)
9. <http://universaldependencies.org/docs/format.html> [↑](#footnote-ref-9)
10. <https://pypi.org/project/langdetect/> [↑](#footnote-ref-10)
11. Les traductions et les valeurs ont été prises depuis cette page, consultée le 21 janvier 2018 : <https://hal.archives-ouvertes.fr/browse/doctype> Nous avons ajouté documents et notices pour faire un total. [↑](#footnote-ref-11)
12. Les guillemets français « » présente l’avantage d’être différenciés entre l’ouvrant et le fermant. **Nous remarquons que 293 titres, soit 0,19% ont des problèmes de cohérence : présence d’un ouvrant sans le fermant ou présence d’un fermant sans l’ouvrant.** Le nombre de titres concernés est néanmoins très faible. [↑](#footnote-ref-12)