Les titres des publications scientifiques en français : Fouille de texte pour le repérage de schémas lexico-syntaxiques

Ludovic TANGUY, Josette REBEYROLLE

Résumé
Nous présentons dans cet article une première étude basée sur corpus visant à établir un panorama des structures que les auteurs d'articles scientifiques en français emploient pour construire les titres de leurs travaux. Nous nous basons sur un ensemble de 340 000 titres (articles de journaux, actes de conférences, chapitres d'ouvrages) extraits de l'archive ouverte institutionnelle HAL et correspondant à l'ensemble des domaines disponibles. Nous proposons une méthode automatique inductive de fouille de texte qui permet de dégager les schémas les plus productifs à différents niveaux de détails (en choisissant de faire apparaître ou non les éléments lexicaux) comme par exemple « la place de X dans X » ou « X : quel X pour X ? ». Le croisement de ces schémas avec les domaines nous permet, dans un second temps, de mettre au jour des configurations contrastées et propres aux disciplines. Nous montrons également comment des méthodes plus ciblées d'interrogation de corpus permettent d'identifier des familles de titres comme les chiasmes.

Mots-clés
Titre, fouille de texte, discours scientifique, schémas lexico-syntaxiques, phraséologie.

Abstract
In this paper we study the titles of academic articles in French, and propose an overview of their syntactic structures. We automated the extraction from the HAL institutional open archive and compiled a corpus of more than 340,000 titles of articles, proceedings and chapters from different academic disciplines. We propose an inductive text mining method that allows us to identify the most productive title structures with varying levels of details (by choosing to mask lexical items or not) such as "la place de X dans X" (*The place of X in X*) or "X: quel X pour X ?" (*X: which X for X?*). We study the distribution of these structures across disciplines and identify several domain-specific title schemes. We also demonstrate how more focused queries can be run on our corpus in order to extract and analyze titles with more specific linguistic phenomena, such as chiasmus.

Keywords
Titles, text mining, academic papers, syntactic patterns, phraseology.

Introduction

Le présent article a pour objectif de rendre compte d'une première expérimentation fondée sur un vaste corpus de titres de publications scientifiques en français. Le choix de cet objet textuel s'inscrit à la suite d'un ensemble de travaux antérieurs qui portaient sur d'autres types de titres, comme les titres d'articles de presse mais surtout les titres de section, qui structurent les documents longs (*cf*. Ho-Dac et al., 2004, Rebeyrolle et al., 2009). Les titres de publications scientifiques ont retenu notre attention pour deux raisons. D'abord, il est extrêmement facile aujourd'hui de constituer rapidement un très large corpus de titres de ce type. Dans un contexte où la diffusion numérique de l'information connait peu de freins, les titres d'articles scientifiques présentent également l'intérêt de constituer des données langagières dont l'usage comme matériau pour la recherche n'est restreint par aucune disposition légale. La seconde raison est que les travaux qui ont été consacrés à ce type de titre portent exclusivement sur l'anglais. Dans l'importante bibliographie, on retiendra notamment Bush-Lauer (2000), Haggan (2004) et, plus récemment, Soler (2011), dont les travaux montrent notamment que la forme des titres de publications scientifiques varie selon les disciplines.

Notre principal objectif, dans cet article, est de mettre au jour des schémas récurrents de titres d'articles scientifiques. Ce faisant, nous inscrivons notre travail dans une conception étendue de la phraséologie, telle que la décrivent Legallois et Tutin (2013). La méthode que nous déployons est inspirée des techniques inductives de fouille de texte et d'identification de motifs (Quiniou et al. 2012). Nous aboutissons à un premier inventaire, à large couverture, des structures lexico-syntaxiques les plus productives : cet inventaire est un premier pas vers une caractérisation des titres dans les discours scientifiques. Notre étude vise également à vérifier, pour le français, la variabilité inter-domaniale déjà mise en évidence dans les travaux portant sur l'anglais.

Cet article se veut avant tout méthodologique : nous y présentons à la fois les détails de l'approche inductive que nous avons appliquée pour faire émerger des schémas génériques, mais aussi un exemple d'investigation plus ciblée portant sur un phénomène spécifique apparaissant dans les titres : le chiasme.

1. Constitution du corpus de titres

1.1. Recueil des données

Les notices bibliographiques des écrits académiques sont désormais facilement accessibles grâce aux nombreux efforts déployés par les auteurs et les éditeurs pour diffuser les productions scientifiques. Ces efforts se manifestent notamment par la mise à disposition de bases de données bibliographiques des grands éditeurs, de sites de prépublications, de catalogues des bibliothèques nationales ou encore par les archives ouvertes. Ces dernières, qui s'inscrivent dans un ensemble de projets internationaux de diffusion libre du savoir, présentent l'avantage de proposer un accès simple et efficace à l'intégralité de leurs bases de données.

Notre objectif étant ici de travailler sur les titres de publications scientifiques en français, nous nous sommes tournés vers l'archive ouverte la plus connue : HAL pour Hyper Article en Ligne (https://hal.archives-ouvertes.fr/). Cette base de données institutionnelle, mise à disposition et gérée par le Centre pour la Communication Scientifique Directe du CNRS, propose l'archivage et la diffusion de notices bibliographiques et de documents mis à disposition par les auteurs eux-mêmes. Ce

service est opérationnel depuis 2001 et, au moment où nous écrivons cet article, la base de données contient plus de 1,6 million de références dans les différentes disciplines académiques. Les notices correspondantes sont saisies par le biais de formulaires contrôlés et vérifiés avant leur dépôt, ce qui garantit un minimum d'intégrité des données (absence de doublons et de notices incomplètes, notamment). Dans la mesure où la fonction principale de HAL est de diffuser les productions scientifiques, la grande majorité de son contenu correspond à des productions contemporaines, mais il peut y avoir également des références plus anciennes, voire des siècles précédents. Les différentes communautés académiques (et donc les disciplines) se sont inégalement appropriées cet outil. On constate des variations en fonction des pratiques déjà mises en place dans les communautés (existence d'autres services de ce type, notamment à l'échelle internationale comme ArXiv pour la physique, les mathématiques et d'autres sciences exactes) et des incitations des acteurs institutionnels (le CNRS incite fortement les chercheurs de ses laboratoires à systématiser ce type de dépôt et l'ANR l'impose pour les travaux qu'elle finance). En ce qui concerne la production francophone et la variété des disciplines, il nous a toutefois paru que cette base de données était la plus à même de nous fournir des données pertinentes adaptées à une étude quantitative exploratoire des formes des titres des publications académiques francophones.

1.2. Description des données recueillies

Nous avons interrogé automatiquement le moteur de recherche de HAL pour en extraire les notices des documents écrits en langue française et publiés sous forme d'articles de journaux, de chapitres d'ouvrages collectifs ou de communications dans les conférences. Nous avons en effet choisi, dans un premier temps, de ne pas traiter les titres des autres types de documents disponibles (ouvrages, thèses ou mémoires, productions audiovisuelles, etc.) dont on peut supposer qu'ils impliquent des choix et des stratégies différentes dans la construction de leurs titres. Cette sélection nous a permis d'obtenir au total près de 340.000 notices. Pour chaque document indexé, diverses informations sont disponibles dans la base de données. Parmi celles-ci, nous avons choisi de conserver : le nom du ou des auteurs, le type de support, la date de publication et les domaines – autant d'informations que nous pourrons mettre en relation avec les caractéristiques linguistiques des titres que nous mettrons en évidence. Ces informations nous permettent de fournir immédiatement une première description des données qui composent notre corpus :

- Le nombre moyen d'auteurs par publication est de 1,95 mais 62 % n'ont qu'un seul auteur.
- S'agissant des supports de publication, 45 % des titres proviennent d'articles de journaux, 35 % de communications dans des conférences et 20 % de chapitres d'ouvrages collectifs.
- Les dates de publication sont concentrées sur la période contemporaine : 99 % des travaux ont été publiés après 1940, et 90 % après 1994. Les publications postérieures à 2000 représentent 85 % de nos données.
- Les fiches bibliographiques de HAL laissent la possibilité aux déposants d'indiquer autant de domaines qu'ils le souhaitent, à choisir dans une taxonomie à 4 niveaux comportant plus de 250 catégories. Nous n'avons pas conservé l'ensemble des informations domaniales disponibles et avons distingué, d'une part, les principales

disciplines des sciences exactes et, d'autre part, les disciplines relevant de la branche intitulée « SHS » (qui regroupe les sciences humaines et sociales mais aussi les arts et lettres). Pour chaque article, nous avons finalement conservé la discipline principale de chaque article telle qu'elle est indiquée par les auteurs. La répartition de ces disciplines par domaine et sous-domaine principaux est fournie dans le tableau 1 ci-dessous :

Domaine (sciences exactes)	**Titres**	**Domaine (sciences humaines)**	**Titres**
Physique	35 538	Sociologie	32 228
Sciences de la vie	24 638	Droit	30 517
Informatique	16 932	Histoire	25 754
Sciences de l'environnement	7 621	Gestion et sciences économiques	23 703
Sciences de l'univers	3 842	Linguistique	16 028
Sciences cognitives	3 141	Littérature	14 938
Mathématiques	3 092	Archéologie	13 745
Chimie	2 836	Sciences de l'éducation	9 696
		Sciences politiques	9 432
		Art et histoire de l'art	8 907
		Philosophie	8 450
		Sciences de l'information et de la communication	7 333
		Anthropologie et ethnologie	7 095
		Architecture	4 761
		Psychologie	2 654
Non renseigné	25 115	Géographie	1 149

Tableau 1 : Répartition des titres selon le domaine et le sous-domaine principal auxquels sont rattachés les articles scientifiques

On remarque un important déséquilibre dans la répartition entre ces principaux domaines. Ce déséquilibre ne traduit pas nécessairement la réalité de la production scientifique francophone contemporaine. Ainsi, le faible nombre de publications relevant de la psychologie ou de la géographie peut surprendre, et pourrait s'expliquer par les biais identifiés dans les archives ouvertes par Kim (2011). Ces biais se situent à des niveaux institutionnels, disciplinaires voire culturels et impactent la décision des chercheurs d'alimenter ou non cette base de données. Mais il ne semble pas à ce stade qu'ils aient affecté le choix des travaux individuels qui sont déposés dans les archives ouvertes institutionnelles, ni donc *a priori* les titres que nous avons recueillis. Nous avons donc choisi, dans cette première étude exploratoire, de nous fonder uniquement sur les données disponibles dans HAL sans recourir à un quelconque échantillonnage ou rééquilibrage du corpus.

Comme cela a déjà été indiqué dans des travaux antérieurs portant sur les titres de publications scientifiques, les différentes caractéristiques que sont le nombre d'auteurs, le type de support, la date de publication et le domaine ne sont pas indépendantes. On sait, notamment grâce à l'étude de Larivière et al. (2015), que le nombre d'auteurs est nettement supérieur dans les disciplines scientifiques et tend globalement à augmenter avec le temps (comme l'a montré Baethge, 2008). Par ailleurs, on sait que certaines disciplines favorisent certains types de support : les chapitres d'ouvrage sont plus fréquents en sciences humaines, par exemple.

1.3. Premier aperçu des titres

Pour pouvoir travailler sur les près de 340 000 titres de notre collection, nous avons commencé par effectuer un prétraitement automatique, à savoir l'étiquetage morphosyntaxique. Cette opération permet de découper chaque titre en *tokens* (mots graphiques et signes de ponctuation), d'attribuer à chacun de ces tokens une étiquette correspondant à sa catégorie morphosyntaxique (nom, verbe, adjectif, etc.) et son lemme (ou forme de citation, par exemple masculin singulier pour les adjectifs, infinitif pour les verbes). Pour réaliser cette tâche, nous avons utilisé le logiciel Talismane (Urieli et Tanguy, 2013) qui est à l'état de l'art pour le français et pour lequel nous avons vérifié en amont son comportement face à ce type particulier de segments de texte.

Une première exploration permet d'examiner la taille des titres et d'identifier les mots les plus fréquents dans l'ensemble du corpus.

1.3.1 Taille des titres

Les titres de notre corpus ont une longueur moyenne de 13,8 mots, mais qui s'étend de 1 à 285. Une rapide observation des titres dont la taille apparaissait tout à fait hors norme a permis de constater que le champ correspondant dans la base de données contenait parfois une tout autre information que le titre lui-même, comme par exemple le résumé de la publication, la référence complète de l'article (titre, auteurs, support, etc.), ou bien encore le titre français suivi de ses traductions dans diverses langues. Ces items ont donc tout simplement été retirés du corpus final qui contient les 99 % des titres qui ont une taille inférieure ou égale à trente-cinq mots. Les titres les plus courts[1] sont pour la plupart de titres de chapitres d'ouvrage en sciences humaines et correspondent souvent à des entrées d'ouvrages encyclopédiques : nom d'une notion, patronyme ou toponyme. Les titres de chapitres étant légèrement plus courts que ceux des deux autres supports (cf. 12 mots contre 14 en moyenne, ce qui est significatif (ANOVA, $p < 0,01$)), on peut émettre l'hypothèse que le contexte introduit par le sujet de l'ouvrage dans lequel ils s'inscrivent permet de réduire l'information nécessaire, ce qui n'est pas le cas pour les actes de conférences ou les articles de journaux.

La taille d'un titre est positivement corrélée au nombre d'auteurs ($\rho = +0,17$, $p<0,01$), ce qui confirme les observations précédentes sur ce point (Yitzhaki, 1994). Malgré l'étroitesse et le déséquilibre de notre période temporelle, on observe une très légère corrélation positive ($\rho =+0,04$, $p<0,01$) avec l'année de publication.

En ce qui concerne les disciplines, on observe des différences importantes avec dans la partie haute du spectre les sciences expérimentales (sciences de la vie, de

[1] On signalera qu'environ un millier des titres de notre corpus n'est constitué que d'un seul mot.

l'univers et de l'environnement, physique et chimie avec plus de 15 mots en moyenne) et le droit, et dans la partie basse la philosophie, les mathématiques et l'informatique avec moins de 12 mots.

Comme nous l'avons dit précédemment, ces différentes caractéristiques n'étant pas indépendantes, il n'est pas possible d'en tirer des conséquences dépassant le simple constat d'une grande variété de formes entre les titres de notre corpus.

1.3.2 Fréquences lexicales

Un outil d'observation classique des données textuelles est le recours aux fréquences lexicales, qui permettent de faire apparaître les éléments de vocabulaire les plus utilisés dans un corpus donné. En nous limitant aux seules classes ouvertes suivantes : noms, verbes et adjectifs, nous avons extrait les lexèmes présents dans le plus grand nombre de titres. Le tableau 2 rassemble les 30 les plus fréquents.

Lexème	**Fréquence (titres)**
étude	13 261
siècle	9 967
nouveau	9 326
droit	9 163
analyse	8 870
France	8 816
social	8 572
cas	8 437
français	7 394
système	6 927
approche	6 727
recherche	6 516
modèle	6 468
application	5 998
travail	5 738
histoire	5 510
développement	5 378
pratique	5 363
politique	5 132
méthode	5 119
exemple	5 114
public	5 050
espace	5 007
effet	4 990
gestion	4 701
enjeu	4 623

Lexème	Fréquence (titres)
européen	4 555
réseau	4 463
urbain	4 454
modélisation	4 409

Tableau 2 : Fréquence des trente lexèmes les plus récurrents dans le corpus de titres

Ainsi classée par fréquence, cette liste des trente lexèmes les plus fréquents dans les titres de notre corpus apparaît assez hétérogène. Plusieurs observations peuvent cependant être faites. D'abord, cette liste est composée quasi exclusivement de noms (dont un nom propre). Elle compte peu d'adjectifs mais l'adjectif *nouveau* est le troisième mot le plus fréquent. La présence massive de cet adjectif dans ces titres ne surprend pas. En effet, dans la mesure où l'une des fonctions du titre est d'attirer l'attention du lecteur, on n'est pas étonné qu'un moyen utilisé pour retenir l'attention consiste à annoncer quelque chose de nouveau. Dans cette liste, un lexème se distingue très nettement des autres par sa fréquence : le nom *étude* que l'on trouve dans près de 13 300 titres différents. Ce lexème appartient à la liste des noms qui constituent le lexique transdisciplinaire des écrits scientifiques mis en évidence par Tutin (2007). Cette liste est organisée en sous-classes distinguant notamment les lexèmes désignant les supports de la rédaction scientifique des lexèmes dont le contenu sémantique est moins fonctionnel. Il s'agit notamment de noms appartenant à la classe des objets construits par l'activité scientifique, comme ici *étude*, *analyse*, *approche*, *recherche*, *modèle application*, *méthode*, *modélisation* et *développement* et de noms appartenant à la classe des observables de l'activité scientifique, comme *cas*, *exemple*, *effet* et *enjeu* (cf. Tutin, 2008). On trouve enfin, dans la liste du Tableau 2 un sous-ensemble de lexèmes liés aux principales thématiques et principaux objets d'étude : *siècle, droit, France, social, français, travail, histoire, politique, pratique, public, espace, gestion, européen, réseau, urbain*[2]. Nous verrons plus loin que ceux-ci se retrouvent employés dans des schémas lexico-syntaxiques productifs de titres, que nous allons maintenant aborder.

2. Mise au jour de schémas récurrents dans les titres d'articles scientifiques

La suite de notre travail se concentre sur le repérage de schémas afin de dégager les principales structures utilisées pour construire un titre d'article. Pour ce faire, nous nous sommes directement inspirés des méthodes de fouille de texte visant l'extraction de séquences de mots récurrentes, comme l'ont opérationnalisé (Quiniou et al. 2012) pour l'analyse de textes littéraires. Le principe général de ce type d'approche consiste à dépasser la simple répétition à l'identique en envisageant les séquences partielles, tant que l'ordre des éléments est respecté (par exemple « méthodes d'analyse des textes » et « méthodes de commentaire de texte » ont en commun la séquence « méthode de * de texte »). Ces approches ont montré leur capacité à repérer des motifs pertinents

[2] Le premier classement que nous proposons ici mériterait d'être affiné et plus précisément mis en relation avec les domaines scientifiques des publications concernées.

correspondant à différents types d'unités syntaxiques ou discursives. Appliquées aux écrits scientifiques, elles permettent également de repérer des expressions phraséologiques (cf. Tutin et Kraif, 2016).

Dans notre cas toutefois, nous avons dû développer une variante de cette méthode pour deux raisons. D'abord, dans la mesure où notre objectif est d'identifier les structures des titres, nous ne pouvons pas nous contenter de rechercher des séquences qui ne les couvriraient pas dans leur intégralité. Ensuite, nous avons choisi de traiter de manière distincte les classes ouvertes et les classes fermées (déterminants, prépositions, conjonctions, adverbes, pronoms), ainsi que les signes de ponctuation (virgules, points, double-points, parenthèses, etc.).

Nous décrivons en détail la procédure que nous avons suivie ci-dessous :

Dans une première étape, nous avons appliqué à chaque titre la série de transformations suivantes :

- le titre est segmenté, étiqueté et lemmatisé (comme indiqué précédemment) ;
- chaque élément d'une classe fermée (ni nom, ni verbe, ni adjectif) est préservé sous sa forme canonique (*une* devient *un*, *la* ou *les* deviennent *le*, etc.), dans le cas des articles contractés avec les prépositions *à* et *de* (*au*, *aux*, *du* et *des*), ils sont retranscrits sous leur forme expansée (*à le* et *de le*) ;
- les unités lexicales, noms, verbes et adjectifs, sont remplacées par une étiquette générique conservant la catégorie (N, V ou A).

Par cette opération, des titres, comme ceux donnés ci-dessous en (1) et (2), vont être représentés sous la forme de la séquence « V le N_1 A : de le N_2 à le N_3 » :

(1) Construire l'espace public : du diagnostic au projet
(2) Mesurer la convergence interactionnelle : de la similarité à l'affiliation

Le schéma ainsi défini est cependant trop strict, notamment parce qu'il ne laisse pas la place à de possibles modifieurs des noms N_2 et N_3 qui le composent. De plus, il restreint la modification de N_1 à la modification adjectivale, excluant notamment les syntagmes prépositionnels.

Dans une seconde étape, pour pallier ces limites, nous procédons à un regroupement des syntagmes nominaux. Ce regroupement s'effectue en utilisant une grammaire locale qui a pour but : 1) d'identifier les séquences incluant des adjectifs (le déterminant étant optionnel et les adjectifs en nombre non restreint), et 2) d'identifier les syntagmes complexes récursifs incluant la préposition *de*, autrement dit des constructions du type « N de N de N ». Chacun des regroupements ainsi effectué correspond à un schéma que nous avons appelé « X ».

Suite à cette opération de regroupement, on passe, pour les titres illustrés en (1) et (2), du schéma initial « V le N_1 A : de le N_2 à le N_3 » au schéma suivant : « V X : de X à X ». Ce schéma, qui est beaucoup plus accueillant, permet de rassembler des titres comme ceux listés ci-dessous :

en (3), un titre très proche du schéma initial mais dans lequel N_2 et N_3 sont modifiés :

(3) Quitter la métropole parisienne : des avantages résidentiels aux avantages économiques

en (4), N_1 est modifié, non par un A, mais par un SP introduit par *de* :

(4) Défendre la cause de l'environnement : de la professionnalisation aux professions

en (5), N_1 n'est pas modifié, en revanche N_3 l'est :

(5) Exploiter les données : du codage au tableau statistique

tel est également le cas en (6), où N_1 n'est pas modifié, mais où N_2 et N_3 le sont, chacun par un SP introduit par *de* :

(6) Dire le génocide : des Antimémoires de Malraux à l'autofiction de Doubrovsky

(7) illustre une autre variation de la modification des noms N_2 et N_3 :

(7) Dialoguer le projet : de la participation des habitants à la programmation générative

et, pour finir, l'exemple (8) donne une illustration de la complexité des syntagmes nominaux présents dans les titres de notre corpus :

(8) Évaluer les impacts sociaux des politiques de mobilité urbaine : de l'accessibilité spatiale à l'accessibilité sociale

Ces schémas abstraits sont très utiles parce qu'ils rendent possibles des regroupements à large spectre. Ils ont cependant un défaut puisqu'ils ne permettent pas de mettre au jour des schémas plus précis qui associeraient par exemple une même structure syntaxique avec des lexèmes particuliers. Pour mettre au jour ce type de configuration, une étape supplémentaire est nécessaire. Cette étape consiste à ajouter, à notre procédure de transformation, la possibilité de maintenir sous sa forme de citation tout nom, verbe ou adjectif à condition que sa fréquence totale dans le corpus atteigne un seuil, que nous avons arbitrairement fixé à 300 titres. Ce seuil a uniquement pour but de limiter l'explosion combinatoire, notamment pour les titres les plus longs. Pour repérer les lexèmes fréquents, la procédure que nous avons définie a pour effet de multiplier le nombre de schémas générés. Cela s'explique par le fait que, à la génération des schémas à base d'étiquettes catégorielles décrite précédemment, s'ajoute le maintien des catégories lexicales, nom, verbe ou adjectif. Pour illustrer cette procédure, nous donnons, ci-dessous, la liste complète des schémas générés à partir du titre (9) :

(9) Architecture d'un serveur multimédia pour les sciences de l'ingénieur

Sachant que, parmi les lexèmes présents dans ce titre, seuls les noms *architecture*, *science* et *ingénieur* dépassent le seuil fixé, nous obtenons les 8 schémas suivants, du plus spécifique au plus abstrait :

a. architecture de X pour le science de le ingénieur
b. architecture de X pour le science de X
c. architecture de X pour X de le ingénieur
d. architecture de X pour X

e. X pour le science de le ingénieur
f. X pour le science de X
g. X pour X de le ingénieur
h. X pour X

Finalement, en appliquant cette procédure aux quelque 340 000 titres de notre corpus, nous générons plus de 41 millions de schémas lexico-syntaxiques différents. Une table de fréquence permet alors simplement d'identifier les schémas les plus productifs. Pour illustrer, nous donnons ci-dessous l'ensemble de ceux dont la fréquence est supérieure ou égale à 500.

Schéma	**Fréquence**
X	19 276
X et X	10 351
X à X	5 726
X dans X	5 466
X en X	4 021
X : X	3 830
X pour X	2 269
X sur X	2 218
X , X	2 027
X , X et X	1 953
X et X : X	1 762
X : X et X	1 753
X et X dans X	1 748
X - X	1 520
X par X	1 450
X et X en X	1 214
X et X à X	1 151
X (X)	826
X .	817
X . X	810
X et de X	692
X : X à X	674
sur X	669
X à X : X	666
V X	604
X entre X et X	597
de X à X	589
X dans X : X	571

Schéma	Fréquence
X V à X	564
X en X : X	546
X chez X	503
X comme X	500

Tableau 3 : Liste des schémas dont la fréquence est supérieure ou égale à 500

En nous limitant à ceux qui réunissent au moins 10 titres, nous identifions 5 182 schémas différents. Le schéma de titre le plus fréquent correspond aux titres constitués d'un syntagme nominal, que nous avons noté « X ». Étant donné la méthode de calcul des schémas que nous avons définie, ce schéma rassemble des réalisations très variées comme l'illustrent (10) et (11) :

(10) L'identité
(11) La nature du droit d'occupation du logement familial

Si l'on examine les fréquences les plus basses, on voit apparaître des schémas dont les formes sont beaucoup plus spécifiques et qui présentent des constantes lexicales remarquables, comme on le voit dans le tableau 4, avec des lexèmes comme *cas, français, place, prisme, question, réseau, identification, influence* et même des séquences comme *développement durable* ou *changement climatique* :

Schéma		Exemple	Fréquence
1	X et X : le cas de X	Sports et identité nationale : le cas du football professionnel	238
2	X français	Le morcellement des exploitations agricoles françaises	167
3	La place de X dans X	La place des éléments radioactifs dans le système périodique	108
4	qu'est -ce que X ?	Qu'est-ce que l'intermédiation algorithmique ?	98
5	X au prisme de X	Les parfums antiques au prisme de l'analyse chimique	75
6	X et la question de X	Les villes nouvelles de Singapour et la question de la démocratie locale	46
7	X dans le réseau X	Le caching proactif dans les réseaux cellulaires 5G	44
8	X et développement durable	Lois de conservation économiques et développement durable	36
9	X à X : X et de X	L'exposition aux pesticides : risque de lymphome et de leucémie	33
10	identification de X par X	Identification de sources multipolaires équivalentes par filtrage spatial	20
11	à qui V X ?	À qui profitent les réseaux de transport ?	10
12	de l'influence de X sur X	De l'influence de l'aimantation sur les propriétés thermoélectriques	10
13	X du changement climatique sur X	Impact du changement climatique sur les populations d'anguilles	7

Tableau 4 : Illustrations de schémas peu fréquents

Ces schémas, plus contraints, semblent correspondre à des classes de titres homogènes du point de vue de leur structure et manifestent des usages typiques de certaines disciplines. Parmi ces schémas, on notera le cas un peu particulier des commentaires d'arrêt en droit (classés dans la base de données de HAL comme des articles) dont les titres obéissent à une grammaire assez rigide qui engendre de nombreux schémas, comme par exemple : *note sous (tribunal / cours / conseil, etc.)*, dont (12) est une occurrence :

(12) Note sous Cour d'appel de Paris, troisième Chambre A, 4 janvier 2005, Groupe Limagrain contre Cariel

Finalement, notre méthode reste quelque peu grossière notamment parce qu'elle n'est pas absolument fidèle à la structure syntaxique des titres. Si l'on examine le schéma 9, dans le tableau 4 ci-dessous, on a un exemple de schéma qui n'est pas complètement satisfaisant, dans la mesure où la structure de la partie du titre qui se trouve à droite des deux points devrait être la suivante : « X de X et de X ». Ces approximations s'expliquent pour deux raisons. D'abord, par le fait que l'analyse

syntaxique automatique de ce type de données est extrêmement difficile pour des raisons liées notamment aux ambiguïtés inhérentes aux structures nominales complexes. Et ensuite, par le fait que les analyseurs automatiques, comme Talismane, sont entraînés sur des données textuelles plus « classiques », où les phrases contiennent généralement une forme finie de verbe, ce qui n'est qu'exceptionnellement le cas dans notre corpus.

Toutefois, l'approche quantitative que nous employons pour filtrer et analyser les schémas a pour effet d'atténuer l'importance de ces imprécisions locales qui finalement ne viennent pas perturber les résultats finaux que nous produisons.

Pour aller au-delà des premières illustrations que nous venons de présenter dans cette première section, nous allons, dans la suite de notre article, nous focaliser sur des applications plus systématiques de notre méthode.

3. Analyses

3.1. Relations entre schémas et domaines

Parmi les différentes caractéristiques extralinguistiques disponibles pour chaque titre de notre corpus, nous avons choisi de nous pencher plus spécifiquement sur le domaine, dont on peut voir dans les exemples de schémas donnés plus haut qu'il joue un rôle important dans le choix de telle ou telle forme de titre. Pour ce faire, nous avons calculé la distribution de nos 5 182 schémas suivant les différentes disciplines. Si certains schémas très génériques (comme par exemple le schéma noté « X ») sont utilisés massivement dans chaque discipline, il apparaît que certains d'entre eux ont, au contraire, des fréquences nettement plus variées. Afin de prendre en compte la variation de fréquence de chaque schéma et les différences de représentation des disciplines dans notre corpus, nous avons calculé les résidus normalisés de Pearson pour chaque couple (schéma, discipline). Cette technique, qui est utilisée notamment dans le test du chi-deux, permet d'identifier les couples dont la fréquence est différente (excédentaire ou déficitaire) par rapport à la fréquence attendue que l'on obtiendrait si les schémas étaient distribués de façon homogène à travers les disciplines. La valeur normalisée ainsi obtenue permet aisément d'identifier les associations privilégiées entre un schéma et une discipline et ainsi, par exemple, de repérer les schémas de titre favorisés par chacun des domaines représentés dans le corpus.

Le tableau ci-dessous présente les 5 titres les plus spécifiques (dont les résidus de Pearson sont les plus élevés) parmi les schémas apparaissant plus de 50 fois au total dans notre corpus.

Chimie	**Informatique**	**Mathématiques**	**Physique**
X par X	X pour X	X pour X	sur X
X sur des X	X à partir de X	sur X	X par X
X pour X	X pour X dans X	X pour des X	X pour X
nouveau X pour X	X : X pour X	X pour l'analyse de X	influence de X sur X
X à base de X	X basé sur X	quelques X	sur X dans X
Sc. cognitives	**Sc. de l'environnement**	**Sc. de l'univers**	**Sc. de la vie**

X et X	X - X	X sur X (X)	X chez X
X	X - X , X	X (X)	X sur X chez X
X par X	X - X et X	X (X , X)	X à X chez X
X et X dans X	X en milieu A	X dans X (X)	X et X chez X
X et X : entre X et X	évaluation de X	X dans X	X au cours de X
Anthropologie	**Archéologie**	**Architecture**	**Art**
X . X	X (X , X)	X urbain	notice X
X et X chez X	X à X (X)	X et X urbain	X
X , X et X en X	X (X)	V X	X , X . X
entretien avec X	X (X) : X	X comme X	X dans X du A siècle
X	X , X	X , entre X et X	X , X
Droit	**Sc. Éducation**	**Géographie**	**Gestion & Sc. Eco**
chronique de X (X)	X en X et X	V X en X	X et X : X
chronique de X civil (X)	X en X : X	X et X : X et de X	X économique de X
X civil (X)	X au service de X	X à X ?	X : X
X de procédure civile (X)	V X à X	X en X : X et X	X et X : le cas de X
X de procédure A (X)	X dans X : de X à X	X à X : de X à X	l'impact de X sur X
Histoire	**Sc. Info. Com.**	**Linguistique**	**Littérature**
X (X)	X numérique	X et X en X	X et X dans X
X au A siècle	X à l'heure de X	X en X	X ou X
X et X dans X (X)	X : entre X et X	X dans X	notice X
X dans X (X)	X comme X	X : le cas de X	X
X dans X au A siècle	X : vers X	X dans X en X	X dans X
Philosophie	**Psychologie**	**Sc. Politiques**	**Sociologie**
X et X	X chez X	X politique de X	V X
X et X chez X	X et X social	V : X	X et X urbain
X , X et X chez X	remarque sur X	X en Europe	X urbain
X chez X et X	X et A dans X	X et X public	X , X et X
qu'est-ce que X ?	X sur X chez X	X européen	X et X

Tableau 5 : 5 schémas les plus spécifiques par discipline

On peut voir qu'à de très rares exceptions chaque discipline présente des schémas très spécifiques qui la distinguent nettement des autres.

Ces différences concernent d'abord la nature de la publication. Certains schémas commencent en effet par un nom qui indique le type de travail académique publié : on trouve le terme de *notice* en littérature, le terme *chronique* en droit, le terme *remarque*

en psychologie, le terme *entretien* en anthropologie, le terme *évaluation* en sciences de l'éducation.

Ces distinctions concernent également, ce que l'on pourrait appeler le point de vue porté sur l'objet d'étude, que l'on peut voir exprimé notamment dans le choix des prépositions. Par exemple, l'emploi de *pour* ou de *par*, dans les sciences dures, nous semble associé à la nature méthodologique des travaux publiés. L'usage de la préposition *chez* (suivi de la dénomination d'une espèce ou d'une famille) est très fréquent dans le domaine des sciences de la vie. On retrouve également cette préposition en philosophie, mais elle est alors suivie du nom de l'auteur étudié. Les titres des travaux publiés dans le domaine de la linguistique présentent deux autres prépositions spécifiques : *en* et *dans*. La préposition *en* sert à introduire la langue étudiée (*en anglais*, *en français*, etc.), quant à *dans*, elle introduit les données (œuvre, type de productions langagières, structure linguistique étudiée).

Les spécificités des schémas par discipline se manifestent également par le fait que chaque discipline possède ses propres thématiques et objets centraux. On trouvera *politique*, *public*, *Europe* en sciences politiques ; *Xième siècle* en histoire et en histoire de l'art ; *urbain* en architecture et en sociologie ; *économique* en gestion; *numérique* en sciences de l'information et de la communication, etc.

Si l'on examine maintenant la forme générale des schémas, au-delà de ce que nous avions déjà noté à savoir que les titres se réalisent très majoritairement sous la forme de syntagmes nominaux, certaines spécificités peuvent être relevées. On notera d'abord le recours à des syntagmes prépositionnels, essentiellement introduits par *sur*, dans les domaines de la physique et des mathématiques, cette préposition servant a priori à limiter la couverture perçue du travail présenté (« *sur* le calcul des forces magnétiques » semble en effet de moindre envergure qu'un titre comme « le calcul des forces magnétiques », que l'on s'attendrait à trouver sur une monographie ou un manuel). Si l'on prend en considération plus de 5 schémas, on voit rapidement apparaître, dans d'autres disciplines, d'autres prépositions ou locutions prépositionnelles avec le même effet de minimisation du propos, comme *à propos de* ou *vers*.

On peut également signaler l'usage de certaines formes associées à certaines disciplines : comme les formes verbales à l'infinitif en sociologie, sciences de l'éducation, architecture et sciences politiques, ou les formes interrogatives en philosophie et en géographie ou encore l'introduction d'un sous-titre par le biais des deux points. Cet usage est d'ailleurs fréquent dans de nombreuses disciplines, et c'est le schéma générique qui a été le plus étudié, notamment pour identifier le lien sémantique entre les deux parties (Anthony 2001, Hartley 2007b).

Dans le but de généraliser encore davantage nos observations, nous avons effectué le même travail de fouille en regroupant les disciplines. Comme il est d'usage, nous avons distingué d'un côté les sciences dures et de l'autre les sciences humaines et sociales (au sein desquelles nous avons inclus les arts et la littérature). Le tableau ci-dessous donne, pour chacun de ces deux groupes, les schémas de titres pour lesquels l'écart des résidus entre les disciplines est le plus important. En suivant le même principe, la colonne de droite, intitulée « schémas communs », fournit la liste des cinq schémas pour lesquels la distinction entre les disciplines est la moins marquée.

Schémas spécifiques		
Sciences dures	**SHS**	**Schémas communs**
X pour X	X et X	X dans des X
X par X	X , X	le rôle de X dans X
sur X	X et X : X	X sur X à X
X sur X	X et X dans X	application de X à X
X pour X en X	X , X et X	étude de X

Tableau 6 : 5 schémas les plus spécifiques et les plus communs, par groupe de disciplines

La distribution complémentaire qui se dégage des données est tout à fait remarquable dans la mesure où elle oppose deux grands types de configuration de titres : 1) les titres des articles publiés dans le domaine des sciences dures où l'articulation entre deux unités (notées « X ») repose sur un sous-ensemble de prépositions : *pour*, *par* et *sur* ; 2) les titres des articles du domaine des SHS où le lien entre les deux (ou trois) unités est fondé sur la conjonction de coordination, *et*, et/ou sur une marque de ponctuation (virgule ou deux points). Cette distribution nous semble pouvoir être associée à des différences de démarches. Les titres des articles publiés dans le domaine des SHS peuvent en effet être décrits comme des titres qui confrontent des concepts, des points de vue ou des méthodes à propos d'un objet d'étude particulier. Ils s'opposent en cela assez nettement aux titres des sciences dures qui se focalisent sur les effets d'une méthode sur un objet. Pour éclairer ce que nous essayons de dire ici, nous proposons de confronter quelques titres issus des schémas « X pour X » (cf. exemples donnés sous 13) et « X et X » (cf. exemples sous 14) :

(13 a) Chaînes de Markov multi-phases floues pour l'évaluation de la performance imprécise des Systèmes Instrumentés de Sécurité (Physique)
(13 b) Adaptation d'une ressource prédicative pour l'extraction d'information (Informatique)
(13 c) Les biopiles hybrides pour la production d'énergie électrique (Chimie)
(13 d) Régression linéaire locale pour variable fonctionnelle (Mathématiques)

(14 a) Généricité du syntagme nominal sujet et modalités (Linguistique)
(14 b) Le Crédit Lyonnais et le financement de l'économie (Sciences économiques)
(14 c) Jules Verne et la Méditerranée (Littérature)
(14 d) Ingénierie de projet et excellence territoriale (Sociologie)

Ces quelques exemples nous paraissent opposer des titres qui, dans la première série (13), explicitent clairement la relation entre deux notions, la première étant généralement un procédé et la seconde son objet d'application ou son objectif. Dans la seconde série (14), le champ reste plus ouvert, et le titre ne permet pas d'identifier directement de relation orientée, que cela soit dû à un choix dans la formulation ou au fait que cette relation est trop complexe ou multiple pour être formulée dans le titre.

Le tableau 6 met également en lumière les principaux schémas communs (dont les résidus sont les plus faibles en valeur absolue). Ce qui retient l'attention ici ce sont les deux types de schémas qui apparaissent. D'un côté, des schémas lexicalement

instanciés par des noms qui relèvent clairement du lexique transdisciplinaire des écrits scientifiques et de l'autre des schémas dans lesquels l'élément de stabilité, le pivot est constitué par une préposition.

Une étude plus fine et détaillée des réalisations de ces schémas serait maintenant nécessaire pour aller plus loin dans l'analyse.

3.2. Comment mettre au jour des schémas réguliers mais spécifiques ? Le cas des chiasmes

La méthode que nous avons présentée permet de mettre au jour les propriétés linguistiques d'un très vaste ensemble de titres d'articles scientifiques mettant ainsi en évidence les habitudes de leurs auteurs. Cette méthode, comme toutes les méthodes inductives de fouille de texte, n'en demeure pas moins inadaptée à faire apparaître des structures dont la spécificité repose sur le lexique et non sur la structure, comme c'est le cas par exemple des chiasmes.

Si nous avons pu mettre au jour l'usage de figures de type chiasmatique dans les titres de notre corpus, ce n'est pas en nous appuyant sur notre méthode, mais en parcourant les titres classés selon le(s) schéma(s) qui le subsume(nt), autrement dit en procédant par sondage. Il n'est en effet pas possible de faire émerger les chiasmes du corpus puisque ceux-ci reposent sur un parallélisme syntaxico-sémantique. Cette figure de style consiste, comme on peut le lire dans le *Trésor de la Langue Française*, à « inverser l'ordre des termes dans les parties symétriques de deux membres de phrase de manière à former un parallèle ou une antithèse », comme dans le célèbre : *il faut manger pour vivre et non pas vivre pour manger*.

Plus précisément, il s'agit de titres dans lesquels une paire de mots, A et B (noms, verbes, adjectifs), occurrent d'abord dans l'ordre AB puis dans l'ordre BA. Une fois la récurrence du phénomène établie, il est tout à fait possible de concevoir un programme informatique simple permettant de le repérer. Cela nécessite toutefois le développement d'un programme complètement *ad hoc* qui recherche les répétitions de lemmes et leur inversion sans autre contrainte sur la structure syntaxique. Nous donnons ci-dessous des exemples des quelques 200 titres correspondant à cette configuration (en faisant apparaître en gras le couple A et B) qui montrent bien la diversité des associations :

(15) **Contacts** de **créoles**, **créoles** en **contact**
(16) L'apprentissage du raisonnement clinique (ARC) en anglais : De l'**anglais** pour la **médecine** à la **médecine** en **anglais**
(17) **Pouvoirs** de la **famille**, **familles** de **pouvoir**. Histoire et anthropologie
(18) **Contraintes** sur le **discours** et genre de **discours contraint :** le commentaire sportif télévisé en direct
(19) Oikiste et tyran : **fondateur** - **monarque** et **monarque** - **fondateur** dans l'Occident grec
(20) Du **hasard** comme **provocation** à la **provocation** comme source de **hasard**
(21) Ce que l'**Europe** fait au **protestantisme**, ce que le **protestantisme** fait à l'**Europe**
(22) L'**orientation politique** des **politiques** d'**orientation**
(23) Racines et floraisons : **tradition poétique**, **poétique** des **traditions** dans l'œuvre de Jacques Roubaud

Ce type de configuration s'observe quasi-exclusivement dans les titres relevant du domaine des SHS. On citera un exemple extrait des sciences dures, dont on peut toutefois penser qu'il n'est pas figural :

(24) Influence de l'état électrique d'une **surface liquide** sur la tension maxima de la vapeur de ce **liquide** en contact avec la **surface**

Perspectives

Au terme de cette première étude, qui a surtout permis de poser les fondements de la méthode et de donner une idée de la richesse du matériau, de nombreuses pistes s'ouvrent pour prolonger et affiner les résultats présentés ici. Il s'agirait notamment d'étudier plus en détail certains schémas – toujours en comparant les disciplines entre elles –, comme par exemple les titres contenant certains signes de ponctuation dont les deux points ou le point d'interrogation. Nous avons par exemple repéré à l'usage de formes interrogatives dans de très nombreux schémas du type de : « vous avez dit X ? », « quel X pour X ? », « que reste-t-il de X ? » ou encore « à qui profite X ? ». L'étude du lexique mériterait également d'être menée plus loin en s'inspirant notamment de l'analyse collostructionnelle proposée par Stefanowitsch et Gries (2003). Dans la mesure où nous disposons désormais des schémas de titres que l'on pourrait considérer comme des constructions, nous pourrions maintenant mesurer l'attraction/répulsion qu'exercent ces schémas sur les unités lexicales. Partant d'une sélection des schémas les plus productifs, il s'agirait d'étudier les parties variables de ces schémas (à savoir le contenu lexical de ce que nous avons appelé « X ») en mettant en évidence les préférences de certains lexèmes pour certains schémas. On pourrait par exemple se focaliser sur les lexèmes qui apparaissent le plus souvent après un deux points en faisant l'hypothèse que les lexèmes « attirés » sont des lexèmes appartenant au lexique dit transdisciplinaire.

Notre excursion dans le domaine du chiasme est là pour illustrer un phénomène récurrent dans les titres parmi d'autres que les schémas lexico-syntaxiques ne permettent pas de mettre au jour. Il permet de rappeler que les approches inductives ne sont pas les seules qui sont disponibles, et qu'un sondage précis reste possible, et est généralement simple et efficace, même si comme ici il demande l'application d'une technique spécifique. La masse critique atteinte permet en revanche d'y appliquer des méthodes d'analyse plus systématique (en l'occurrence, le lien exclusif avec les SHS).

Il apparaît donc clairement que ces données peuvent être abordées de différentes façons, en mobilisant différents niveaux de description, et permettront à terme de proposer des typologies plus larges, comme celle qu'a proposée Hartley (2007a), mais sur la base d'un travail de fouille systématique et reproductible.

Ludovic Tanguy, Josette Rebeyrolle
CLLE : CNRS & Université de Toulouse
ludovic.tanguy@univ-tlse2.fr ; josette.rebeyrolle@univ-tlse2.fr

Bibliographie

ANTHONY, Laurence (2001), « Characteristic features of research article titles in computer science », *IEEE Transactions on Professional Communication*, *44* (3), pp. 187-194.

BAETHGE, Christopher (2008), « Publish together or perish: the increasing number of authors per article in academic journals is the consequence of a changing scientific culture », *Deutsches Arzteblatt international*, 105 (20), pp. 380-383.

BENDINELLI, Marion (2017), « Segments phraséologiques et séquences textuelles », *Corpus*, 17.

BUSH-LAUER, Ines (2000), « Titles of English and German Research Papers in Medicine and Linguistics Theses and Research Articles », In Trosborg A. (ed.) *Analysing Professional Genres*. Amsterdam/Philadelphia: John Benjamins, pp. 77-94.

HAGGAN, Madeline (2004), « Research paper titles in literature, linguistics and science: dimensions of attraction », *Journal of Pragmatics*, 36 (2), pp. 293-317.

HARTLEY, James (2007a), « There's more to the title than meets the eye: Exploring the possibilities », *Journal of Technical Writing and Communication*, 37 (1), pp. 95-101.

HARTLEY, James (2007b), « Planning that title : Practices and preferences for titles with colons in academic articles », *Library and Information Science Research*, 29 (4), pp. 553-568.

HO-DAC, Lydia-Mai., JACQUES, Marie-Paule, REBEYROLLE, Josette (2004), « Sur la fonction discursive des titres », In S. Porhiel & D. Klingler (Eds). *L'unité texte*, Pleyben, Perspectives, pp. 125-152.

KIM, Jihyun (2011), « Motivations of Faculty Self-archiving in Institutional Repositories », *Journal of Academic Librarianship*, 37 (3), pp. 246-254.

LARIVIÈRE, Vincent, GINGRAS, Yves, SUGIMOTO, Cassidy R., TSOU, Andrew (2015), « Team size matters: Collaboration and scientific impact since 1900 », *Journal of the Association for Information Science and Technology*, 66 (7), pp. 1323-1332.

LEGALLOIS, Dominique, TUTIN, Agnès (2013), « Présentation : vers une extension du domaine de la phraséologie », *Langages*, 189, pp. 3-25.

MAGRI, Véronique, PURNELLE, Gérald. LEGALLOIS, Dominique (2016), « Mot à mot, brin par brin : les suites [Nom préposition Nom] comme indices de littérarité ? », *Actes des Journées d'Analyse des Données Textuelles (JADT)*, Nice, pp. 365-376.

QUINIOU, Solen, CELLIER, Peggy, CHARNOIS, Thierry, LEGALLOIS, Dominique (2012), « What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics ? », *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'12)*, New Delhi, pp. 166-177.

REBEYROLLE, Josette, JACQUES, Marie-Paule, PÉRY-WOODLEY, Marie-Paule (2009), « Titres et intertitres dans l'organisation du discours », *Journal of French Language Studies*, 19, pp. 269-290.

SOLER, Viviana (2011), « Comparative and contrastive observations on scientific titles written in English and Spanish », *English for Specific Purposes*, 30 (2), pp. 124-137.

STEFANOWITSCH, Anatol, GRIES, Stefan Th. (2003), « Collostructions : investigating the interaction between words and constructions », *International Journal of Corpus Linguistics*, 8 (2), pp. 209-243.

TUTIN, Agnès (2007), « Autour du lexique et de la phraséologie des écrits scientifiques », *Revue Française de Linguistique Appliquée*, 12 (2), pp. 5-14.

TUTIN, Agnès (2008), « Sémantique lexicale et corpus : l'étude du lexique transdisciplinaire des écrits scientifiques », *Lublin studies in modern languages and littérature*, 32, pp. 242-260.

TUTIN, Agnès, KRAIF, Olivier (2016), « Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents », *Lidil*, 53.

URIELI, Assaf, TANGUY, Ludovic (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.

YITZHAKI, Moshe (1994), « Relation of title length of journal articles to number of authors », *Scientometrics*, 30, pp. 321-332.