Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Détection automatique de l'innovation lexicale dans des corpus diachroniques

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Solveig PODER

Directeur de mémoire :

Kata Gabor

Encadrant :

Gaël Lejeune

Année universitaire 2020/2021

Résumé

Le figement lexical est un phénomène central du langage et les expressions figées représentent une importante proportion du lexique de toute langue. Leur détection demeure un des enjeux du TAL. Après avoir effectué un résumé détaillé de l'état de l'art dans ce domaine, ce mémoire présente une méthode non supervisée permettant de détecter la formation de nouvelles expressions figées au sein de corpus diachroniques d'articles de presse en français. On utilisera un modèle LDA (allocation de Dirichlet latente) pour extraire les expressions les plus représentatives du corpus que l'on considèrera comme des candidats au figement. Puis, partant du principe établi par les linguistes qu'une expression est figée si ses termes n'admettent pas d'être remplacés par des synonymes, nous utiliserons des plongements de mots pour établir une liste d'expressions synonymes pour chaque expression candidate. Enfin, nous calculerons l'évolution au fil du corpus du taux d'apparition d'une expression par rapport à ses « synonymes » (nous dresserons automatiquement pour chaque candidat une liste de potentiels synonymes en remplaçant les mots de l'expression candidate par des mots dont la représentation en vecteur est similaire).

Mot clés : *expressions polylexicales, collocations, figement lexical, LDA*

Remerciements

Je tiens en premier lieu à remercier mon tuteur de stage Gaël Lejeune pour sa confiance et sa disponibilité, ainsi que mon encadrante de mémoire Kata Gabor pour ses conseils et son écoute.

Je souhaite également remercier mon entreprise qui m'a autorisé une absence de deux ans pour effectuer mon Master et à l'organisme *Transitions Pro Île-de-France* pour avoir accepté de financer intégralement cette formation qui m'a passionnée et m'offre aujourd'hui de nouvelles perspectives professionnelles.

Je veux également remercier Lucas Eberhardt qui m'a poussée à faire ma reconversion professionnelle et m'a soutenue pendant ces deux années, ainsi que Johanna Garnier qui m'a appris l'existence du CPF de transition professionnelle.

Enfin, je veux remercier tous les professeurs de la formation pour les connaissances qu'ils m'ont apportées, ainsi que tous mes camarades (et plus particulièrement Camille Rey, Juliette Caron, Chinatsu Kuroiwa et France Cazelles, qui sont devenue mes amies) pour leur solidarité et leur humour qui m'ont permis de ne jamais me décourager, même dans les moments les plus difficiles.

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

Introduction

Objectif

L'objectif de ce travail sera de détecter les phénomènes de figements lexicaux dans des corpus diachroniques centrés autour d'un même sujet. En effet, nous émettons l'hypothèse que, sur un sujet donné, l'on peut fréquemment voir émerger des expressions que les locuteurs (ou certains groupes de locuteurs) vont s'approprier, ce qui va créer un fort lien sémantique entre ladite expression et le sens qu'elle revêt dans le cas de cette thématique particulière au point d'éclipser les autres interprétations possibles de l'expression, mais également d'exclure d'autres formulations possibles du concept ainsi exprimé.

Applications et enjeux

Ce travail pourra trouver diverses applications liées à l'analyse de corpus diachroniques. En effet, la recherche en sciences humaines est de plus en plus amenée à traiter de très vastes corpus qui peuvent quelquefois s'avérer difficiles à aborder du fait de leur grande taille et parce que, dans le cadre d'une démarche dite « corpus-driven », le chercheur ne souhaite pas confirmer une théorie mais faire parler le corpus sans a priori. Les outils textométriques, faciles à prendre en main, peuvent donner de premières pistes d'analyse mais s'avèrent parfois insuffisants pour détecter des phénomènes subtils et progressifs comme l'apparition de figements lexicaux, phénomène qui peut pourtant se révéler très intéressant lorsque l'on analyse l'évolution du traitement d'un sujet donné.

Le figement lexical est un phénomène inhérent au langage naturel : une séquence fréquemment répétée peut finir par fonctionner de manière automatique comme une seule unité sémantique, le sens de chaque mot de cette séquence finissant par se perdre au profit du sens global. Ce phénomène se retrouve dans presque toutes les langues du monde et environ 30% des mots d'un texte appartiennent à des expressions dites « figées » ou « polylexicales ». C'est pourquoi le figement a été beaucoup étudié par les linguistes. Cependant, le traitement de ces expressions représente un véritable défi pour le traitement automatique du langage, du fait notamment de la « non-compositionnalité » d'une grande partie d'entre elles (le sens de chaque mot pris séparément ne renseigne pas sur le sens de l'expression) qui peut être à l'origine, par exemple, d'une mauvaise traduction automatique (car littérale), de réponses non pertinentes en recherche d'information, ou encore d'un mauvais calcul de la valence en fouille d'opinion.

Démarche

L'avantage de notre travail comparé à un travail classique d'extraction d'expressions polylexicales est que nous ne serons probablement pas confrontés au problème

de non-compositionnalité cité plus haut. En effet, il s'agit d'une caractéristique propre aux expressions ayant atteint un fort degré de figement. Or étant donné que nous recherchons la naissance de figements lexicaux, ces derniers seront très probablement encore décomposables. En revanche, la difficulté sera de délimiter la frontière entre simple cooccurrence et véritable expression figée. Il nous faudra également décider du traitement des collocations. Ces trois notions de cooccurrences, collocation et expressions polylexicales devront être définies avec précision. Il est en effet primordial de connaître la nature exacte et les caractéristiques propres des unités lexicales que nous souhaitons extraire.

Une fois ces définitions posées et après avoir pris connaissance des différents travaux existants en lien avec notre projet, nous mettrons en place une méthode d'extraction non supervisée et pouvant être réutilisée pour des corpus de thèmes et de genres aussi variés que possible. Nous testerons notre méthode sur deux corpus de presse, centrés sur des sujets de société ayant donné lieu à une certaine intensité éditoriale qui, nous l'espérons, aura engendré une certaine intensité terminologique pouvant être à l'origine de figements. L'évaluation se fera manuellement sur un échantillon des candidats préalablement extraits, ainsi qu'en comparant les résultats de notre méthode avec ceux d'autres méthodes, à savoir les formules statistiques standards ainsi qu'une méthode exposée dans l'état de l'art et dont nous nous sommes inspirés. Cela nous permettra de savoir dans quelles mesures la méthode que nous proposons offre une réelle plus-value aux méthodes existantes.

Première partie

Contexte général

CHAPITRE 1

EXPRESSION FIGÉE : DÉFINITION

Sommaire

1.1 Ses caractéristiques

« Une expression figée est une unité phraséologique constituée de plusieurs mots, contigus ou non, qui présentent un certain degré de figement sémantique, un certain degré de figement lexical et un certain degré de fixité morphosyntaxique. » Voici la définition que donne [Lamiroy, 2008]. En effet, il s'agit d'un concept difficile, voire impossible à définir avec précision tant les expressions figées peuvent revêtir des formes diverses et présenter des propriétés hétérogènes : non-actualisation d'un élément (*prendre ombrage*), non-référentialité d'un élément (*lever l'ancre*), infraction des restrictions sélectionnelles (*manger son chapeau*), traces de la langue ancienne (*entrer en lice*), impossibilité de traduire (expression anglaise *kick the bucket*, 'mourir' impossible à traduire littéralement). Ainsi la définition que donne [Savary, 2019], qui parle de l'ensemble des expressions figées comme d'un « amas hétéroclite », n'est pas beaucoup plus précise : « Combinaisons de plusieurs mots qui possèdent des propriétés irrégulières au niveau du lexique, de la grammaire, de la sémantique, etc. ».

Lamiroy n'est pas la seule à essayer de définir ce phénomène reconnu unanimement par les linguistes comme un phénomène central du langage. Rappelant les propriétés considérées comme essentielles que sont la **non-compositionnalité du sens** (on ne peut déduire le sens de l'expression à partir de la somme des sens de ses constituants : *tenir la dragée haute*), la **non-substituabilité paradigmatique** (on ne peut remplacer un des mots de l'expression par un mot de sens équivalent : **avoir du fer dans l'aile*) et la **non-modifiabilité** (on ne peut jouer avec les caractéristiques morphosyntaxiques de l'expression, comme par exemple changer le singulier en pluriel : **avoir du plomb dans les ailes*), elle en montre cependant les faiblesses.

En effet, aucun de ces critères n'est nécessaire ni suffisant pour considérer une expression comme figée : certaines expressions les présentent tous tandis que d'autres n'en présentent qu'un ou deux voire aucun. D'autre part, chacun de ces critères s'avère en réalité non spécifique aux expressions figées et leur présence ne peut déterminer à elle seule le caractère figé d'une expression. Pour ce qui est par exemple de l'opacité sémantique, elle repose assez souvent sur le fait que beaucoup d'expressions

se basent sur le procédé de la métaphore (*apporter sa pierre à l'édifice*, *avoir le bras long*). Or comment établir une frontière précise entre expression figée et simple métaphore ? Cette frontière s'avère délicate à trouver comme en témoignent les désaccords entre linguistes sur les expressions métaphoriques qu'ils considèrent ou non comme figées (*se changer les idées*, par exemple). La différence entre figement et collocation est également impossible à établir sur la seule base du phénomène de solidarités lexicales : l'influence d'un mot sur les mots qui l'entourent n'est en effet pas propre au phénomène de figement lexical. Enfin, les contraintes morphosyntaxiques concernent également la phrase libre dans la mesure où les propriétés lexicales conditionnent la grammaire.

[Mejri, 2013] remet quant à lui en question le critère de l'opacité sémantique et considère comme caractéristique essentielle d'une expression figée sa capacité à être défigée. Dans son article, il définit le figement lexical comme un phénomène complexe impliquant toutes les dimensions linguistiques : prosodie, phonologie, morphologie, lexique, syntaxe, pragmatique, etc. L'auteur en propose une représentation sous forme de trames comportant des points de fixité relatifs aux niveaux d'analyse. Le défigement serait alors appréhendé en termes de désactivation de ces points, désactivation qui pourrait être partielle ou totale.

1.2 Cooccurrences, collocations et expressions figées

Lorsque l'on parle d'expressions figées, il convient de définir les notions proches de **cooccurrences** et de **collocations**, afin de préciser ce qui les en différencie.

Les collocations comme les expressions figées représentent des cas particuliers de cooccurrences. En effet, une cooccurrence n'est rien d'autre qu'une combinaison fréquente de mots. La définition de cette notion ne se base que sur des aspects purement statistiques, ce qui en fait une notion assez simple à appréhender. Notre travail consistera justement à exploiter d'autres critères que les critères statistiques afin de ne pas extraire de simples cooccurrences, ce qui donnerait un rappel élevé mais une précision médiocre.

La collocation est, comme le rappelle Lamiroy dans l'article que nous avons déjà cité, un cas particulier de cooccurrence binaire constituée d'une base, qui conserve généralement son sens habituel, et d'un collocatif dont l'usage est conditionné par la base et qui présente une certaine idiomaticité. Par exemple, dans la collocation *peur bleue*, l'adjectif est appelé par le nom *peur* (on ne dira pas **une frayeur bleue*) et n'a son sens intensif qu'avec ce nom. D'un point de vue logique, le collocatif d'une collocation est considéré comme un prédicat et la base son argument.

D'après cette définition communément admise par les linguistes, il nous semble difficile de distinguer collocations et expressions figées. En effet, comme nous l'avons déjà évoqué, la solidarité lexicale, considérée comme une des caractéristiques majeures des expressions figées, concerne également les collocations (*peur bleue* mais **frayeur bleue*). De plus, l'opacité sémantique se retrouve dans un certain nombre de collocations (un apprenant en français ne connaissant pas l'expression peinera à deviner le sens de *bleu* dans *peur bleue* de même qu'il lui sera difficile de comprendre le sens d'une collocation comme *nuit blanche*). La principale différence d'après nous tient au fait qu'une collocation nous semble généralement contenir deux unités sémantiques distinctes (*peur* et *bleue* représentent chacun une unité sémantique, *peur* étant un synonyme de *frayeur* et *bleue* un intensif) là où une expression figée est souvent considérée comme une seule unité sémantique (en réalité, nous verrons plus

loin que beaucoup d'expressions figées sont sémantiquement décomposables). Malgré cette différence (qui comme on l'a évoqué n'est pas un critère infaillible de différenciation), il nous semble pertinent de traiter les collocations comme des expressions figées et de chercher à extraire indifféremment les deux. En effet, au-delà de la proximité des deux notions, dans le cadre de notre objectif d'analyse de l'évolution du traitement d'un thème dans un corpus diachronique, la collocation nous semble un phénomène tout aussi intéressant que l'expression figée.

1.3 Les différents types d'expressions figées

Un concept important pour décrire les expressions figées est la notion de degré de figement. En effet, le figement est un phénomène de créativité lexicale progressif qui s'opère dans le temps. Ainsi Mejri considère la mesure du degré de figement d'une expression comme une problématique centrale et plus importante que la détection de ces expressions elle-même. De nombreux auteurs s'accordent à distinguer plusieurs types d'expressions polylexicales en fonction de leur degré de figement. Lamiroy, reprenant [Manguin and François, 2006], parle d'**expressions figées, semi-figées et quasi-figées**.

[Baldwin et al., 2003], dans un article que nous détaillerons plus loin, distinguent également trois types d'expressions figées en fonction de leur décomposabilité : **non-decomposable**, **idiosyncratically decomposable** et **simple decomposable** (ou **institutionalised**). Les expressions non décomposables sont caractérisées par leur opacité sémantique et la seule variation syntaxique possible est l'inflexion verbale et la réflexivisation pronominale. Quant aux expressions dites « idiosyncratiquement décomposables », elles apparaissent décomposables dans le sens où l'on peut associer un sens à chacun des mots d'une telle expression, cependant les mots en question ne peuvent se voir attribuer ce sens qu'au sein de ladite expression (par exemple « avoir d'autres *chats* à *fouetter* » pour « avoir d'autres *choses* à *faire* »). Elles sont également plus libres de varier syntaxiquement. Pour finir, les expressions institutionnalisées sont sémantiquement transparentes et présentent un degré élevé de variabilité syntaxique. Ce qui distingue ces dernières de simples collocations, d'après les auteurs (mais il nous semble que le terme *collocation* est plutôt utilisé ici dans le sens de *cooccurrence*), est leur capacité à bloquer toute expression alternative susceptible d'exprimer la même chose (l'article reprend le terme d'**anti-collocation** défini par [Pearce and Qh, 2001]).

Enfin, [Sag et al., 2002] adaptent dans leur article la terminologie de [Bauer, 1983] en divisant les expressions figées en deux groupes principaux basés sur leur degré de figement, qu'ils nomment respectivement **lexicalized phrases** et **institutionalized phrases**. La première catégorie (expressions lexicalisées) est à son tour divisée en trois groupes distincts : **fixed expressions**, **semi-fixed expressions** et **syntactically-flexible expressions**. Les expressions semi-figées comprennent les idiomes non décomposables, les noms composés et noms propres, tandis que les expressions à la syntaxe flexible sont les constructions verbales à particule, les idiomes décomposables et les expressions à verbe support. À noter que le statut d'expression figée de ces dernières est discuté : Mejri explique en effet qu'il s'agit de constructions à structure binaire où l'élément verbal ne joue que le rôle d'un actualisateur de l'élément nominal qui est en réalité un nom prédicatif. Aussi exact que nous semble cette définition, nous considérons que cela ne remet pas en cause leur statut

d'expressions figées puisque le verbe support associé à un nom ne peut en aucun cas être remplacé par un autre : il y a donc bien figement.

Nous pouvons rapprocher cette tentative de classement des expressions lexicales de la notion de **classes naturelles**, utilisée principalement en phonologie. En effet, les sons sont regroupés dans des classes dites « naturelles » en fonction de traits qu'ils partagent et chaque classe est soumise à des règles qui lui sont propres. Définir des règles applicables au type d'expression que l'on souhaite extraire permettra une meilleure automatisation de leur extraction. Ainsi, par exemple, comme nous l'avons évoqué, les expressions classées comme « non décomposables » ont comme trait commun leur opacité sémantique et comme règle commune découlant de ce trait commun l'impossibilité de varier syntaxiquement autrement que par l'inflexion verbale ou la réflexivisation pronominale : pour vérifier qu'une expression appartient à cette classe, on pourra donc vérifier qu'elle ne connaît pas d'autres variations syntaxiques que celles-ci.

1.4 Conclusion : que cherchons-nous finalement ?

En ce qui nous concerne, nous serons probablement amenés à rencontrer majoritairement des expressions « institutionnalisées ». En effet, le figement est un phénomène progressif qui passe par tous les stades énumérés plus haut (même si le découpage peut différer d'un auteur à l'autre). Dans la mesure où nous recherchons les nouveaux figements, il nous faudra détecter le moment où une cooccurrence peut être considérée comme figée et nous aurons donc à traiter des expressions au degré de figement minimal, c'est-à-dire ce que les auteurs appellent *expressions institutionnalisées*. Selon la définition de [Baldwin et al., 2003], il nous faudra donc trouver une méthode qui nous permette de repérer la disparition des anti-collocations des expressions que nous aurons retenues comme possibles expressions figées via une première sélection de cooccurrences basée sur des critères statistiques.

CHAPITRE 2

EXTRACTION D'EXPRESSIONS POLYLEXICALES : ÉTAT DE L'ART

Sommaire

2.1 Introduction

Comme nous l'avons évoqué précédemment, la reconnaissance d'expressions figées représente un défi pour le domaine du TAL. C'est pourquoi de nombreuses approches ont été expérimentées, basées sur les différentes caractéristiques connues des expressions figées (sémantiques, lexicales, morphosyntaxiques) et sur différents types de méthodes (symboliques, statistiques, mixtes). En effet, comme nous l'avons vu dans le chapitre précédent, les expressions figées partagent un certain nombre de caractéristiques, qui couvrent différents domaines.

Tout d'abord, un certain nombre d'expressions (dont le degré de figement est le plus élevé) sont sémantiquement opaques : le sens de l'expression ne correspond pas à la somme des sens de chacun de ses termes (*pousser le bouchon*).

Ensuite, il existe pour toutes les expressions un phénomène de solidarité lexicale : deux termes seront utilisés ensemble de façon préférentielle, chacun ne pouvant pas être remplacé par un synonyme (**triste comme les pierres*, **malheureux comme les cailloux*).

Enfin, des contraintes morphosyntaxiques s'appliquent aux expressions figées, plus ou moins fortes selon le degré de figement de l'expression qui pourra ou non varier en nombre, admettre des insertions (*donner beaucoup de fil à retordre* mais **avoir beaucoup de plomb dans l'aile*), etc.

Après avoir fait le point sur les différentes mesures traditionnelles d'associations lexicales, nous passerons en revue un certain nombre de méthodes intéressantes qui constitueront une base de réflexion pour notre travail.

2.2 Mesures statistiques d'association lexicale

2.2.1 Introduction

Il existe de nombreuses mesures permettant de repérer les cooccurrences significatives, susceptibles de constituer des expressions figées, que nous allons exposer ici. Nous nous appuierons sur [Manning and Schütze, 1999], qui exposent de façon très claire et détaillée la plupart de ces mesures dans leur ouvrage.

2.2.2 Fréquence

La méthode la plus simple pour repérer les collocations consiste à compter le nombre d'occurrences de chaque n-gramme et de conclure au figement pour les n-grammes les plus fréquents. S'il peut s'agir d'un point de départ pour sélectionner les candidats au figement, cette méthode est bien trop simple et ne prend pas en compte de nombreux facteurs comme la fréquence de chaque terme pris séparément. En effet, deux termes très fréquents ont plus de chance d'apparaître ensemble sans pour autant constituer des collocations. Ainsi, si l'on ne filtre pas les éléments à compter, on obtiendra de nombreux n-grammes dont un ou plusieurs éléments seront des mots outils ou autres mots par ailleurs très fréquents. En filtrant grâce à une analyse morphosyntaxique sur les *patterns* les plus susceptibles de constituer des collocations, on augmentera significativement la précision. Cependant, le rappel restera à améliorer, dans la mesure où des bigrammes peu fréquents peuvent tout de même constituer des collocations. Ce qui nous intéresse en effet pour repérer une collocation est de savoir si les termes d'un bigramme sont associés plus souvent que le hasard ne le voudrait. On considère que la probabilité de rencontrer un bigramme particulier est le produit des probabilités de rencontrer chacun de ses termes. Pour détecter des bigrammes dont la fréquence d'apparition dépasse de façon significative cette probabilité, il existe plusieurs mesures statistiques. Nous allons tenter d'expliquer certaines d'entre elles comme l'information mutuelle, le test t, le test du χ^2, ou encore la fonction de vraisemblance.

2.2.3 Information mutuelle

Une des mesures les plus fréquemment mentionnées et les plus simples pour repérer les collocations dans un texte est l'**information mutuelle ponctuelle** (*pointwise mutual information*) qui permet de déterminer la corrélation entre deux termes

en comparant leur fréquence de cooccurrences par rapport à la probabilité que ces termes apparaissent ensemble par chance. La formule est la suivante :

$$\text{PMI(x,y)} = log_2 \frac{\text{P(x,y)}}{\text{P(x)P(y)}} \tag{2.1}$$

Le score est égal à 0 si les deux termes X et Y sont complètement indépendants. Les problèmes pouvant se poser avec cette mesure sont notamment la nécessité d'avoir un corpus de grande taille et le cas de mots rares pour lesquels le score sera surévalué. Ce dernier problème peut être réglé en augmentant la probabilité d'apparition de ces mots ou en appliquant la mesure qu'aux mots dont la fréquence dépasse un certain seuil. De plus, l'objectif de cette mesure est de calculer le degré d'information qu'apporte la présence d'un mot à côté d'un autre et ce n'est pas toujours adapté à la tâche de recherche de collocation. C'est pourquoi les mesures qui suivent lui sont de plus en plus préférées.

2.2.4 Test t

Le **test de Student** ou **test t** compare la moyenne d'un ensemble avec celle d'un échantillon (en prenant en compte sa variance) dont on souhaite évaluer la représentativité. Sa formule est :

$$\text{t} = \frac{\overline{x}\text{-}\mu}{\sqrt{\frac{s^2}{\text{N}}}} \tag{2.2}$$

$\overline{x}$, s^2 et N sont respectivement la moyenne, la variance et la taille de l'échantillon tandis que μ est la moyenne attendue. Au-delà d'un certain seuil (fixé en général à 2.576), l'échantillon est considéré comme non représentatif. Pour appliquer cette formule à la recherche de collocation, on considère le nombre de bigrammes d'un texte comme la taille N de l'échantillon : $\overline{x}$ est la fréquence à laquelle les deux termes apparaissent ensemble divisée par N tandis que μ est la probabilité qu'ils apparaissent ensemble s'ils sont indépendants (c'est-à-dire P(w1) x P(w2)). La variance s^2 est égale à μ.

Le test de t peut également être utilisé pour trouver les cooccurrents permettant de distinguer le mieux les sens de deux quasi-synonymes. Il a été critiqué du fait qu'il présuppose une distribution normale (ou gaussienne, c'est-à-dire symétrique autour de la moyenne) des probabilités.

2.2.5 Test du χ^2

Le **test du χ^2 de Pearson** ou test du χ^2 d'indépendance est un test statistique qui compare la répartition d'un ensemble de données entre différentes catégories avec la répartition attendue si ces données étaient indépendantes. En d'autres termes, il s'agit pour la détection des collocations de répartir l'ensemble des bigrammes du texte en fonction de la présence d'un ou plusieurs des termes dont on souhaite mesurer l'associativité, ou de leur absence, puis comparer cette répartition avec celle que le texte devrait montrer en cas d'indépendance totale de ces termes. La formule est la suivante :

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.3}$$

Les lettres i et j font référence aux coordonnées dans un tableau qui représenterait le nombre de bigrammes pour chaque configuration possible (présence ou absence

du premier terme en colonnes et présence ou absence du second en lignes). *O* représente les données observées et *E* les données attendues en cas d'indépendance des deux termes. [Manning and Schütze, 1999] donnent un exemple de tableau pour le bigramme *new company* que nous reproduisons ici (table 2.1).

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g. *old companies*)
$w_2 \neq companies$	15820 (e.g. *new machines*)	14287181 (e.g. *old machines*)

TABLE 2.1 – Tableau pour calcul du score χ^2 de *new company* ([Manning and Schütze, 1999])

Pour le calcul des données attendues, on utilise la loi de probabilité marginale en partant du nombre total de bigrammes dont le premier terme correspond à celui de notre candidat à la collocation (dans le tableau ci-dessus, 8 + 15 820 = 15 828) et du nombre total de bigrammes dont le second terme correspond au second terme de notre potentielle collocation (8 + 4667 = 4675) convertis en proportions. Ainsi, on aura par exemple 5.2 comme nombre attendu d'occurrences du bigramme *new company* si les deux termes sont indépendants : $\frac{15828}{14307676} * \frac{4675}{14307676} * 14307676$ (14 307 676 étant le nombre total de bigrammes).

La formule ci-dessus peut être simplifiée ainsi lorsque l'on cherche à évaluer une collocation de deux termes uniquement :

$$\chi^2 = \frac{\mathbf{N}(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \tag{2.4}$$

Un résultat supérieur à 3.841 sera en faveur d'une collocation.

Le test du χ^2 peut également être utilisé pour détecter des paires de traduction dans des corpus parallèles ou comme mesure de similarité entre deux corpus. Son utilisation est déconseillée pour des données trop réduites.

2.2.6 Fonction de vraisemblance

Le **log de la fonction de vraisemblance** (*log-likelihood* en anglais) est également fréquemment utilisé. En statistique, la vraisemblance mesure une adéquation entre la distribution observée sur un échantillon aléatoire et une loi de probabilité supposée décrire une réalité sur la population dont l'échantillon est issu. Plus la vraisemblance est proche de zéro, moins l'adéquation à la loi est bonne.

Pour déterminer si deux termes forment une collocation, on formulera deux hypothèses : les deux termes sont indépendants si la probabilité de rencontrer le second terme après le premier est la même que celle de le rencontrer après un autre mot, en revanche on a affaire à une collocation si elle lui est supérieure.

— **Hypothèse 1.** $P(w^2|w^1) = p = P(w^2|\neg w^1)$

— **Hypothèse 2.** $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

On calculera ensuite la vraisemblance de chaque hypothèse à l'aide d'une loi binomiale appliquée aux fréquences du corpus étudié. Pour rappel, une binomiale permet de calculer la probabilité de rencontrer un nombre k de succès en effectuant un nombre n de tests, la probabilité p d'obtenir un succès pour un test étant connue. Ici elle permettra de calculer les chances que l'on a d'obtenir les fréquences effectivement

observées dans le corpus pour le candidat à la collocation, sachant les fréquences de chacun de ces termes.

— **Indépendance.** $L(H_1) = b(k = c_{12}; n = c_1; p = \frac{c_2}{N})b(k = c_2 - c_{12}; n = N - c_1; p = \frac{c_2}{N})$

— **Collocation.** $L(H_2) = b(k = c_{12}; n = c_1; p = \frac{c_{12}}{c_1})b(k = c_2 - c_{12}; n = N - c_1; p = \frac{c_2 - c_{12}}{N - c_1})$

c_{12}, c_1 et c_2 désignent les fréquences respectives du bigramme complet, du premier terme et du second terme, et N le nombre total de bigrammes. Le log de la vraisemblance de la première hypothèse divisée par celle de la seconde est ensuite calculé :

$$log\lambda = log\frac{\mathrm{L}(H_1)}{\mathrm{L}(H_2)} \tag{2.5}$$

Cette méthode est particulièrement efficace pour des données peu nombreuses contrairement aux autres méthodes exposées dans cette partie.

2.2.7 Évaluation des mesures

Dans toute tâche de TAL se pose la question de l'évaluation. Dans des domaines tels que la recherche d'information ou la classification, il est important de pouvoir évaluer la pertinence et la robustesse d'une méthode. Cependant cela nécessite en général des données annotées qui font parfois défaut. Il arrive fréquemment dans ce cas que le chercheur évalue son travail sur un échantillon qu'il fait annoter dans ce but spécifique. Cela peut poser un problème de représentativité. Dans leur article, [Evert and Krenn, 2001] présentent leur propre méthode d'évaluation avec laquelle ils comparent les cinq mesures utilisées que nous venons d'exposer : fréquence, information mutuelle, test t, test du χ^2, fonction de vraisemblance.

Les cinq mesures à évaluer sont appliquées à deux ensembles de données annotées : des bigrammes lemmatisés Adjectif+Nom extraits d'un corpus de textes juridiques en allemand et des triplets préposition-nom-verbe extraits du *Franckfurter Rundschau Corpus* (préposition et nom doivent se trouver dans le même syntagme et le verbe doit être dans la même phrase). Pour chaque mesure, les expressions sont triées du plus haut au plus bas degré de collocativité, puis l'évaluation est présentée sur des graphiques dont l'ordonnée représente la mesure d'évaluation (précision ou rappel) et l'abscisse le nombre de résultats sur lesquels elle est calculée. Les auteurs constatent que la meilleure précision est donnée par le **log de la fonction de vraisemblance** pour les bigrammes (voir figure 2.1 ci-dessous) et par le **test de t** pour les triplets. Pour toutes les mesures, la précision chute dans la première moitié des résultats tandis que le rappel augmente.

Les auteurs ont également souhaité comparer les résultats en fonction de la fréquence des candidats. Les mesures les plus performantes demeurent les mêmes pour les candidats très fréquents tandis que toutes les mesures se valent pour les candidats peu fréquents. Les hapax et doublons constituent un cas à part : ils sont souvent mis de côté car les méthodes statistiques ne permettent pas de tirer des conclusions sur des données si peu fréquentes. Cependant, ils représentent la majorité des candidats et les auteurs ont souhaité savoir si cette décision de ne pas les traiter était ou non justifiée. Pour cela, ils ont estimé le nombre de vrais positifs parmi eux à partir d'un échantillon grâce à la loi binomiale. La proportion de vrais positifs étant très faible, leur conclusion est que leur mise à l'écart est justifiée.

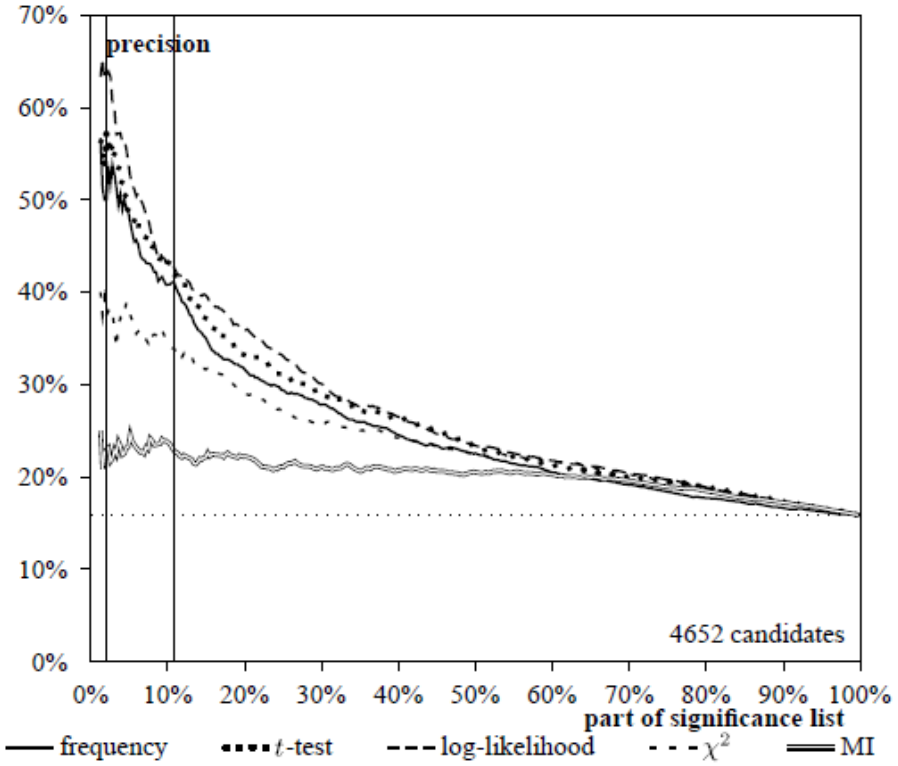

FIGURE 2.1 – Graphique montrant l'évolution de la précision des mesures statistiques de collocation ([Evert and Krenn, 2001])

2.2.8 Conclusion sur les mesures standards et transition vers les autres méthodes

En conclusion, les mesures statistiques standards constituent une première approche acceptable pour détecter les collocations, avec des résultats corrects. Cependant, de nombreux chercheurs ont constitué leurs propres méthodes, basées sur les caractéristiques des expressions figées que nous avons déjà mentionnées, dans l'espoir d'obtenir de meilleurs résultats. Nous allons donc en détailler quelques unes que nous pensons particulièrement intéressantes. Nous commencerons par les méthodes s'appuyant sur les caractéristiques sémantiques, puis lexicales et enfin morphosyntaxiques.

2.3 Identification du figement sur des critères sémantiques

2.3.1 Calcul de similarité entre une expression et ses constituants avec analyse sémantique latente (LSA)

Dans leur article déjà cité plus haut, [Baldwin et al., 2003] cherchent à distinguer trois types d'expressions polylexicales en fonction de leur degré de figement sémantique. Partant du principe qu'une expression dite « institutionnalisée » (qui possède le degré de figement le plus bas) constitue un hyponyme de sa tête, la méthode présentée consiste à calculer la similarité sémantique entre une expression

et ses constituants. La méthode est testée sur un corpus de noms composés extraits du *Wall Street Journal* et de verbes à particule extraits du *British National Corpus* et utilise le procédé d'**analyse sémantique latente** (LSA pour *Latent Semantic Analysis*). Ce procédé a l'avantage d'être indépendant de la langue et de la composition syntaxique des données d'entrée. Développé à l'origine pour la recherche d'informations, il consiste à représenter les mots comme des points dans un vecteur spatial : les auteurs ont ainsi créé une matrice permettant de représenter les 50 000 mots ou groupes de mots (parmi lesquels figurent les candidats au figement) les plus fréquents selon la fréquence à laquelle ils cooccurrent avec une sélection de 1 000 mots pleins fréquents. Ils ont ensuite calculé la similarité cosinus entre le vecteur de chaque expression et ceux de ses constituants.

Pour rappel, la similarité cosinus est une des nombreuses manières de calculer la similarité entre deux vecteurs numériques. Elle est très utilisée en recherche d'information et permet de mesurer l'angle entre deux vecteurs. Sa formule est la suivante :

$$\frac{A * B}{\|A\| * \|B\|} \tag{2.6}$$

Dans cette formule, $A * B$ correspond au produit scalaire des vecteurs $\vec{A}$ et $\vec{B}$ (somme des produits des valeurs des deux vecteurs prises deux à deux), et $\|A\|$ et $\|B\|$ la norme des deux vecteurs (la norme d'un vecteur est la racine carrée de la somme de ses éléments élevés au carré).

En calculant la similarité entre des vecteurs représentant chacun le contexte lexical d'un mot ou d'une expression, on peut se faire une idée de la similarité sémantique existant entre les mots ou expressions ainsi représentés. En effet, comme l'a écrit John Rupert Firth en 1957, « You shall know a word by the company it keeps » (*Vous connaîtrez un mot par ses fréquentations*). Cette technique est basée sur l'hypothèse dite « de Harris » (ou *distributional hypothesis*), qui veut que les mots apparaissant dans des contextes similaires ont des significations apparentées, et qui constitue la base de ce qu'on appelle les plongements lexicaux (ou *word embeddings*).

L'évaluation a été faite à l'aide de Wordnet : la distance sémantique entre deux mots sur Wordnet a d'abord été calculée en fonction de la distance relative de leur sens dans le lexique. Cependant, la corrélation entre les similarités de la LSA et celles obtenues avec Wordnet s'est avérée très basse et une autre méthode d'évaluation a finalement été suivie : trier et partitionner les expressions en fonction de leur degré de similarité avec leurs constituants et vérifier que les partitions à la similarité la plus haute ont le plus grand nombre d'expressions présentées dans Wordnet comme hyponymes d'un de leurs constituants. Cette seconde évaluation a donné des résultats encourageants uniquement pour les expressions peu fréquentes. Il est possible, d'après les auteurs, que ces résultats peu concluants s'expliquent par la polysémie des expressions plus fréquentes ou par les imprécisions de Wordnet.

2.4 Identification du figement sur des critères lexicaux

2.4.1 Critères distributionnels pour détection d'hapax

Ces résultats encourageants sur des expressions peu fréquentes auraient cependant pu intéresser [Lapata and Lascarides, 2003] qui ont proposé une méthode

permettant de détecter des noms composés constituant des hapax. En effet, partant de la constatation que les méthodes statistiques souvent utilisées sont incapables de détecter des mots composés rares mais valides (puisque ces méthodes reposent sur une fréquence élevée de l'expression par rapport à celle de chacun de ses termes), ils proposent de dégager à partir de ces méthodes les tendances distributionnelles des noms composés lexicalisés (et donc fréquents) afin de classifier les hapax avec un algorithme de *Machine Learning*. L'intérêt de proposer une telle méthode tient au fait que les noms composés constituent un type d'expressions particulièrement productif et les hapax sont nombreux : il semble donc important d'être en mesure de les repérer. Voici dans le détail la méthode utilisée dans l'article.

Suivant l'heuristique de [Lauer, 1995], tous les bigrammes de noms qui ne sont ni précédés ni suivis d'un nom sont extraits d'une version POS-taguée du *British National Corpus*, puis on évalue ensuite un échantillon comprenant des bigrammes classés comme noms composés et d'autres écartés, avec comme référence une annotation manuelle effectuée à l'aide d'un concordancier. La précision est bonne pour les noms composés fréquents mais mauvaise pour les hapax qui représentent pourtant la moitié des noms composés valides. On modélise alors les tendances distributionnelles des noms composés fréquents pour classifier les hapax.

Les paramètres passés au classifieur sont au nombre de quatre : trois paramètres numériques et un paramètre contextuel. On calcule d'abord la fréquence du nom qui gouverne le bigramme en tant que tête de syntagme, puis on fait de même pour le modifieur (fréquence à laquelle le nom est trouvé comme modifieur dans les noms composés fréquents). À partir de ces fréquences, on calcule les probabilités. Le troisième paramètre s'appuie sur la notion de **règles lexicales** : les mots sont remplacés par les concepts qui leur sont associés dans une taxonomie et on calcule la fréquence parmi les noms composés lexicalisés du couple de concepts. Les taxonomies utilisées sont **Wordnet** et Roget. Lorsqu'un mot est associé à plusieurs concepts (on trouve en moyenne 11.5 concepts par mot dans Wordnet), on modélise toutes les combinaisons de concepts possibles et on calcule la moyenne des fréquences de chaque combinaison. On pondère également la fréquence de chaque nom composé lexicalisé correspondant à une des combinaisons en divisant sa fréquence par le nombre de combinaisons de concepts que lui-même peut recouvrir. Enfin, le dernier paramètre est la prise en compte du contexte via l'encodage de l'étiquette morpho-syntaxique (*part-of-speech*) et de la position des mots précédents et/ou suivants.

L'outil **Weka** est utilisé pour classifier les hapax à l'aide d'une partie ou de tous les paramètres précédemment énumérés. L'entraînement et l'évaluation s'effectuent sur un corpus de 1000 candidats annotés manuellement par deux personnes et répartis en dix strates (*folds*) pour une validation croisée. Deux algorithmes sont utilisés et comparés : arbre de décisions et Naive Bayes. Les résultats sont très satisfaisants pour les deux algorithmes. Naive Bayes atteint sa meilleure performance (précision de 72.3%) avec trois paramètres : fréquences des concepts avec Wordnet, probabilité que la tête de l'hapax soit la tête d'un nom composé, fenêtre contextuelle de taille 1 de chaque côté. L'arbre de décision atteint quant à lui sa meilleure performance avec quasiment tous les paramètres réunis (précision de 72%). Les auteurs concluent sur leur souhait d'approfondir les critères conceptuels et contextuels et de tester d'autres algorithmes de classifications tels que SVM ou AdaBoost.

2.4.2 Calcul de solidarité lexicale par clustering distributionnel

Faisant l'hypothèse qu'une des caractéristiques principales des expressions figées est l'impossibilité de remplacer un de ses termes par un synonyme, [Van de Cruys and Villada Moirón, 2007] proposent une méthode calculant les préférences lexicales d'une expression pour un nom par rapport aux noms de sens proches, rassemblés grâce à un **clustering** effectué à partir de critères distributionnels. L'objectif des auteurs est de pouvoir alimenter les lexiques avec les expressions figées extraites grâce à cette méthode non supervisée et applicable à tous les types d'expressions.

Les candidats aux figements choisis pour l'expérience sont des syntagmes verbaux de type verbe + groupe prépositionnel extraits d'un corpus néerlandais parsé, le *Twente Nieuws Corpus*, et sont représentés dans une matrice de 5 000 combinaisons les plus fréquentes de verbe + préposition sur 10 000 noms les plus fréquents du corpus.

Pour effectuer le clustering de noms, la méthode consiste à extraire des triplets représentant des relations de dépendance (de type <*pomme, objet, manger*>) et à générer pour chacun des 10 000 noms un vecteur donnant le score de PMI (*pointwise mutual information*, une des mesures statistiques que nous avons décrites plus haut, qui calcule les chances qu'ont des termes d'apparaître ensemble par rapport à leur chance d'apparaître séparément) entre le nom et chacune des 100 000 relations de dépendance extraites. Le clustering est appliqué à ces vecteurs grâce à l'algorithme de partitionnement de données K-moyennes (plus connu sous son nom anglais **k-means**) et génère 1 000 clusters.

Cinq mesures sont alors utilisées pour calculer la solidarité lexicale d'une expression. Les trois premières mesures permettent de calculer la préférence d'un verbe pour un nom à partir de la divergence entre la probabilité P(N) de rencontrer le nom et la probabilité P(N | V) de rencontrer le nom avec le verbe : plus le ratio de la préférence d'un nom par rapport aux autres noms du cluster auquel il appartient est proche de 1, plus la probabilité est élevée pour que l'expression soit lexicalement figée. Les deux dernières mesures calculent à l'inverse la préférence d'un nom pour un verbe selon le même procédé.

L'évaluation est effectuée sur les expressions figées présentes dans les ressources lexicographiques (*Referentie Bestand Nederlands* et *Van Dale Lexicographical Information System*) et compare les résultats pour plusieurs mesures (et plusieurs seuils pour chaque mesure) afin de déterminer la configuration optimale (la préférence du verbe pour le nom donne de meilleurs performances mais celle du nom pour le verbe permet d'exclure les simples collocations). Une comparaison est également effectuée avec les résultats du travail de [Fazly and Stevenson, 2006], qui consistait à comparer la PMI d'une expression avec la PMI de la même expression dans laquelle les mots avaient été remplacés par des synonymes. Le travail de Van de Cruys et Moirón obtient une meilleure précision mais un rappel inférieur. Les résultats peuvent être biaisés par la qualité du clustering et la syntaxe (des restrictions grammaticales peuvent engendrer une surévaluation du degré de figement), et le clustering ne permet pas de prendre en compte la polysémie car un mot ne peut figurer que dans un cluster. En conclusion, les auteurs conseillent de combiner cette méthode avec d'autres méthodes prenant en considération la syntaxe ou avec des classifieurs, d'utiliser des mesures de similarité à la place d'un clustering et d'effectuer une évaluation

manuelle du travail.

2.5 Identification du figement sur des critères morphosyntaxiques

2.5.1 Calcul des préférences morphosyntaxiques

Si beaucoup d'articles traitant de la détection d'expressions figées choisissent une approche sémantique ou lexicale, il en est qui abordent le problème en s'intéressant au figement morphologique de ces expressions. C'est le cas de [Evert et al., 2004], qui s'intéressent aux préférences morphosyntaxiques dans les collocations (préférence du singulier sur le pluriel, par exemple), notamment dans les langues possédant une morphologie flexionnelle riche comme l'allemand. L'analyse statistique des cooccurrences de lemmes est selon eux insuffisante tandis que les préférences morphosyntaxiques constituent un fort indice du degré de figement d'une expression.

Pour effectuer ce travail, les auteurs ont extrait des paires adjectif + nom dans un large corpus de presse allemande (300 millions de mots) tokenisé, POS-tagué, lemmatisé avec Tree-Tagger et chunké avec YAC, en se concentrant sur quatre mots : Tag, Zeit, Schritt et Kraft. Ils comptent ensuite le nombre total de cooccurrences quelle que soit la forme prise par les constituants du bigramme puis pour chaque bigramme la fréquence de chaque cas ou nombre (une loi binomiale est appliquée afin de pondérer cette fréquence en fonction de la fréquence globale du cas ou du nombre concerné). Une préférence marquée est signe d'une idiomatisation de la collocation.

Un des problèmes rencontrés par les auteurs est l'ambiguïté du cas (une forme peut correspondre à plusieurs cas) qui concerne 80% du corpus malgré une désambiguïsation partielle effectuée grâce à l'accord nom-adjectif. Dans ce cas, des classes ambiguës ont été ajoutées sous la forme d'ensembles de valeurs.

Les auteurs conseillent de combiner cette méthode avec des méthodes standards d'identification de collocations, soit pour identifier préalablement des candidats soit au contraire pour identifier les collocations à partir de candidats déjà sélectionnés. Des critères comme le caractère défini ou indéfini de l'article peuvent être ajoutés à ceux du cas et du nombre.

2.5.2 Utilisation d'une grammaire formelle

Parmi les articles les plus importants sur le sujet des expressions figées figure l'article de [Sag et al., 2002] déjà cité précédemment. Les auteurs, après avoir fait le point sur les différents types d'expressions figées et les méthodes utilisées pour les traiter ainsi que leurs défauts, y proposent une nouvelle méthode basée sur une **grammaire HPSG** (*head-driven phrase structure grammar*). La grammaire HPSG est un type de grammaire d'unification (les grammaires d'unification sont des grammaires formelles apparues à partir des années 1980 pour analyser le langage naturel) utilisant une structure de traits, c'est-à-dire que chaque élément du lexique est représenté par un ensemble de paires attribut-valeur qui donne en quelque sorte le programme linguistique du mot. L'objectif à terme est, après avoir testé la pertinence de la grammaire ainsi construite, de l'incorporer à la LinGO ERG (English Resource Grammar), une grammaire de la langue anglaise à large

couverture.

Les expressions complètement figées, qui n'admettent aucune variation syntaxique ou modification interne, peuvent être tout simplement traitées comme des **mots avec espaces**. En termes de grammaire HPSG, cela signifie qu'une liste de mots se voit attribuer le même type que des mots simples (par exemple l'expression « ad hoc » et l'adjectif « pretty » ont tous les deux le type *intransitive adjective*).

Concernant les expressions semi-figées, un attribut est ajouté afin d'indiquer les positions des constituants de l'expression susceptibles de varier syntaxiquement. D'autres mécanismes un peu complexes sont utilisés, notamment pour les expressions à la syntaxe flexible (plus complexes à traiter), que nous ne détaillerons pas ici, car nous ne retiendrons pas cette méthode pour notre travail. En effet, il s'agit d'un travail long et fastidieux que nous n'aurons pas le temps de mener à bien dans le temps imparti pour la rédaction de ce mémoire, et qui semble plus adapté à l'extraction d'expressions connues qu'à la détection de nouvelles expressions en formation.

2.6 Conclusion de l'état de l'art

Nous avons pu voir que les méthodes de détection d'expressions figées étaient extrêmement variées, tout en s'appuyant toujours sur les caractéristiques de figement que nous avons énumérées au chapitre 1 (compositionnalité du sens, non-substituabilité paradigmatique, non-modifiabilité). Toutes les approches ne seront pas pertinentes pour notre sujet puisque, comme nous l'avons déjà évoqué, nous rechercherons principalement des expressions au degré de figement peu élevé, voire très bas. Elles constituent néanmoins des pistes très intéressantes. Les méthodes sémantiques sont sans doute à écarter (l'opacité sémantique sera probablement rare pour les expressions que nous chercherons à extraire, puisqu'il s'agit d'une caractéristique d'un degré de figement avancé), tandis que les méthodes s'appuyant sur la solidarité lexicale seront à privilégier : en effet, comme nous l'avons déjà constaté plus haut, la disparition des « anti-collocations » semble être la première étape du figement.

Deuxième partie

Expérimentations

CHAPITRE 3

PRÉSENTATION DES DONNÉES : DEUX CORPUS DE PRESSE DIACHRONIQUES

Sommaire

3.1 Introduction : choix et origine des corpus

Comme nous l'avons évoqué en introduction, un des objectifs de ce travail est de fournir une méthode non supervisée permettant à des chercheurs en sciences humaines d'obtenir des pistes d'analyses concernant les larges corpus diachroniques qu'ils sont amenés à traiter. Nous évaluerons donc notre méthode sur deux corpus de tailles différentes et portant sur des thèmes différents. Les deux corpus ont été constitués à partir d'une recherche de mot clé sur le site *Europresse*. Ils font tous deux l'objet d'un sujet de recherche en sciences humaines à la Maison de la Recherche de la Sorbonne et m'ont été fournis dans le cadre du stage que j'ai effectué là-bas. Il s'agit de corpus d'articles de presse en français. Certains articles (minoritaires) ont fait l'objet d'une océrisation.

3.2 Premier corpus : le corpus *Sackler*

Le premier corpus porte sur la famille Sackler et leur implication dans la crise des opioïdes aux États-Unis. Ce corpus a été constitué via Europresse [1] dans le cadre d'une collaboration avec Pierre-Marie Chauvin, maître de conférences en sociologie au GEMASS (Sorbonne Université). J'ai pu échanger avec lui pendant le stage et il s'intéresse aux Sackler dans le cadre d'une étude sur la sociologie des réputations. Le corpus *Sackler* comporte 437 articles de presse publiés de 1978 à 2021 (dont 174 en 2019, année où la presse française en a le plus parlé), provenant de 93 journaux différents. Les journaux les plus représentés sont *Art Media Agency* (39 articles), *Le Figaro* (29 articles), *Libération* (22 articles), *Le Monde* (20 articles) et *Les Échos* (19 articles). Le tableau 3.1 donne quelques statistiques sur le corpus (le nombre de

1. http://www.europresse.com

mots et la taille du vocabulaire ont été calculés à l'aide d'une tokenisation et d'une lemmatisation effectuées avec la librairie SpaCy).

Année	Nombre de documents	Nombre de mots	Nombre moyen de mots par document	Taille du vocabulaire
1978	1	1995	1995	790
1987	2	10387	5193.5	4330
1989	5	15465	3093	4275
1995	1	1409	1409	589
1996	2	270	135	146
1997	9	4664	518.22	1540
1998	4	3604	901	1247
1999	1	415	415	219
2001	1	1105	1105	559
2002	1	942	942	482
2003	1	325	325	195
2004	3	2773	924.33	1049
2005	3	1952	650.67	848
2007	3	1997	665.67	875
2008	7	4318	616.86	1446
2009	10	4381	438.1	1502
2010	8	3401	425.12	1333
2011	19	8488	446.74	2394
2012	5	1755	351	696
2013	20	7375	368.75	2162
2014	20	9977	498.85	3392
2015	18	10682	593.44	3059
2016	16	40332	2520.75	7195
2017	17	16415	965.59	3855
2018	29	27738	956.48	5124
2019	174	105640	607.13	9080
2020	29	15178	523.38	2989
2021	11	10203	927.55	2466
Total	437	327505	749.44	27735

TABLE 3.1 – Quelques chiffres concernant le corpus Sackler

On peut constater en l'observant un certain manque d'homogénéité du corpus. La taille des documents est très variable, allant de 242 mots pour l'unique article de 1996 à 5193.5 mots en moyenne pour 1987. Mais ce qui doit surtout attirer notre attention est la répartition des articles dans le temps, qui est loin d'être optimale. En effet, l'année 2019 est nettement sur-représentée tandis que certaines années sont sous-représentées voire pas du tout représentées (notamment la période allant de 1979 à 1986). Étant donnée la place importante de la diachronie dans l'objectif que nous avons défini pour notre travail, cela pourra s'avérer problématique et devra être pris en compte dans l'interprétation des résultats obtenus lors de nos différentes expérimentations.

3.3 Second corpus : le corpus *PMA*

Le second corpus porte sur la procréation médicalement assistée. Il s'agit d'un sujet étudié par une autre collègue, Virginie Julliard professeur des universités en sciences de l'information et de la communication au CELSA (Sorbonne Université). Il comporte 2 138 articles de presse publiés de 1994 à 2021 (dont 395 en 2013, année de promulgation de la loi française du « Mariage pour tous »), provenant de seulement neuf journaux différents. Les journaux les plus représentés sont *Le Figaro* (469 articles), *La Croix* (417 articles), *Le Monde* (417 articles), *Libération* (352 articles), et *Aujourd'hui en France* (256 articles).

Année	Nombre de documents	Nombre de mots	Nombre moyen de mots par document	Taille du vocabulaire
1994	3	1866	622	628
1995	3	1845	615	659
1996	2	1111	555.5	506
1997	3	2916	972	932
1998	5	2056	411.2	743
1999	7	5928	846.86	1455
2000	5	4881	976.2	1336
2001	4	2920	730	990
2002	6	4084	680.67	1159
2003	3	1503	501	553
2004	13	9850	757.69	2194
2005	6	4576	762.67	1312
2006	10	5298	529.8	1498
2007	12	9249	770.75	1931
2008	19	17902	942.21	3299
2009	6	4106	684.33	1209
2010	12	7523	626.92	1798
2011	14	10455	746.79	2131
2012	161	85527	531.22	6391
2013	395	255336	646.42	12171
2014	263	170317	647.59	10548
2015	61	36952	605.77	4617
2016	91	62579	687.68	6471
2017	221	139137	629.58	9616
2018	343	233454	680.62	12229
2019	352	257120	730.45	12293
2020	97	68856	709.86	6773
2021	20	13689	684.45	2620
Total	2138	1421755	664.99	29155

TABLE 3.2 – Quelques chiffres concernant le corpus PMA

Les mêmes remarques émises au sujet du premier corpus valent également pour ce second corpus. En effet, en observant le tableau 3.2, on constate que les années précédant 2012 comprennent un nombre très limité d'articles (moins de 20 articles

par an et même moins de 10 avant 2004), tandis que les années 2013, 2018 et 2019 en comptent plus de 300 chacune. En revanche, la taille des articles semble un peu plus homogène.

3.4 Constitution des corpus

Les deux corpus nous ont été fournis au format HTML (format de sortie des requêtes effectuées sur Europresse). Les documents HTML comportaient de nombreuses métadonnées inutiles pour notre travail et étaient difficiles à manipuler du fait de noms d'éléments peu transparents et d'un manque de cohérence de la structure d'ensemble. À l'aide de scripts Python utilisant la librairie BeautifulSoup (documentation ici : https://www.crummy.com/software/BeautifulSoup/bs4/doc/) nous avons donc converti les fichiers HTML en json, afin d'avoir des données propres et facilement exploitables, avec une structure identique pour les deux corpus.

Chaque fichier json comprend, pour chaque article, un identifiant numérique unique, le nom du ou des journaux dans lequel a été publié l'article, sa ou ses dates de publication, son titre, son contenu « brut »sans les balises HTML pour les différentes manipulations, mais aussi son contenu enrichi avec les balises HTML pour pouvoir conserver la structure visuelle notamment pour l'exploitation dans une application de parcours de corpus, notamment via *topic modeling*, que j'ai développée durant mon stage.

Dans la figure 4.6 figure un exemple d'article au format HTML produit par Europresse.

```
<article>                  <img
↪  src="./sackler_files/Docviewer(2).aspx"/><span></span>
↪  <div> <span> Libération   </span> </div> <div> <span>
↪  CULTURE,  vendredi 1 mars 1996 171 mots, p. 35 </span>
↪  </div>             <div>   <p>Un mécène pour le
↪  Louvre.</p>  </div>          <div>    <div> <p>C'est une
↪  nouveauté à plus d'un titre dans le monde des musées
↪  français que cet accord conclu hier entre le musée du
↪  Louvre et la fondation britannique Mortimer et Theresa
↪  Sackler. [...] C'est en connaissant son intérêt pour les
↪  antiquités orientales que le Louvre a approché Mortimer
↪  Sackler: avec ses deux frères, il a financé la Sackler
↪  Wing du Metropolitan Museum of Art de New York et la
↪  fondation a fait des donations en Angleterre, à la
↪  National Gallery et la Tate Gallery.</p>     </div>
↪  </div>                  <a name="0.2_complement"></a>
↪  <div></div>   <div>  <small>© 1996 SA Libération. Tous
↪  droits réservés.</small>      
↪  <small>Le présent document est protégé par les lois et
↪  conventions internationales sur le droit d'auteur et son
↪  utilisation est régie par ces lois et
↪  conventions.</small> </div>  <div> <div> <div><br/><img
↪  width="60" src="./sackler_files/Public_vert.png"
↪  border="0"/><br/></div> <div>Certificat émis le <b>14
↪  avril 2021</b> à
↪  <b>Bibliothèque-de-Sorbonne-<wbr>Université </b> à des
↪  fins de visualisation personnelle et
↪  temporaire.</div><br/> </div>
↪  <div>news·19960301·LI·44141</div> </div></article>
```

FIGURE 3.1 – Exemple d'article au format HTML extrait sur Europresse

À l'issue de notre script de *scraping* cet article sera représenté tel que présenté dans la figure 3.1.

```
"9": {
        "journal": "Libération",
        "date": "vendredi 1 mars 1996",
        "texte_html": "<p>C'est une nouveauté à plus d'un
        ↪  titre dans le monde des musées français que cet
        ↪  accord conclu hier entre le musée du Louvre et
        ↪  la fondation britannique Mortimer et Theresa
        ↪  Sackler. [...] C'est en connaissant son intérêt
        ↪  pour les antiquités orientales que le Louvre a
        ↪  approché Mortimer Sackler: avec ses deux frères,
        ↪  il a financé la Sackler Wing du Metropolitan
        ↪  Museum of Art de New York et la fondation a fait
        ↪  des donations en Angleterre, à la National
        ↪  Gallery et la Tate Gallery.</p>",
        "titre": "Un mécène pour le Louvre.",
        "texte": "Un mécène pour le Louvre. C'est une
        ↪  nouveauté à plus d'un titre dans le monde des
        ↪  musées français que cet accord conclu hier entre
        ↪  le musée du Louvre et la fondation britannique
        ↪  Mortimer et Theresa Sackler. [...] C'est en
        ↪  connaissant son intérêt pour les antiquités
        ↪  orientales que le Louvre a approché Mortimer
        ↪  Sackler: avec ses deux frères, il a financé la
        ↪  Sackler Wing du Metropolitan Museum of Art de
        ↪  New York et la fondation a fait des donations en
        ↪  Angleterre, à la National Gallery et la Tate
        ↪  Gallery."
}
```


FIGURE 3.2 – Exemple d'article converti au format json

Nous avons montré ici un extrait assez lisible du document HTML, mais de nombreux éléments facultatifs contenant diverses métadonnées peuvent figurer dans l'élément *Article*. La plupart des éléments sont simplement nommés *div* ou *span* et aucun attribut n'indique la nature du contenu, ce qui rend le document très difficile à scraper. De plus, les deux corpus n'ayant pas été extraits au même moment, la structure HTML était légèrement différente et il a fallu produire un script différent pour chaque corpus. C'est pourquoi cette partie simple du travail a requis plus de temps qu'initialement prévu.

Les corpus comportaient des doublons (un même article publié dans plusieurs journaux et/ou à plusieurs dates). Cependant, le contenu n'étant pas nécessairement identique (la ponctuation peut être différente, une phrase peut avoir été ajoutée à la fin, etc.), nous avons élaboré un script en Python permettant de repérer et supprimer ces doublons en calculant la distance de Levenshtein entre les articles en fonction de leur longueur (grâce à la librairie *Levenshtein*) : deux articles de tailles similaires (la différence en nombre de caractères ne doit pas excéder 500) et dont la distance de Levenshtein divisée par le nombre de caractères moyen des deux articles était inférieur à 0.1 ont été considérés comme des doublons.

Ainsi, 65 doublons ont été repérés dans le corpus Sackler et seulement cinq dans le corpus PMA. Lorsque des articles en doublons sont repérés, on conserve le contenu du plus ancien et on ajoute la date et le journal des plus récents à ses méta-données. Voici un exemple de résultat :


```
"36": {
        "journal": [
            "L'Histoire, no. 316",
            "Le Figaro"
        ],
        "date": [
            "lundi 1 janvier 2007",
            "samedi 24 février 2007"
        ],
        "texte_html": "<p>La Sackler Gallery et la Freer
        ↪  Gallery of Art, situées sur le National Mall, à
        ↪  Washington, jouent le rôle de musée national des
        ↪  arts asiatiques des États-Unis d'Amérique (même
        ↪  si une des deux galeries possède un fonds d'art
        ↪  américain). </p> [...] <p> Le catalogue du musée
        ↪  en ligne (situé sous la rubrique « collections
        ↪  ») est à la hauteur de l'institution : 6 000
        ↪  objets sont reproduits. </p><p><a
        ↪  href=\"http://www.asia.si.edu/\"
        ↪  target=\"_blank\">
        ↪  http://www.asia.si.edu/</a></p>",
        "titre": "Focus Merveilles d'Asie à Washington",
        "texte": "Focus Merveilles d'Asie à Washington La
        ↪  Sackler Gallery et la Freer Gallery of Art,
        ↪  situées sur le National Mall, à Washington,
        ↪  jouent le rôle de musée national des arts
        ↪  asiatiques des États-Unis d'Amérique (même si
        ↪  une des deux galeries possède un fonds d'art
        ↪  américain). [...] Le catalogue du musée en ligne
        ↪  (situé sous la rubrique « collections ») est à
        ↪  la hauteur de l'institution : 6 000 objets sont
        ↪  reproduits.   http://www.asia.si.edu/"
    }
```


FIGURE 3.3 – Exemple d'article en doublon

Le même article a été publié dans les deux journaux à deux mois d'écart. Cependant, si nous n'avions fait que vérifier si le contenu de l'élément *texte* était identique pour les deux, nous n'aurions pas pu détecter ce doublon car le *Figaro* a ajouté un titre à l'article, que la revue *L'Histoire* n'avait pas (ou du moins le titre n'était pas précisé par la sortie Europresse pour cette revue).

Ce script est utilisable sur nos corpus, mais sera beaucoup trop long sur des corpus plus conséquents (la distance de Levenshtein est calculée entre chaque document du corpus, ce qui représente une immense quantité de calculs). Il serait

donc utile d'y ajouter une étape préalable permettant d'effectuer une pré-selection de potentiels doublons via le calcul de similarité (cosinus, par exemple) entre les documents représentés sous forme de vecteurs (plongements de mots, idéalement).

Comme mentionné plus haut, certains articles ont fait l'objet d'une océrisation (reconnaissance optique de caractères). Les techniques OCR n'étant pas encore complètement au point, cela a donné lieu à des erreurs, dans la reconnaissance de caractères d'une part mais également dans la segmentation du texte en mots. On trouvera par exemple dans l'article de 1978 du corpus Sackler des phrases comme « No n au "super-parlement" réDondent 5 1 % des personnes interroaées ANS leur grande majorité ». Ce type de phrase ne pourra pas être parsé correctement par les outils que nous utiliserons et il y a de fortes chances pour que la tokenisation, la lemmatisation, le POS taggging et autres techniques de traitement automatique du langage naturel s'avèrent inefficaces sur ce type de documents.

La segmentation du contenu océrisé d'un journal en différents articles étant, elle aussi, automatique, ces documents contiennent du bruit (par exemple, des encadrés ou autres contenus se situant autour de l'article et ayant été considéré à tort comme en faisant partie).

Fort heureusement, la proportion de textes océrisés dans nos corpus est extrêmement faible (elle concerne en effet les plus vieux articles qui, comme nous l'avons vu, sont beaucoup moins nombreux).

3.5 Conclusion : des données imparfaites

Lors de l'interprétation des résultats obtenus au cours des expériences que nous détaillerons dans le chapitre qui suit, il nous faudra garder à l'esprit les différents « défauts » des corpus que nous venons d'exposer. En effet, nos corpus sont de taille réduite, présentent un fort déséquilibre temporel avec un petit nombre d'années surreprésentées et beaucoup d'années sous-représentées, et enfin les documents les plus anciens ont fait l'objet d'une océrisation donnant lieu à de nombreuses coquilles et à des mots mal segmentés susceptibles de poser problème lors du parsing du texte.

Néanmoins, il est rare d'avoir des données parfaites et il est intéressant de travailler sur de « vraies données de terrain », qui sont souvent hétérogènes et bruitées, en espérant pouvoir en retirer des informations intéressantes malgré tout.

CHAPITRE 4

APPLICATION DE TROIS MÉTHODES D'EXTRACTION D'EXPRESSIONS FIGÉES

Sommaire

4.1 Introduction : trois méthodes distinctes

Avant de proposer notre propre méthode et afin d'avoir quelques éléments de comparaison, nous avons commencé par tester quelques méthodes décrites dans l'état de l'art.

Tout d'abord, nous avons appliqué les différentes mesures statistiques que nous avons passées en revue au début du chapitre 2. Puis nous avons tenté de reproduire l'expérience de [Van de Cruys and Villada Moirón, 2007] avec quelques adaptations. En effet, comme nous l'avions évoqué en conclusion de l'état de l'art, la solidarité lexicale est probablement le critère le plus intéressant à exploiter dans le cadre de notre travail et, de tous les articles que nous avons décrits, celui-ci s'avère donc le plus pertinent pour notre tâche.

Nous présenterons ensuite notre méthode personnalisée qui utilise dans un premier temps un modèle de topic modeling pour présélectionner des candidats, puis dans un second temps s'inspire de la méthode Van de Cruys-Villada en calculant la préférence de chaque candidat sur ses potentiels synonymes.

4.2 Mesures statistiques

La librairie NLTK possède un module *collocations* qui offre la possibilité de calculer les scores de solidarité lexicale entre tous les mots d'un texte. On peut ainsi obtenir les bigrammes et trigrammes aux scores les plus élevés en choisissant la mesure que l'on souhaite utiliser. Toutes les mesures statistiques que nous avons décrites dans le chapitre précédent sont disponibles. La documentation de ce module se trouve à cette adresse : `https://www.nltk.org/howto/collocations.html`.

Pour chaque corpus, nous allons donc comparer les résultats obtenus avec les différentes mesures (fréquence, PMI, test t, test du χ^2 et fonction de vraisemblance).

Voici un exemple de fonction permettant de récupérer les **n** bigrammes possédant le score le plus haut pour chaque mesure statistique (on filtre les bigrammes en supprimant ceux dont la fréquence est inférieure à 3 ainsi que ceux dont un des termes au moins est un stopword) :

```
def get_ME(text, n):

    results = {}
    tokens = nltk.wordpunct_tokenize(text)

    bigram_measures =
    ↪  nltk.collocations.BigramAssocMeasures()
    finder = BigramCollocationFinder.from_words(tokens)
    stop_filter = lambda w1, w2: w1 in stopwords or w2 in
    ↪  stopwords
    finder.apply_freq_filter(3)
    finder.apply_ngram_filter(stop_filter)
    results["frequence"] =
    ↪  sorted(finder.nbest(bigram_measures.raw_freq, n))
    results["pmi"] = finder.nbest(bigram_measures.pmi, n)
    results["student_t"] =
    ↪  finder.nbest(bigram_measures.student_t, n)
    results["chi_sq"] = finder.nbest(bigram_measures.chi_sq,
    ↪  n)
    results["likelihood_ratio"] =
    ↪  finder.nbest(bigram_measures.likelihood_ratio, n)

    return results
```

FIGURE 4.1 – Fonction détectant les collocations d'un texte avec le module **nltk.collocations**

S'agissant de corpus diachroniques, nous avons concaténé les textes de chaque année. Nous avons ensuite extrait les 10 bigrammes obtenant le meilleur score pour chaque année puis procédé de même pour les trigrammes. Ensuite, afin de vérifier la stabilité de l'utilisation de ces collocations à travers le temps, nous avons conservé uniquement celles qui figurent dans ce top 10 pour plusieurs années du corpus. Nous présenterons pour chaque corpus les résultats de chaque mesure afin de les comparer et juger de leur pertinence.

Nous avons jugé préférable de ne pas effectuer de prétraitement (pas de lemmatisation ou de stopwordisation) afin de conserver toute l'intégrité du texte. En effet, altérer le texte lorsque l'on cherche des collocations nous a semblé une assez mauvaise idée : comme on l'a vu, le figement est également morpho-syntaxique, de ce fait lemmatiser ou retirer des mots outils présente un risque de passer à côté de certains figements ainsi que de générer un certain nombre de faux positifs. Nous n'avons donc pas touché au texte avant calcul des scores mais nous avons pu écarter

automatiquement certains résultats. En effet, le module *collocation* de NLTK permet d'appliquer un filtre sur les n-grammes : nous avons donc exclu les bigrammes dont un des termes est un mot outil ainsi que les trigrammes dont le dernier terme est un mot outil. En effet, bien qu'on puisse s'attendre à ce que les mesures utilisées (excepté la fréquence) soient assez efficaces pour les exclure d'elles-mêmes, nous avons constaté que ce n'était pas le cas.

n-grams	frequence	pmi	student_t	chi_sq	likelihood_ratio
La famille	1989,2019,2020,2021		1989,2019,2020,2021		2021
art contemporain	1989,2004	2004	1989,2004	2004	1989,2004
New York	1995,2004,2005,2007, 2009,2011,2012,2013, 2018,2019,2020,2021	1989,1995,1997,2004, 2005,2007,2009,2012	1995,1997,2004,2005, 2007,2009,2011,2012, 2013,2018,2019,2020, 2021	1989,1995,1997,2004, 2005,2007,2009,2011, 2012,2013,2021	1995,1997,2004,2005, 2007,2009,2011,2012, 2013,2018,2019,2020, 2021
Sean Connery		1989,2015		1989,2015	
Museum of	1995,2011,2012,2014	1995	1995,2011,2012,2014	1995,2012	1995,2011,2012,2020
Grand Louvre	1997,1998	1998	1998	1998	1998
Metropolitan Museum	1997,2012,2019	1997,2012	1997,2012,2019	1997,2012	1997,2012,2019
XIXe siècle	1997	1997,2010	1997,2014	1997,2010	1997,2010,2014
aile Sackler	1997,1998	1998	1997,1998	1998	1997,1998
antiquités orientales	1997,1998	1998	1997,1998	1997,1998	1997,1998
Pierre Rosenberg	1998	1997,1998	1998	1997,1998	1998
British Museum	2012	1997,2012	2012	1997	2012
Marc Mayer	2004,2018		2004,2018	2004	2004,2018
Sackler Gallery	2008,2009,2011,2012, 2013,2014,2015	2008	2008,2009,2011,2012, 2013,2014,2015	2008,2009,2012	2008,2009,2011,2012, 2013,2014
Lena Dunham	2013		2013		2013,2015
Serpentine Sackler	2013,2014		2013,2014		2014
Zaha Hadid	2013		2013	2013,2017	2013,2016,2017
Villar Rojas		2013,2014		2013,2014	
Royal Academy	2014,2018		2014,2018	2015,2016	2014,2015,2018
Adam Driver	2015,2016		2015,2016		2015,2016
Star Wars	2015		2015	2015,2016	2015
chirurgie bariatrique	2016,2017		2016,2017		2016,2017
patients obèses	2016,2017		2016,2017		
santé publique	2017		2017		2017,2019
Nan Goldin	2018,2019	2020	2018,2019		2018,2019
Purdue Pharma	2018,2019,2020,2021		2018,2019,2020,2021		2018,2019,2020,2021
famille Sackler	2018,2019,2020,2021		2018,2019,2020,2021		2018,2019,2020,2021

FIGURE 4.2 – Bigrammes détectés sur plusieurs années (corpus Sackler)

On obtient des résultats assez similaires entre les mesures de PMI et du χ^2 ainsi qu'entre les trois autres mesures. Les résultats pour le premier corpus (Sackler, voir les figures 4.2 et 4.3) sont plutôt décevants. En effet, peu de collocations reviennent plusieurs années de suite excepté quelques entités nommées, principalement des noms de personnes, de villes ou de musées (qui sont d'une certaine manière des types de collocations, mais qui ne nous intéressent pas vraiment ici) : on ne détecte de ce fait l'installation d'aucune expression figée si ce n'est celle qui donne son nom à l'affaire, *crise des opioïdes* / *opiacés*. On constate d'ailleurs que l'installation de cette expression se fait assez tardivement puisqu'aucune des mesures utilisées ne la relève dans son top 10 avant 2018. Cependant, cela peut être dû tout simplement au nombre trop limité de données avant cette période, qui rend moins fiables les mesures statistiques. Malgré tout, même s'ils semblent peu pertinents pour notre tâche car on ne peut pas à proprement parler utiliser le terme d'*expressions figées* concernant les collocations trouvées, les résultats ne sont pas inintéressants pour autant. On constate que les n-grammes extraits pour les années plus anciennes sont du domaine artistique (les n-grammes contiennent des mots comme *musée*, *galerie*...) et qu'un glissement s'opère peu à peu vers des thématiques médicales (*chez les patients*, *perte de poids*, *prise en charge*...) : on distingue avec une grande netteté le

n-grams	frequence	pmi	student_t	chi_sq	likelihood_ratio
millions de francs	1987,1997,1998	1987,1997,1998	1987,1997,1998	1987,1997,1998	1987,1997,1998
département des antiquités	1997,1998	1997,1998	1997,1998	1997,1998	1997,1998
millions de dollars	1998,2009,2018,2019, 2020,2021	1998,2009	1998,2009,2018,2019, 2020,2021	1998,2009	1998,2009,2018,2019, 2020,2021
plus en plus	2004,2018,2019	2004	2004,2019	2004	2004
Musée des beaux	2005,2018	2005	2005,2018	2005	2005
arts du Canada	2005,2018	2005	2005,2018	2005	2005
Gallery of Art	2008,2009,2011	2008,2009,2011	2008,2009,2011	2008,2009,2011	2008,2009,2011
Art Media Agency	2011,2012,2013	2011,2012,2013	2011,2012,2013	2011,2012,2013	2011,2012,2013
Freer Gallery of	2011,2013	2011,2013	2011,2013	2011,2013	2011,2013
Freer and Sackler	2011,2015	2015	2015	2015	2015
Serpentine Sackler Gallery	2013,2014,2015	2013,2015,2017	2013,2014,2015	2013,2014,2015,2017	2013,2014,2015
Biennale de Venise	2015,2018	2015	2015,2018	2015	2015
chez les patients	2016,2017		2016,2017		2016,2017
perte de poids	2016,2017		2016,2017		2017
prise en charge	2016,2017		2016,2017	2017	2016,2017
crise des opiacés	2018,2019,2020,2021		2018,2019,2020,2021		2018,2019,2020,2021
milliard de dollars	2018,2019,2020		2018,2019,2020		2018,2020,2021
milliards de dollars	2018,2019,2020,2021		2018,2019,2020,2021	2021	2018,2019,2020,2021
Addiction Intervention Now		2018		2018,2019	
New York Times	2020		2020		2018,2020
crise des opioïdes	2019,2020,2021		2019,2020,2021		2018,2020,2021
La famille Sackler	2019,2020,2021	2021	2019,2020,2021	2021	2020,2021
Hard Truth Ain		2020,2021		2020,2021	2021

FIGURE 4.3 – Trigrammes détectés sur plusieurs années (corpus Sackler)

moment où le nom Sackler, longtemps associé au mécénat, devient le symbole de la crise des opioïdes aux États-Unis.

Concernant le second corpus (PMA, voir les figures 4.4 et 4.5), nous montrons ici uniquement les collocations détectées par toutes les mesures (les tableaux auraient été trop volumineux). On y trouve également des entités nommées (principalement de personnes, plus particulièrement de personnages politiques). Nous rencontrons le même problème que pour le premier corpus, avec beaucoup d'expressions détectées de manière très sporadique (sur une année seulement ou au mieux deux ou trois années éloignées les unes des autres). Quelques expressions apparaissent plusieurs années, puis disparaissent du top 10 (*fécondation in vitro*, par exemple, n'est plus relevé après 2011 par la plupart des mesures, tandis que *couples homosexuels* n'apparaît plus après 2016) mais peu d'entre elles s'installent définitivement dans le corpus. On peut tout de même relever *gestation pour autrui* qui arrive en 2006 et semble progressivement prendre la place de la fécondation in vitro dans le débat, s'installant dans le top 10 des collocations jusqu'en 2021 et, bien sûr, *procréation médicalement assistée* (rien d'étonnant puisque *PMA* est le mot clé ayant servi à récolter le corpus sur Europresse).

Afin d'avoir une vue plus complète et une meilleure compréhension des résultats, nous avons également généré des graphiques de type **stream graph** (graphique en aires superposées) à l'aide de la librairie *altair* : un graphique par corpus, par type de n-gramme et par mesure. On peut constater qu'il y a en fait peu de résultats pour le corpus Sackler, qui est probablement trop petit pour ce type de mesures statistiques. De fait, en sélectionnant les dix collocations obtenant le meilleur score, on obtient en fait souvent pour ce corpus la totalité des collocations trouvées, qui ont donc parfois un score assez faible. Les résultats pour ce corpus sont donc à considérer avec un certain recul.

D'autre part, on constate également que pour les deux corpus, la fréquence fournit des résultats assez peu satisfaisants pour les bigrammes du fait de la mauvaise toke-

n-grams	frequence	pmi	student_t	chi_sq	likelihood_ratio
diagnostic prénatal	1994,1999,2005	1994,1999,2005	1994,1999,2005	1994,1999,2005	1994,1999,2005
fécondation in	1994,1995,1996,1997, 1999,2006,2007,2010, 2011	1994,1995,1996,1997, 2002,2006	1994,1995,1996,1997, 1999,2006,2007,2008, 2010,2011	1994,1995,1996,1997, 2006,2010	1994,1995,1996,1997, 1999,2006,2007,2008, 2010,2011
in vitro	1994,1995,1996,1997, 1999,2001,2006,2007, 2008,2010,2011	1994,1995,1996,1997, 2000,2001,2002,2006, 2010	1994,1995,1996,1997, 1999,2001,2006,2007, 2008,2010,2011	1994,1995,1996,1997, 1999,2000,2001,2002, 2006,2007,2008,2010, 2011	1994,1995,1996,1997, 1999,2000,2001,2002, 2006,2007,2008,2010, 2011,2015,2016
médicalement assistée	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2012,2013,2014, 2015,2016,2017,2018, 2019,2020,2021	1994,1995,1998,2000, 2002,2005,2009	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2012,2013,2014, 2015,2016,2017,2018, 2019,2020,2021	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2008,2009,2010	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2012,2013,2014, 2015,2016,2017,2018, 2019,2020,2021
procréation médicalement	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2012,2013,2014, 2015,2016,2017,2018, 2019,2020,2021	1994,1995,1998,2000, 2002,2005,2009	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2012,2013,2014, 2015,2016,2017,2018, 2019,2020,2021	1994,1995,1998,1999, 2000,2002,2005,2006, 2009	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2012,2013,2014, 2015,2016,2017,2018, 2019,2020,2021
Axel Kahn	1996,2000	1996,2000	1996,2000	1996,2000	1996,2000
deux ans	1997,2000	1997,2000	1997,2000	1997,2000	1997,2000
Christine Boutin	1998	1998,2011	1998	1998,2011	1998
couples homosexuels	1998,2004,2007,2012, 2013,2014,2015,2016	1998	1998,2004,2007,2012, 2013,2014,2015,2016	1998	1998,2004,2012,2013, 2014,2016
peut être	1998,2000,2009	1998	1998,2000	1998	1998,2000
Israël Nisand	2007	1999,2007	2007	2007,2018	2007
assistance médicale	2007	1999,2000	2007	2000	2007
projet parental	2000,2002	2000,2002,2010	2000,2002	2000,2002,2010	2000,2002
Lionel Jospin	2001	2000,2001	2001	2000,2001	2000,2001
François Hollande	2004,2012,2013,2014, 2015,2016,2017	2004	2004,2012,2013,2014, 2015,2016,2017	2004	2004,2012,2013,2014, 2015,2016,2017
autorité parentale	2004	2004	2004	2004	2004,2014
insémination artificielle	2007,2008,2010	2004,2010,2021	2007,2008,2010	2004,2007,2008,2010, 2021	2007,2008,2010
La Croix	2009,2011	2009	2009,2011	2009	2009,2011
états généraux	2018	2010	2018	2010	2013,2018
Irène Théry	2011,2016	2011,2012,2017	2011,2016	2011,2012,2013,2018	2011,2014,2016
première lecture	2020,2021	2011	2020,2021	2011	2020,2021
Manuel Valls	2014	2012	2014	2012,2015	2014,2015
Comité consultatif	2013,2017,2018	2020	2013,2017,2018	2020	2013,2017,2018

FIGURE 4.4 – Bigrammes détectés sur plusieurs années et par toutes les mesures (corpus PMA)

nisation de la ponctuation par NLTK, qui considère certaines séquences comme «). » ou « ... » comme des mots : ainsi ces séquences apparaissent comme premier terme de bigrammes et leur fréquence est particulièrement élevée dans le corpus PMA.

Les graphiques se trouvent dans la partie Annexe B.1. Nous avons produit un total de 20 graphiques (cinq mesures, deux tailles de n-grammes, deux corpus), cependant nous n'avons pas voulu surcharger la partie annexe, c'est pourquoi nous y présentons seulement les trigrammes du corpus PMA, qui nous semblent constituer les résultats les plus intéressants.

En conclusion, les résultats des mesures statistiques standards ne sont pas très concluants mais cela peut être dû à la taille trop limité des corpus choisis. Cependant, bien qu'ils ne soient pas concluants pour notre tâche de recherche d'expressions figées, ils n'en font pas moins parler les corpus de manière intéressante.

n-grams	frequence	pmi	student_t	chi_sq	likelihood_ratio
fécondation in vitro	1994,1995,1996,1997, 1999,2002,2006,2007, 2008,2010,2011	1994,1995,1996,1997, 1999,2002,2006,2007, 2008,2010,2011,2012, 2013,2014,2015,2016, 2017	1994,1995,1996,1997, 1999,2002,2006,2007, 2008,2010,2011	1994,1995,1996,1997, 1999,2002,2006,2007, 2008,2010,2011,2012, 2013,2014,2015,2016, 2017,2018,2019,2020	1994,1995,1996,1997, 1999,2002,2006,2007, 2008,2010,2011
procréation médicalement assistée	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2012,2013,2014, 2015,2016,2017,2018, 2019,2020,2021	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2011,2021	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2011,2012,2013, 2014,2015,2016,2017, 2018,2019,2020,2021	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2011,2012,2013, 2015,2016,2017,2021	1994,1995,1998,1999, 2000,2002,2004,2005, 2006,2007,2008,2009, 2010,2011,2012,2013, 2014,2015,2016,2017, 2018,2019,2020,2021
avoir un enfant	1995,2002,2008	1995,2002	1995,2002,2008	1995,2002	1995,2002,2008
procréation médicale assistée	1997	1997	1997	1997	1997,2012,2013,2015, 2016,2017,2018,2019
proposition de loi	1998,2014	1998	1998,2014	1998	1998
plus en plus	1999,2004,2005,2013	1999,2004,2005	1999,2004,2005,2013	1999,2004,2005	1999,2004,2005
projet de loi	2002,2010,2011,2012, 2013,2014,2018,2019, 2020,2021	2002,2010	2002,2010,2011,2012, 2013,2014,2018,2019, 2020,2021	2002,2010	2002,2010,2011,2012, 2013,2019,2020,2021
différence des sexes	2004	2004,2008	2004	2004,2008	2004
gestation pour autrui	2006,2007,2008,2009, 2012,2013,2014,2015, 2016,2017,2018,2020, 2021	2006,2007,2008,2009, 2021	2006,2007,2008,2009, 2012,2013,2014,2015, 2016,2017,2018,2020, 2021	2006,2007,2008,2009, 2021	2006,2007,2008,2009, 2016,2017,2018,2020
artificielle avec donneur	2007,2010	2007,2010	2007,2010	2007,2010	2007,2010
don de sperme	2007,2008,2010,2011, 2015	2010	2007,2008,2010,2011, 2015	2010	2007,2010
Procréation médicalement anonyme	2010,2011	2008,2010,2011	2010,2011	2008,2010,2011	2008,2010,2011
lois de bioéthique	2009,2011,2017	2009	2009,2011,2017	2009,2011	2009,2011
don de gamètes	2010,2019	2010	2010,2019	2010	2010
Comité consultatif national	2013,2017,2018	2015,2016,2020	2013,2017,2018	2013,2014,2015,2016, 2017,2020	2017,2018
prise en charge	2020,2021	2021	2020,2021	2021	2021

FIGURE 4.5 – Trigrammes détectés sur plusieurs années et par toutes les mesures (corpus PMA)

4.3 Implantation de la méthode de détection de Van de Cruys-Villada

Nous avons ensuite tenté de reproduire l'expérience de [Van de Cruys and Villada Moirón, 2007], décrite au chapitre précédent. Nos corpus étant beaucoup plus petits que le corpus utilisé pour l'expérience originale, nous avons légèrement adapté la méthode. En effet, là où les auteurs de l'article avaient présélectionné des candidats en se concentrant sur un schéma syntaxique particulier (verbe + préposition + nom) et en conservant les combinaisons des 10 000 noms les plus fréquents avec les 5 000 couples verbe + préposition les plus fréquents, nous avons décidé d'appliquer la méthode à tous les bigrammes contenant un des 500 noms les plus fréquents, dont le cooccurrent n'est pas un mot outil.

Nous avons utilisé SpaCy (lien ici : `https://spacy.io/`) pour la tokenisation, le POS-taging (nécessaire pour sélectionner les noms) et l'analyse syntaxique (le clustering des noms nécessaire à cette méthode se base sur leurs relations syntaxiques). Nous avons supprimé les mots outils lors de la tokenisation du texte afin de limiter le bruit dans les résultats. En effet, pour le clustering, nous avons représenté chaque nom par un vecteur contenant le score PMI entre ce nom et chacun des mots en relation syntaxique avec un des 500 noms sélectionnés. Les relations qu'un nom peut entretenir avec un mot outil comme un article ou une préposition ne nous a pas semblé pertinent pour un clustering censé regrouper les noms par proximité sémantique.

De plus, nous recherchons ici des bigrammes et un bigramme dont un des termes est un mot outil ne présente que peu d'intérêt.

La fonction ayant permis de calculer le score PMI est la suivante (*unigram_freq* est un dictionnaire Counter donnant la fréquence de chaque mot du corpus, *dep_freq* est la fréquence à laquelle les deux mots sont en relation de dépendance dans le corpus) :

```
def get_pmi(word1, word2, unigram_freq, dep_freq):
  prob_word1 = unigram_freq[word1] /
   ↪  float(sum(unigram_freq.values()))
  prob_word2 = unigram_freq[word2] /
   ↪  float(sum(unigram_freq.values()))
  prob_word1_word2 = dep_freq /
   ↪  float(sum(unigram_freq.values()))
  try:
    return
     ↪  math.log(prob_word1_word2/float(prob_word1*prob_word2),2)
  except:
    return 0
```

FIGURE 4.6 – Fonction calculant le score PMI de deux termes

Le module *cluster* de la librairie Scikit-Learn nous a permis d'implémenter très facilement l'algorithme de clustering k-moyennes. Le clustering effectué, nous avons formé une liste de tous les bigrammes contenant au moins un des noms sélectionnés et apparaissant au minimum trois fois, puis nous avons implémenté les différentes formules de l'article afin de calculer, pour chaque bigramme, s'il existe une préférence marquée du nom sélectionné sur ses *synonymes* (noms appartenant au même cluster). Les fonctions implémentant les formules sont expliquées en détail en annexe A.2.

Comme pour les mesures statistiques standards, nous avons appliqué la méthode sur les textes concaténés de chaque année, puis nous avons conservé les colocations détectées sur plusieurs années du corpus.

Probablement du fait de la taille réduite des corpus, le clustering ne produit pas de très bons résultats (la répartition entre clusters étant très déséquilibrée), ce qui impacte les résultats finaux. En effet, comme on peut le voir dans les extraits de résultats montrés en figure 4.7 et 4.8, ces derniers ne sont pas très satisfaisants. La plupart des collocations trouvées dans le corpus Sackler ne sont détectés qu'en 2019. Du fait que nous ne conservons que les collocations détectées sur plusieurs années, les résultats sont peu nombreux et, malheureusement, ne sont pas très convaincants (sur 24 résultats, seulement trois ou quatre peuvent vraiment être considérés comme des collocations). Concernant le corpus PMA, les résultats sont beaucoup plus nombreux (291) mais comprennent également une grande majorité de simples cooccurrences ou d'entités nommées. Cela est dû à la taille réduite des corpus et à la mauvaise qualité du clustering : beaucoup de clusters ne contenant qu'un seul mot, un nombre élevé de noms se sont retrouvés sans concurrent et la mesure de la préférence lexicale d'un mot pour ces noms n'a donc pas pu se faire. De fait ces mots se sont retrouvés avec un score maximal et ont gonflé le nombre de faux positifs, faisant grandement baisser la précision. La méthode est donc peu

avantageuse (pour nos corpus, tout du moins) étant donné sa complexité et le peu de valeur ajoutée qu'elle apporte.

Quoi qu'il en soit, ces méthodes n'ont pas été conçues pour des corpus diachroniques : elles ont pour objectif de détecter des expressions figées mais ne permettent pas de différencier des expressions préexistantes et de nouvelles expressions propres au corpus. Or nous souhaiterions pouvoir détecter ces nouvelles expressions formées au fil d'un corpus dont les textes recouvrent une période plus ou moins étendue.

ngram	1978	1987	1989	1995	1996	1997	1998	1999	2001	2002	2003	2004	2005	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
conseil administration	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	
art of	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	O	N	N	O	N	N	N	N	N	N	N
sackler gallery	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	O	O	N	N	N	N	O	N	N	N	N
xixe siècle	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	N	N	N	N	O	N
m. sackler	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N	N	N	N	N	O	N	N
freer and	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	O	N	N	N	N	N	N
new york	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	N	N	N	O	O	N	N
metropolitan museum	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	O	O	N	N
ans imc	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N	N	N
millions dollars	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	O	N
type diabète	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N	N	N
milliards dollars	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	O	N	N
crise opiacés	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	O
londres gallery	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N
états américains	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N
antiquités orientales	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N
purdue pharma	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
mettre fin	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
morts overdose	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
fortune estimée	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
loi faillites	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
loi protection	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
espérance vie	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
crise opioïdes	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O

FIGURE 4.7 – Collocations détectées avec méthode Van de Cruys-Villada (corpus Sackler)

ngram	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
embryon recherche	N	N	N	N	N	O	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	N	N
assistance médicale	N	N	N	N	N	O	O	N	N	N	N	N	N	O	N	N	N	N	N	N	N	O	O	N	O	N	O	O
bioéthique loi	N	N	N	N	N	N	N	N	O	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O
égalité sexes	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N
droits égalité	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N
différence sexes	N	N	N	N	N	N	N	N	N	N	O	N	N	N	O	N	N	N	N	N	N	N	N	O	N	N	N	N
enfants nés	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	O	N	N	O	O	O	N	O	O	N	O	N
pma assistée	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	O	O	O	O	O	O	N	N	O	N
homosexuels couples	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	O	N	O	N	N	N	N	N	N	N
hommes femmes	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	O	O	N	O	N	N	O	N	N	N	N
insémination artificielle	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	O	O	O	O	O	O	N	O	N	N
venue monde	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	N	N	N	N	N	O	N	N
sperme don	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	N	O	O	N	N	N	N	N
gamètes don	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	O	N	N	N	N	N	O	N	N
levée anonymat	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	N	N	N	O	O	N	N
gamètes dons	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	O	N	N	N	N	O	N	N
gamètes donneurs	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	N	N	O	N	N
identité donneur	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N	N	N	N	N	N	N	N
bioéthique lois	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	O	N	N	O	N	N	N	N
insémination donneur	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	N	N	N	N	N	N	N
révision lois	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	N	N	O	O	N	O	N
ministre déléguée	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N	N	N	N	N	N	N
assemblée nationale	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	O	O	O	N	O	N
personnes sexe	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	O	N	O	O	O	N	N	N
vote étrangers	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	N	N	N	N	N	N	N
couple homosexuel	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	O	N	N	N	N
temps débat	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	N	N	N	N	N	O	N	N	N
conseil ministres	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	N	N	N	O	O	O	N
vie fin	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	N	O	O	O	N	N	N
état chef	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N	O	N	N	N	N	N

FIGURE 4.8 – Collocations détectées avec méthode Van de Cruys-Villada (corpus PMA, 30 premiers résultats)

4.4 Nouvelle méthode : topic modeling avec LDA et détection de la disparition progressive des synonymes

4.4.1 Sélection des candidats

Lors de mon stage, j'ai été amenée à effectuer du **topic modeling** sur les corpus en utilisant un modèle LDA (**Latent Dirichlet Allocation**) sur des clusters obtenus avec un algorithme de propagation d'affinités appliqué sur les articles de chaque année. La LDA génère pour chaque cluster d'articles un certain nombre de n-grammes représentatifs des thèmes abordés dans le cluster. Ces n-grammes nous ont semblé pouvoir constituer une bonne base pour présélectionner de potentiels figements.

Dans un premier temps, nous effectuons un **clustering** des articles de chaque année du corpus. Pour cela nous utilisons la librairie *sklearn* et son module *cluster*. Nous avons choisi la méthode de clustering par propagation d'affinités car elle permet de ne pas avoir à déterminer à l'avance le nombre de clusters : en effet, la méthode identifie elle-même les éléments exemplaires d'un cluster par un procédé itératif au cours duquel chaque élément recherche parmi les autres celui avec lequel il partage le plus d'affinités. Nous avons représenté chaque document, préalablement stopwordisé avec Spacy, par un vecteur avec la méthode **doc2vec** (développée par [Le and Mikolov, 2014]) de la librairie Gensim, qui utilise les plongements de mots (**word2vec**) pour produire un vecteur représentant un document. Pour chaque année, nous avons soumis ces vecteurs à l'algorithme de clustering. Nous avons ensuite regroupé en un seul cluster les clusters ne comptant qu'un seul document.

Un modèle LDA a ensuite été appliqué aux résultats. Pour chaque année, nous avons concaténé les articles d'un même cluster afin d'avoir un unique document par cluster, que nous avons représenté par un vecteur de TF-IDF, le vocabulaire étant constitué de l'ensemble des n-grammes des articles de l'année (**n** étant un paramètre du script, nous avons lancé le script à plusieurs reprises avec une valeur de **n** différente, allant de 2 à 4). Pour les raisons déjà évoquées plus haut, nous avons conservé les mots outils et n'avons pas lemmatisé le texte avant passage de l'algorithme de LDA (et nous avons redéfini le paramètre *token_pattern* de TfidfVectorizer qui par défaut ne prend pas en considération les mot d'une seule lettre), mais nous avons par la suite écarté les résultats commençant ou se terminant par un mot outil.

Le modèle LDA génère un nombre de topics (thèmes) prédéfini (ici nous avons fixé un nombre de topics égal au nombre de clusters de l'année), en créant une matrice qui donne une mesure d'association entre chaque mot du vocabulaire et chaque topic. Il donne ensuite pour chaque document (ici donc pour chaque cluster) la probabilité de son appartenance à chaque topic. Nous avons conservé les 50% premiers topics de chaque cluster, que nous avons représentés par les 10 n-grammes les plus représentatifs de ces topics (pour chaque topic, 10 n-grammes divisé par le nombre de topics retenus). La fonction utilisée pour effectuer le topic modeling via la méthode *LatentDirichletAllocation* du module *decomposition* de scikit-learn se trouve en annexe A.3, où elle est expliquée plus en détail. La documentation en ligne se trouve à cette adresse : https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html.

Enfin, nous n'avons gardé que ceux qui représentent au moins deux clusters. Le tableau qui suit (4.1) donne le nombre de candidats ainsi retenus :

Corpus	2-grammes	3-grammes	4-grammes	Total
Sackler	30	45	49	124
PMA	20	35	49	104

TABLE 4.1 – Nombre de candidats retenus par corpus et type de n-grammes avec méthode clustering + LDA

En jetant un œil à cette première sélection, on trouve des candidats intéressants comme *crise des opioïdes*, *crise des opiacés*, *droit à l'enfant*, *accès aux origines*, mais également de simples cooccurrences (*millions de dollars*, *loi sur la famille*) ou des entités nommées (*New York*, *musée des Beaux Arts*, *Arthur Kermalvezen*...). On constate que les candidats sont moins nombreux pour le corpus PMA, pourtant plus grand que le corpus Sackler. Cela peut s'expliquer par une plus grande homogénéité du corpus : les thèmes abordés y sont moins nombreux malgré le nombre plus élevé de documents et, de ce fait, les n-grammes représentant les clusters reviennent probablement plus régulièrement.

4.4.2 Détection des figements

Comme nous l'avons déjà exposé à plusieurs reprises, la première étape du figement est la disparition de ses « anti-collocations », c'est-à-dire que les potentiels synonymes disparaissent progressivement au profit de l'expression qui se fige et que l'on constate une préférence mutuelle entre les termes de ladite expression. La suite de notre méthode consiste donc à vérifier pour chaque candidat si ce dernier devient majoritaire au fil des années par rapport à des expressions comprenant un synonyme à la place d'un des termes du dit candidat.

Pour commencer nous avons donc recherché tous les synonymes, quasi-synonymes ou co-hyponymes de chaque terme de chaque candidat, en utilisant la méthode *similarity* de la librairie SpaCy, qui calcule la similarité cosinus entre les vecteurs de mots. Spacy utilise, pour le français, les vecteurs de mots *FastText*, une des trois principales architectures de plongements lexicaux avec *word2vec* et *GloVe*. Ces vecteurs de mots représentent les mots en fonction des contextes dans lesquels ils apparaissent habituellement. On considère que deux mots apparaissant dans les mêmes contextes sont sémantiquement proches. Après quelques tests, nous avons fixé à 0.6 la similarité minimale entre deux vecteurs de mots permettant de les considérer comme des synonymes ou co-hyponymes. On fait ensuite une liste, pour chaque candidat, de potentielles « anti-collocations », en remplaçant chaque terme par chacun de ses synonymes (par exemple, pour le candidat *crise des opioïdes*, on obtient l'expression synonyme *crise des opiacés* en remplaçant le dernier terme par un synonyme), puis on compte pour chaque année du corpus la fréquence de chaque candidat et de chacune de ses « anti-collocations ». Le script effectuant cette tâche produit en sortie un fichier au format json avec la structure suivante : {collocation : {année : {collocation synonyme : nombre d'occurrences}}} (parmi les synonymes se trouve la collocation elle-même, puisqu'on a besoin de connaître son nombre d'occurrences par année également). La fonction permettant de faire la liste des synonymes de chaque

candidat est la suivante (*candidates* est un dictionnaire associant taille de n-gramme et liste des n-grammes de cette taille tandis que *vocab* est une liste de tokens Spacy) :

```
def get_synonymes(candidates, vocab):

    """Pour chaque candidat, génère une liste d'expressions
    ↪  a priori sémantiquement équivalentes (candidat dont
    ↪  on a échangé un terme avec un synonyme)"""

    synonymes = {}
    for ngram_size, candidates_list in candidates.items():
        print("Recherche de synonymes des ", ngram_size,
        ↪  "-grams")
        for candidate in candidates_list:
            parsed_candidate = nlp(candidate)
            synonymes[candidate] = []
            for i in range(in(ngram_size)):
                if parsed_candidate[i].is_stop:
                    continue
                for token in vocab:
                    if token.text !=
                    ↪  parsed_candidate[i].text and
                    ↪  token.similarity(parsed_candidate[i])
                    ↪  > 0.6:
                        synonymes[candidate].append("
                        ↪  ".join([token.text if word ==
                        ↪  parsed_candidate[i] else
                        ↪  word.text for word in
                        ↪  parsed_candidate]))
    return synonymes
```

FIGURE 4.9 – Fonction de topic modeling avec LDA

Voici un extrait intéressant du fichier de sortie avec l'exemple de la collocation *crise des opiacés* :

```
"crise des opiacés": {
        "2017": {
            "crise des opioïdes": "1"
        },
        "2018": {
            "crise des opiacés": "13",
            "crise des opioïdes": "5",
            "épidémie des opiacés": "3"
        },
        "2019": {
            "catastrophe des opiacés": "1",
            "crise des opiacées": "1",
            "crise des opiacés": "160",
            "crise des opioïdes": "104",
            "crise des surdoses": "1",
            "épidémie des opiacés": "1"
        },
        "2020": {
            "crise des opiacés": "16",
            "crise des opioïdes": "16"
        },
        "2021": {
            "crise des opiacés": "12",
            "crise des opioïdes": "8"
        }
    }
```


FIGURE 4.10 – Extrait du fichier json comptant les synonymes de chaque candidat

Enfin on vérifie avec un dernier script si la proportion du candidat augmente avec le temps. La fonction suivante (figure 4.11) permet de calculer pour chaque année le taux d'apparition d'une expression candidate par rapport à ses synonymes, puis de vérifier si cette expression valide les critères que nous avons jugés importants afin de décider s'il s'agit ou non d'une expression figée : l'expression doit apparaître sur au moins quatre années différentes (pas nécessairement consécutives), la moyenne de ses taux d'apparition par année pour la première moitié des années du corpus doit être inférieure à celle calculée pour la seconde moitié. On vérifie également que le candidat représente plus de la moitié des occurrences de l'expression par rapport à ses synonymes sur les quatre dernières années.

```
def is_ME(candidate, candidate_syns):
    ratios = []
    for year, cnt_syns in candidate_syns.items():
        try:
            candidate_freq = int(cnt_syns[candidate])
        except:
            candidate_freq = 0
        syn_freq = sum([int(cnt_syns[syn]) for syn in
        ↪  cnt_syns.keys() if syn != candidate])
        if syn_freq == 0 and candidate_freq != 0:
            ratios.append(1)
        elif syn_freq != 0:

            ↪  ratios.append(candidate_freq/(syn_freq+candidate_freq))
        else:
            ratios.append(0)
    if len([ratio for ratio in ratios if ratio != 0]) > 3
    ↪  and statistics.mean(ratios[:int(len(ratios)/2)]) <
    ↪  statistics.mean(ratios[int(len(ratios)/2):]) and
    ↪  statistics.mean(ratios[-4:]) > 0.5:
        return True
    else:
        return False
```

FIGURE 4.11 – Fonction finale de notre méthode

Nous obtenons 17 résultats pour le corpus Sackler et 47 pour le corpus PMA.

Le principal problème rencontré avec cette méthode est la difficulté à trouver des synonymes. Comme avec la méthode Van de Cruys-Villada, du fait de la taille réduite des corpus, beaucoup de candidats se sont retrouvés sans synonymes dans le corpus et il a donc été impossible d'évaluer la réduction ou disparition progressives des anti-collocations, ces dernières étant absentes dès le départ. Cela a généré un certain nombre de faux positifs.

On constate que notre méthode a fortement réduit le nombre de candidats pour le corpus Sackler alors qu'il en reste près de la moitié pour le corpus PMA. Cela confirme ce que nous avions dit précédemment sur l'homogénéité du corpus PMA par rapport au corpus Sackler. Notre algorithme a eu moins de difficulté à faire le tri dans le corpus Sackler car beaucoup de n-grammes n'apparaissaient en fait que deux ou trois années dans le corpus et ils n'ont donc pas été retenus. De plus, ce corpus voit ses thématiques changer assez radicalement au fil du temps. En effet, les premières années ont pour thème le rôle de mécène de la famille Sackler, tandis que la fin du corpus est axée sur la crise des opiacés. De ce fait, de nombreux n-grammes qui figurent au début du corpus disparaissent par la suite et ils ont donc été écartés par notre méthode. À l'inverse, nous obtenons beaucoup de résultats pour le corpus PMA parmi lesquels beaucoup de bruit, car le corpus évolue moins dans ces thématiques et, notre première étape ayant proposé un nombre de candidats déjà réduits, notre méthode connaît des difficultés pour les réduire encore plus.

Nous étudierons plus en détails ces résultats dans le prochain chapitre.

CHAPITRE 5

ÉVALUATION ET ANALYSE DES RÉSULTATS

Sommaire

5.1 Introduction : comment évaluer notre méthode ?

Dans ce chapitre, nous évaluerons les résultats de notre méthode sur un des corpus. Pour cela, nous annoterons les candidats pré-selectionnés par le modèle LDA, puis évaluerons la seconde partie de notre méthode (détection des collocations via la disparition des anti-collocations) en calculant précision, rappel et F-mesure. Nous comparerons également nos résultats avec ceux des deux autres expériences menées (mesures standards d'association lexicale et adaptation de la méthode de [Van de Cruys and Villada Moirón, 2007]) afin de déterminer si notre méthode apporte une plus-value aux méthodes existantes.

5.2 Évaluation de la seconde partie de notre méthode

5.2.1 Annotation des données

Afin d'évaluer la seconde partie de notre méthode, consistant à rechercher des expressions synonymes aux candidats préalablement sélectionnés et à déduire de leur absence ou de leur diminution dans le temps le degré de figement du dit candidat, nous avons dû annoter les candidats sélectionnés par la première partie de notre méthode (qui utilise un modèle LDA pour produire les n-grammes les plus représentatifs des thématiques du corpus). Cela a soulevé diverses problématiques.

Tout d'abord, idéalement, une tâche d'annotation doit être effectuée par plusieurs annotateurs afin d'optimiser la qualité et la fiabilité de l'annotation via le calcul de l'accord inter-annotateurs (pour lequel il existe plusieurs mesures dont l'une des plus

connues est le kappa de Cohen). D'autre part, la ou les personnes qui annotent ne doivent pas être celles à l'origine du projet. Ayant annoté moi-même et étant l'unique annotatrice, aucune de ces deux préconisation n'a été respectée et la démarche scientifique peut donc être légitimement critiquée. Malheureusement, je n'ai disposé ni du temps ni des moyens suffisants pour mettre en place un protocole d'évaluation plus conforme à l'éthique.

Concernant l'annotation en elle-même, plusieurs difficultés sont apparues. Nous avons choisi d'annoter les candidats du corpus PMA, le corpus étant plus fourni et de ce fait probablement plus à même de fournir des résultats intéressants. Pour chaque n-gramme candidat, nous avons choisi entre cinq étiquettes : **expression spécifique, expression générale, entité nommée, mot composé, cooccurrence**. Les trois premières étiquettes peuvent se voir associer la mention *incomplète* (dans le cas où il manque un token, par exemple *procréation médicalement*).

L'étiquette *expression spécifique* concerne les expressions figées fortement liées au thème du corpus (comme *procréation médicalement assistée* ou *fécondation in vitro*) : il s'agit des n-grammes que l'on souhaite extraire avec notre méthode.

L'étiquette *expression générale* est attribué aux expressions figées qui n'ont aucun lien particulier avec le thème du corpus. Nous avons par exemple donné cette étiquette à des expressions comme *projet de loi* ou *fin de vie*. Il n'a pas toujours été facile de choisir entre ces deux étiquettes et certains de nos choix pourront sans doute être discutés. Par exemple, nous avons considéré *théorie du genre* comme une expression spécifique car, bien que pas directement liée au débat sur la PMA, il s'agit d'un débat apparu à une période et dans un contexte relativement proches. En revanche, nous avons annoté *vrais jumeaux* comme expression générale car, bien que la gémellité soit liée à la PMA, le terme est connu en dehors du contexte de ces débats et, de plus, il s'agit plus d'une collocation que d'une véritable expression.

Les autres étiquettes ont des noms plus explicites. Il n'était cependant pas toujours facile de choisir l'étiquette adéquate. Par exemple, nous avons décidé d'annoter les noms de lois en entités nommées (*loi de bioéthique*, par exemple, ou *mariage sexuellement neutre*, qui est le nom d'une loi suédoise), mais avons considéré les n-grammes *loi sur la famille* et *loi sur le mariage* comme de simples cooccurrences car il ne s'agit pas de réels noms de lois, mais plutôt de termes généraux. Du fait de la suppression de la ponctuation lors du prétraitement des données (et donc de la suppression des traits d'union), parmi les candidats figuraient quelques noms composés pour lesquels nous avons donc dû ajouter une étiquette.

En annotant les candidats, nous avons pu constater qu'il y avait très peu d'expressions valides parmi les bigrammes mais beaucoup d'entités nommées, tandis que les 4-grammes contenaient un très grand nombre de cooccurrences, et que c'était parmi les trigrammes que l'on trouvait le plus d'expressions à extraire. Il serait donc probablement intéressant de se focaliser sur les trigrammes dans un éventuel futur travail sur les figements lexicaux.

5.2.2 Méthode d'évaluation

Concernant la méthode d'évaluation, nous utiliserons les mesures standards que sont la précision, le rappel et la F-Mesure.

La précision consiste à calculer, parmi tous les items associés à une classe, la proportion des items appartenant effectivement à cette classe. Elle mesure donc la

capacité de l'algorithme à limiter le bruit. Sa formule est :

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux positifs}} \tag{5.1}$$

Le but de notre méthode est bien entendu de détecter les n-grammes que nous avons labellisés en tant qu'*expressions spécifiques*. Cependant, nous avons décidé de nous montrer relativement souple, considérant que la détection d'expressions que l'on a appelées *générales*, des entités nommées et des mots composés n'était pas complètement hors de propos dans la mesure où il s'agit également de figements (même si les entités nommées constituent un cas à part, on peut dire qu'elles partagent un certain nombre de caractéristiques avec les expressions figées), bien que pas du type recherché. Nous considérerons ces n-grammes, dans le cas où ils sont extraits par notre méthode, comme des demi-vrais positifs. Par exemple, si la méthode repère une expression spécifique et une expression générale, sa précision sera de 0.75 ($\frac{1+0.5}{2}$). D'autre part, nous retirerons 0.25 à toute expression (ou entité nommée) incomplète. En revanche les simples cooccurrences seront de véritables faux positifs.

Le rappel mesure quant à lui le nombre d'items trouvés par rapport au nombre d'items à trouver. C'est ici le silence que l'on cherche à limiter. La formule est :

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux négatifs}} \tag{5.2}$$

Notre calcul sera beaucoup plus simple cette fois : on calculera le nombre d'expressions spécifiques trouvées par rapport au nombre d'expressions spécifiques parmi les candidats.

Enfin la F-Mesure est constitue une sorte de moyenne entre précision et rappel. Sa formule est la suivante :

$$\text{F-Mesure} = 2 * \frac{\text{Précision} * \text{Rappel}}{\text{Précision + Rappel}} \tag{5.3}$$

5.2.3 Résultats

Voici les résultats obtenus avec le corpus PMA :

Précision	Rappel	F-Mesure
0.47	0.79	0.59

TABLE 5.1 – Évaluation des résultats sur corpus PMA

Le rappel est plutôt bon, cependant la précision est plutôt décevante, surtout si l'on considère la souplesse dont on a fait preuve pour calculer cette dernière.

5.3 Comparaison avec les résultats des deux autres expériences

Nous n'avons pu évaluer que la seconde partie de notre méthode, à partir des résultats de la première partie. Afin de se faire une idée de la performance globale de

notre méthode, nous allons comparer ses résultats avec les résultats obtenus en utilisant les mesures d'association lexicale standards ainsi qu'avec ceux de la méthode Van de Cruys-Villada.

Pour chacune des deux méthodes, nous classerons les résultats sous forme de trois tableaux : un tableau avec les résultats communs avec notre méthode, un tableau avec les résultats trouvés uniquement avec notre méthode et un tableau avec les résultats trouvés uniquement avec l'autre méthode. Dans chaque tableau, les résultats seront classés selon qu'ils sont jugés *pertinents* (expressions figées jugées représentatives du corpus), *discutables* (entités nommées ou n-grammes présentant un certain degré de figement mais qui ne semblent pas particulièrement représentatif du corpus) ou *non pertinents* (simples cooccurrences). Dans la catégorie *discutables* (qui reprend grosso modo les mêmes types de n-grammes que ceux qui donnaient 0.5 de précision dans le chapitre précédent) se trouvent aussi les expressions mal segmentées.

Nous montrerons ici les tableaux de comparaison portant sur le corpus PMA. Les tableaux du corpus Sackler se trouvent en annexe B.2.

Pour commencer, nous avons récupéré pour chaque corpus les n-grammes extraits pour au moins deux années par les cinq mesures standards. Ce sont avec ces résultats que nous avons comparé les résultats de notre expérience.

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
procréation médicalement assistée, fécondation in vitro	in vitro, procréation médicalement, médicalement assistée, projet de loi	

TABLE 5.2 – Résultats communs avec mesures standards (corpus PMA)

D'après les tableaux 5.2, 5.3 et 5.4, d'un point de vue purement statistique, les résultats des cinq mesures statistiques d'association lexicale combinées semblent au premier abord très légèrement meilleurs que ceux de notre méthode. En effet, le taux de résultats pertinents pour les n-grammes détectés uniquement par les mesures standards est à peu près équivalent à celui calculé pour les n-grammes détectés uniquement par notre méthode mais le bruit (taux de résultats non pertinents) est légèrement inférieur (0.22 contre 0.3).

Cependant, si l'on regarde les résultats dans le détail, ceux de notre méthode nous semblent plus intéressants. Ils sont également plus nombreux sans pour autant générer trop de bruit. D'ailleurs, *gestation pour autrui* apparaît comme résultat trouvé uniquement par les méthodes statistiques mais notre méthode a trouvé *gestation pour autrui gpa* (où *GPA* devait être entre parenthèses dans le texte) : les mesures statistiques n'engendrent donc en réalité qu'un seul résultat pertinent que notre méthode n'a pas trouvé, tandis que notre méthode en a huit. Concernant les résultats non pertinents, ils sont certes bien plus nombreux avec notre méthode mais demeurent en proportions raisonnables. De plus, bien qu'étant de simples cooccurrences, ils demeurent assez intéressants si l'on souhaite seulement faire parler le texte : en effet des cooccurrences telles que *contre l'homophobie* ou *extension de la PMA* donnent beaucoup d'informations sur le texte là où *plus en plus* n'en donne aucune.

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
anonymat du don, désir d enfant, don de gamètes, accès aux origines, procréation médicalement assistée pma, pma procréation médicalement assistée, levée de l anonymat, droit à l enfant, gestation pour autrui gpa	porte parole, jean marc, manuel valls, médicalement assistée pma, loi de bioéthique, national d éthique, européenne des droits, cour de cassation, fin de vie, conseil des ministres, médicale à la procréation, ludovine de la rochère, ministre de l intérieur, consultatif national d éthique, chef de l état, national d éthique ccne, généraux de la bioéthique, projet de loi bioéthique, libération de la parole	contre l homophobie, couples de femmes, députés et sénateurs, président du groupe, recherche sur l embryon, recherche sur les embryons, couples de même sexe, nés d un don, loi sur le mariage, personnes de même sexe, pma à l étranger, extension de la pma

TABLE 5.3 – Résultats non trouvés par les mesures standards (corpus PMA)

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
procréation médicalement anonyme, gestation pour autrui	axel kahn, artificielle avec donneur, comité consultatif national, diagnostic prénatal, fécondation in	avoir un enfant, plus en plus

TABLE 5.4 – Résultats trouvés uniquement par les mesures standards (corpus PMA)

En conclusion, bien que les résultats des deux méthodes soient assez proches en termes de précision, il nous semble que notre méthode, en plus d'avoir un rappel légèrement supérieur, génère des résultats un peu plus intéressants que les mesures statistiques.

Concernant la méthode Van de Cruys-Villada, cette dernière ne recherchant que des bigrammes, nous n'avons effectué notre comparaison que sur les bigrammes. Cependant, pour les raisons évoquées au chapitre précédent du fait de la nature de cette méthode, nous avions supprimé les mots outils du texte et certains résultats peuvent correspondre à des trigrammes de notre méthode (plus précisément les trigrammes dont le second terme est une préposition) voire à des 4-grammes. Nous avons donc ajouté les n-grammes concernés à la comparaison.

Pour plus de clarté, nous avons ajouté à la main les prépositions ou conjonctions manquantes lorsque celles-ci étaient évidentes dans le tableau 5.7. Par ailleurs, ce tableau étant trop volumineux, nous avons dû retirer un certain nombre de résultats (comme nous l'avions noté lors de l'expérience, cette méthode a généré beaucoup de bruit).

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
désir d enfant, don de gamètes, accès aux origines, levée de l anonymat	procréation médicalement, manuel valls, loi de bioéthique, national d éthique, cour de cassation, fin de vie, conseil des ministres, chef de l état	députés et sénateurs, recherche sur l embryon

TABLE 5.5 – Résultats communs avec la méthode Van de Cruys-Villada (corpus PMA)

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
	in vitro, porte parole, jean marc, médicalement assistée	

TABLE 5.6 – Résultats non trouvés par la méthode Van de Cruys-Villada (corpus PMA)

Comme le confirme l'observation des tableaux 5.5, 5.6 et 5.7, la méthode Van de Cruys-Villada produit des résultats extrêmement bruités. Ainsi, cette méthode produit un taux beaucoup plus élevé de résultats non pertinents, sans pour autant détecter plus de résultats pertinents (tous les résultats pertinents sont communs aux deux méthodes). On peut donc conclure rapidement à la supériorité de notre méthode, dont le rappel est équivalent mais la précision beaucoup plus élevée et les résultats s'en trouvent donc beaucoup plus lisibles.

5.4 Conclusion : des résultats encourageants

En conclusion, notre méthode obtient des résultats plutôt satisfaisants. En effet, la seconde partie de notre méthode obtient un score de F-Mesure légèrement audessus de la moyenne, mais surtout cette méthode semble apporter une légère plusvalue par rapport aux deux autres méthodes testées, avec une précision beaucoup plus élevée que celle de la méthode Van de Cruys-Villada et des résultats légèrement plus intéressants que ceux obtenus avec les mesures statistiques standards.

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
insémination artificielle, don de sperme, dons de gamètes, identité du donneur, projet parental, mère porteuse, marchandisation du corps, mères porteuses, modèle bioéthique, anonymat des donneurs, donneur anonyme, théorie du genre, intérêt de l enfant	assistance médicale, égalité des sexes, égalité de droits, assistée pma, lois bioéthique, ministre déléguée, assemblée nationale, commission de lois, gpa gestation, justice christiane, bruno roux, opinion publique, secrétaire général, code civil, président de la république, sceaux christiane, liberté de conscience, prise de position, candidat hollande, intérêt supérieur...	différence de sexes, enfants nés, couples homosexuels, hommes et femmes, venue au monde, révision des lois, personnes de sexe, vote étrangers, couple homosexuel, temps de débat, sein du couple, insémination donneur, grande réforme, associations de défense, associations lgbt, père et mère, majorité des parlementaires, moment du pacs, questions de bioéthique, point de vue, rapporteur binet, légalisation pour les mères, gays et lesbiennes, gays et lesbiens, connaître l identité, semaine dernière, hétéros et homos...

TABLE 5.7 – Résultats trouvés uniquement par la méthode Van de Cruys-Villada (corpus PMA)

Conclusion générale

L'objectif de ce travail était de proposer une méthode de détection d'émergence de figements lexicaux au sein de corpus diachroniques.

Pour commencer, il était important de définir la notion de figement. En effet, le terme d'*expression figée* regroupe un large panel d'expressions au degré de figement plus ou moins élevé et aux caractéristiques variées. Parmi ces caractéristiques figurent notamment l'opacité sémantique (on ne peut pas déduire le sens d'une expression d'après le sens de ses constituants) ou la solidarité lexicale (on ne peut pas remplacer un terme d'une expression par un synonyme).

Notre objectif étant de trouver de nouvelles expressions, il nous a semblé pertinent de chercher à détecter les expressions sur le critère de la solidarité lexicale plutôt que sur celui de l'opacité sémantique, ce dernier étant plus caractéristique d'une expression déjà bien ancrée dans la langue et au degré de figement élevé.

Nous avons également décidé d'inclure les collocations dans notre travail. En effet, elles partagent avec les expressions figées le critère de solidarité lexicale et, étant donné que nous avons choisi de nous baser sur ce critère pour détecter les expressions figées, il semble logique que les collocations soient comprises dans notre champ de recherche, d'autant plus qu'elles présentent tout autant d'intérêt vis-à-vis de l'objectif de notre méthode, qui est d'observer l'évolution du traitement d'un thème dans un corpus diachronique.

Nous n'avons pas trouvé de travaux parlant de la détection en diachronie des expressions figées mais nombreux sont ceux sur la détection en synchronie. Au-delà des mesures standards d'association lexicales, que nous avons testées sur nos propres corpus et dont on a démontré l'efficacité, de nombreuses techniques de détection d'expressions polylexicales ont été mises au point par les chercheurs en TAL.

Nous nous sommes particulièrement intéressés à celles se fondant sur la solidarité lexicale de ces expressions, notamment la méthode de [Van de Cruys and Villada Moirón, 2007] qui calcule la préférence d'un verbe pour un nom par rapport aux autres noms du corpus considérés comme ses synonymes (un clustering des noms du corpus ayant été préalablement effectué et tous les noms d'un même cluster étant considérés comme synonymes).

Nous nous sommes inspirés de cette méthode pour mettre au point la seconde partie de notre propre méthode. En effet, nous avons proposé une méthode en deux temps. Nous avons commencé par appliquer un modèle LDA à nos données préalablement clusterisées afin de générer un certain nombre de n-grammes représentatifs du corpus, considérant que les n-grammes les plus représentatifs des thèmes abordés dans le corpus étaient susceptibles de constituer de bons candidats au figement.

Dans un second temps, nous avons généré pour chaque candidat de potentielles expressions synonymes en remplaçant un terme par un autre terme dont la représen-

tation en vecteur de mot était similaire. Nous avons considéré comme une expression figée une expression dont le taux d'apparition par rapport à ses synonymes était en hausse au fil du temps et suffisamment élevé sur les dernières années du corpus.

Les résultats de notre méthode se sont avérés relativement satisfaisants et en tout cas légèrement supérieurs à ceux des mesures statistiques standards et de la méthode Van de Cruys-Villada que nous avons essayé d'appliquer à nos corpus mais qui a donné des résultats très bruités. Les n-grammes retenus qui ne constituent pas des expressions figées à proprement parler demeurent intéressants quant aux informations qu'ils donnent sur le corpus.

Reste à confirmer ces résultats encourageants sur d'autres types de corpus. En effet, nos deux corpus sont assez similaires, étant tous deux des corpus de presse française récoltés sur le même site web et recouvrant une période similaire. Il serait donc intéressant de tester la méthode sur des corpus de genre différents et d'époques différentes afin de vérifier son applicabilité.

BIBLIOGRAPHIE

[Baldwin et al., 2003] Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan. Association for Computational Linguistics. – Cité pages 15, 16 et 22.

[Bauer, 1983] Bauer, L. (1983). *English Word-Formation*. Cambridge Textbooks in Linguistics. Cambridge University Press. – Cité page 15.

[Evert et al., 2004] Evert, S., Heid, U., and Spranger, K. (2004). Identifying morphosyntactic preferences in collocations. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA). – Cité page 26.

[Evert and Krenn, 2001] Evert, S. and Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France. Association for Computational Linguistics. – Cité pages 7, 21 et 22.

[Fazly and Stevenson, 2006] Fazly, A. and Stevenson, S. (2006). Automatically constructing a lexicon of verb phrase idiomatic combinations. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 337–344, Trento, Italy. Association for Computational Linguistics. – Cité page 25.

[Lamiroy, 2008] Lamiroy, B. (2008). Le figement: à la recherche d'une définition. *ZFSL, Zeitschrift für französische Sprache und Literatur*, 36:85–99. – Cité page 13.

[Lapata and Lascarides, 2003] Lapata, M. and Lascarides, A. (2003). Detecting novel compounds: The role of distributional evidence. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics. – Cité page 23.

[Lauer, 1995] Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Compound Nouns*. PhD thesis, Macquarie University. – Cité page 24.

[Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. – Cité page 47.

[Manguin and François, 2006] Manguin, J.-L. and François, J. (2006). Dispute théologique , discussion oiseuse et conversation téléphonique : les collocations adjectivo-nominales au cœur du débat. – Cité page 15.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). Collocations. In *Foundations of Statistical Natural Language Processing*. The MIT Press. – Cité pages 7, 18 et 20.

[Mejri, 2013] Mejri, S. (2013). Figement et défigement : problématique théorique. *Pratiques*, 159-160. – Cité page 14.

[Pearce and Qh, 2001] Pearce, D. and Qh, B. (2001). Using conceptual similarity for collocation extraction. In *Proceedings of the Fourth annual CLUK colloquium*. – Cité page 15.

[Resnik, 1996] Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159. – Cité page 65.

[Resnik, 1993] Resnik, P. S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, USA. UMI Order No. GAX94-13894. – Cité page 65.

[Sag et al., 2002] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg. – Cité pages 15 et 26.

[Savary, 2019] Savary, A. (2019). Expressions polylexicales dans la linguistique computationnelle: on n'est pas sorti de l'auberge. – Cité page 13.

[Van de Cruys and Villada Moirón, 2007] Van de Cruys, T. and Villada Moirón, B. (2007). Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics. – Cité pages 6, 25, 39, 44, 53, 61 et 65.

ANNEXE A

CODE

A.1 Introduction

Nous nous attarderons ici sur quelques extraits des scripts utilisés au cours de notre travail qui nous paraissent intéressants à expliquer.

A.2 Fonctions implémentant les formules de [Van de Cruys and Villada Moirón, 2007]

L'article de [Van de Cruys and Villada Moirón, 2007] présente cinq équations permettant de calculer la préférence d'un verbe pour un nom ou d'un nom pour un verbe, afin d'en déduire le degré de figement d'une expression. Ces équations sont inspirées des travaux de Philip Resnik ([Resnik, 1993] et [Resnik, 1996]).

Les équations 1, 2 et 3 fonctionnent ensemble pour déterminer la préférence du verbe pour le nom tandis que les équations 1.2 (nous appellerons ainsi l'équation 1 dans laquelle nous avons inversé les variables n et v), 4 et 5 permettent de calculer la préférence du nom pour le verbe. Les auteurs présentent ces équations ainsi :

$$S_v = \sum_n p(n|v) log \frac{p(n|v)}{p(n)} \tag{A.1}$$

$$A_{v \to n} = \frac{p(n|v) log \frac{p(n|v)}{p(n)}}{S_v} \tag{A.2}$$

$$R_{v \to n} = \frac{A_{v \to n}}{\sum_{n' \in C} A_{v \to n'}} \tag{A.3}$$

$$A_{n \to v} = \frac{p(v|n) log \frac{p(v|n)}{p(v)}}{S_n} \tag{A.4}$$

$$R_{n \to v} = \frac{A_{n \to v}}{\sum_{n' \in C} A_{n' \to v}} \tag{A.5}$$

La première équation calcule la somme des probabilités pour chaque nom candidat n de se trouver en cooccurrence avec un mot donné (v, un verbe dans l'expérience d'origine, mais n'importe quelle classe grammaticale dans notre version). Il s'agit de la divergence de Kullback-Leibler, une mesure de dissimilarité entre deux distributions de probabilités. Elle est utilisée comme normalisation pour l'équation 2.

L'équation 2 calcule la probabilité pour un des noms candidats n d'apparaître avec un mot particulier v, par rapport à la somme des probabilités de tous les noms candidats d'apparaître avec ce même mot v (calculée par l'équation 1).

L'équation 3 calcule un ratio entre le résultat de l'équation 2 et la somme des résultats de cette même équation 2 pour tous les noms n' appartenant au même cluster que le nom n (autrement dit ses synonymes).

Les équations 1.2, 4 et 4 opèrent les mêmes calculs en inversant n et v.

Voici comment nous les avons implémentées en Python :

```
def equation1(v, cnt_cooc, unigram_freq):
    result = 0
    for n in [key for key, value in cnt_cooc.items() if value[v]
    ↪  > 0]:
        proba_nv = cnt_cooc[n][v]/unigram_freq[v]
        proba_n = unigram_freq[n]/sum(unigram_freq.values())
        result += proba_nv*math.log(proba_nv/proba_n,2)
    return result

def equation1_v2(n, cnt_cooc, unigram_freq):
    result = 0
    for v in [key for key, value in cnt_cooc[n].items() if value
    ↪  > 0]:
        proba_vn = cnt_cooc[n][v]/unigram_freq[n]
        proba_v = unigram_freq[v]/sum(unigram_freq.values())
        result += proba_vn*math.log(proba_vn/proba_v,2)
    return result

def equation2(n, v, cnt_cooc, unigram_freq):
    proba_nv = cnt_cooc[n][v]/unigram_freq[v]
    proba_n = unigram_freq[n]/sum(unigram_freq.values())
    try:
        return (proba_nv*math.log(proba_nv/proba_n,
        ↪  2))/equation1(v, cnt_cooc, unigram_freq)
    except:
        return 0

def equation3(n, v, cnt_cooc, unigram_freq, most_fqt_nouns,
↪  clusters):
    probas_synonymes = 0
    for synonyme in [most_fqt_nouns[i] for i in
    ↪  range(len(most_fqt_nouns)) if clusters[i] ==
    ↪  clusters[most_fqt_nouns.index(n)] and i !=
    ↪  most_fqt_nouns.index(n)]:
        probas_synonymes += equation2(synonyme, v, cnt_cooc,
        ↪  unigram_freq)
    try:
        return equation2(n, v, cnt_cooc,
        ↪  unigram_freq)/probas_synonymes
    except:
```

```
        return 0

def equation4(n, v, cnt_cooc, unigram_freq):
    proba_vn = cnt_cooc[n][v]/unigram_freq[n]
    proba_v = unigram_freq[v]/sum(unigram_freq.values())
    try:
        return (proba_vn*math.log(proba_vn/proba_v,
        ↪ 2))/equation1_v2(n, cnt_cooc, unigram_freq)
    except:
        return 0

def equation5(n, v, cnt_cooc, unigram_freq, most_fqt_nouns,
↪ clusters):
    probas_synonymes = 0
    for synonyme in [most_fqt_nouns[i] for i in
    ↪ range(len(most_fqt_nouns)) if clusters[i] ==
    ↪ clusters[most_fqt_nouns.index(n)] and i !=
    ↪ most_fqt_nouns.index(n)]:
        probas_synonymes += equation4(synonyme, v, cnt_cooc,
        ↪ unigram_freq)
    try:
        return equation4(n, v, cnt_cooc,
        ↪ unigram_freq)/probas_synonymes
    except:
        return 0
```

Nous avons conservé n et v comme noms de variables bien que, lors de notre reconstitution de l'expérience, nous n'ayions pas limité v aux verbes du corpus.

La variable *cnt_cooc* fait référence à un dictionnaire ayant pour clé un des 1 000 noms sélectionnés et comme valeur un dictionnaire de type Counter donnant pour chacun de ses cooccurrents la fréquence de leur apparition ensemble (quel que soit l'ordre des termes dans le bigramme).

Utilisé dans les équations « finales » 3 et 5, *clusters* est un vecteur attribuant à chacun des 1 000 noms un cluster représenté par un nombre. Les éléments ayant le même label appartiennent au même cluster et sont donc considérés comme synonymes. Leur indices permettent de retrouver les noms correspondant dans la liste de noms *most_fqt_nouns*.

Enfin, *unigram_freq* est un dictionnaire Counter donnant la fréquence de chaque mot du texte.

A.3 Extraits du script *create_clustered_data.py*

Ce script, trop long pour être présenté dans son intégralité, prend un corpus au format json en entrée, effectue un clustering sur les articles de chaque année puis une recherche de thématiques pour chaque cluster. Il génère un nouveau fichier json avec ces informations. Il nous semble intéressant d'expliquer un peu plus précisément la fonction utilisant le LDA pour générer les ngrams représentatifs des thématiques de chaque cluster.

```
def get_lda(cluster_docs, annee, taille_ngrams):

    """Prend en entrée la liste des clusters (articles
    ↪  concaténés) d'une année, puis retourne pour chaque
    ↪  cluster les ngrams représentant les topics principaux
    ↪  (10 ngrams en tout)"""

    vectorizer = TfidfVectorizer(token_pattern='(?u)\\b\\w+\\b',
    ↪  ngram_range=(taille_ngrams,taille_ngrams))
    count_data = vectorizer.fit_transform(cluster_docs)
    lda = LDA(n_components=len(cluster_docs), n_jobs=-1)
    clusters_lda = lda.fit_transform(count_data)
    words = vectorizer.get_feature_names()

    clusters_topics = []
    for j, cluster in enumerate(clusters_lda):
        topics_words = []
        topics_index = cluster.argsort()[::-1]
        selected_topics = []
        i = prob = 0
        while prob <= 0.5 and len(selected_topics) < 10:
            selected_topics.append(topics_index[i])
            prob += cluster[topics_index[i]]
            i+=1

        terms_per_topic = ceil(10/len(selected_topics))
        for topic in selected_topics:
            for id_term in
            ↪  lda.components_[topic].argsort()[::-1][:terms_per_topic]:
                if words[id_term] not in topics_words:
                    topics_words.append(words[id_term])

        clusters_topics.append(topics_words)

    return clusters_topics
```

Afin d'utiliser le modèle LDA, il nous faut d'abord formater les données sous forme d'une matrice de nombres : nous utilisons *TfidfVectorizer* pour représenter chaque cluster d'une année (tous les articles d'un même cluster sont concaténés pour former un seul texte) par un vecteur de la taille du vocabulaire et donnant le score **TF-IDF** de chaque token. Le TF-IDF est une mesure permettant de pondérer la fréquence d'un terme dans un texte par sa fréquence dans le corpus : on multiplie la fréquence d'un terme dans un texte (TF : Term Frequency) par l'IDF de ce terme dans l'ensemble du corpus (l'IDF ou Inverse Document Frequency consiste à diviser le nombre de documents du corpus par le nombre de documents du corpus dans lesquels le terme apparaît). Ici le document est un cluster et le corpus tous les clusters d'une même année.

Une fois la matrice créée, il nous faut paramétrer le nombre de topics que nous souhaitons avoir par cluster. Il s'agit du paramètre *n_components* que nous avons décidé de fixer au nombre de clusters de l'année, ce qui semble logique si l'on consi-

dère que le clustering a rapproché des documents probablement proches dans leurs thématiques.

En appliquant le modèle LDA à nos clusters, nous récupérons deux matrices : *clusters_lda* et *lda.components_*. La première matrice contient un vecteur par cluster de la taille du nombre de topics (ici égale au nombre de clusters de l'année) : pour chaque topic, elle donne une probabilité que le document soit à ce sujet. La seconde matrice contient un vecteur par topic de la taille du vocabulaire et donne pour chaque ngram la probabilité qu'il soit associé au topic.

La fonction *argsort* nous permet de trier en ordre décroissant les index en fonction de leur valeur : on obtient donc une liste des labels des topics (leur position dans le vecteur) du topic le plus probable du document au topic le moins probable. Nous avons décidé, pour chaque vecteur, de récupérer les premiers topics jusqu'à avoir atteint 50% de représentativité (c'est-à-dire jusqu'à atteindre 0.5 en additionnant leurs probabilités) tout en limitant le total de topics sélectionnés à dix afin de ne pas avoir un nombre de topics trop important si l'algorithme a été trop hésitant, et également pour avoir au moins un ngram par topic (puisque nous avons choisi de représenter le cluster par un total de dix ngrams).

Pour chaque topic sélectionné, nous utilisons une nouvelle fois *argsort* sur le vecteur de la matrice *lda.components_* correspondant au topic afin de récupérer les positions des ngrams les plus souvent associés à ce topic (au nombre de dix divisé par le nombre de topics sélectionnés pour représenter le cluster, afin d'avoir un total de dix ngrams par cluster). On récupère les ngrams grâce à leur position dans la liste *words* générée avec *TfidfVectorizer*.

ANNEXE B

GRAPHIQUES ET TABLEAUX

B.1 Stream graphs de la section 4.2

Ces graphiques présentent, pour chaque mesure statistique, les scores obtenus chaque année par les trigrammes obtenant pour au moins deux années un des dix scores les plus elevés. Le *Stream Graph* (ou *Stacked Area Graph*) est un type de graphique en aires superposées construit autour d'un axe central. Il permet ici de visualiser les évolutions des scores obtenus par les ngrams au fil des années. Le tutoriel m'ayant aidée à implémenter ces graphiques se trouve ici : https://towardsdatascience.com/a-quick-introduction-into-stream-graphs-with-python-33ff4493ccc. La documentation du module altair.Chart peut être consultée à cette adresse : https://altair-viz.github.io/user_guide/generated/toplevel/altair.Chart.html.

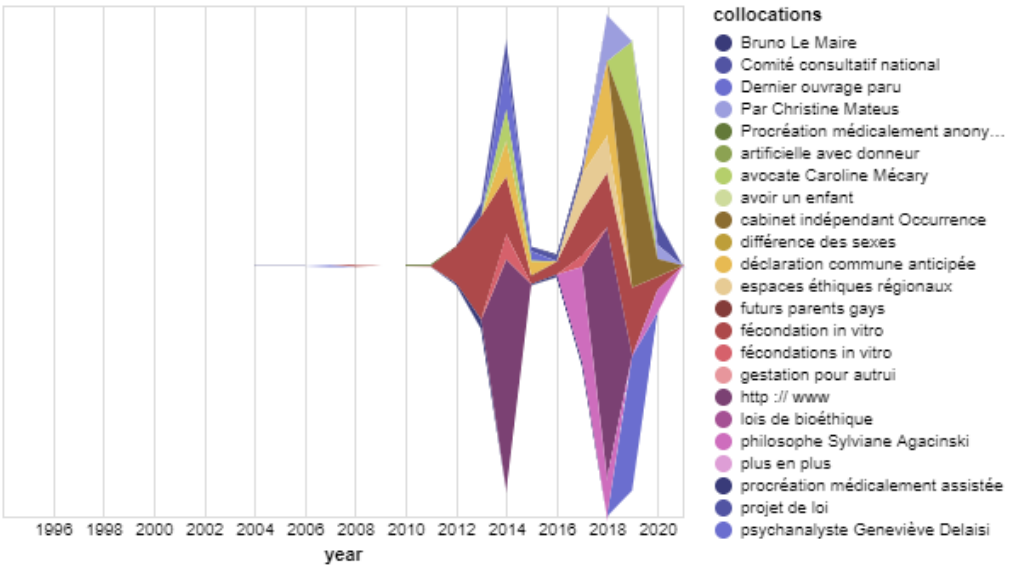

FIGURE B.1 – Trigrammes détectés avec χ^2 (corpus PMA)

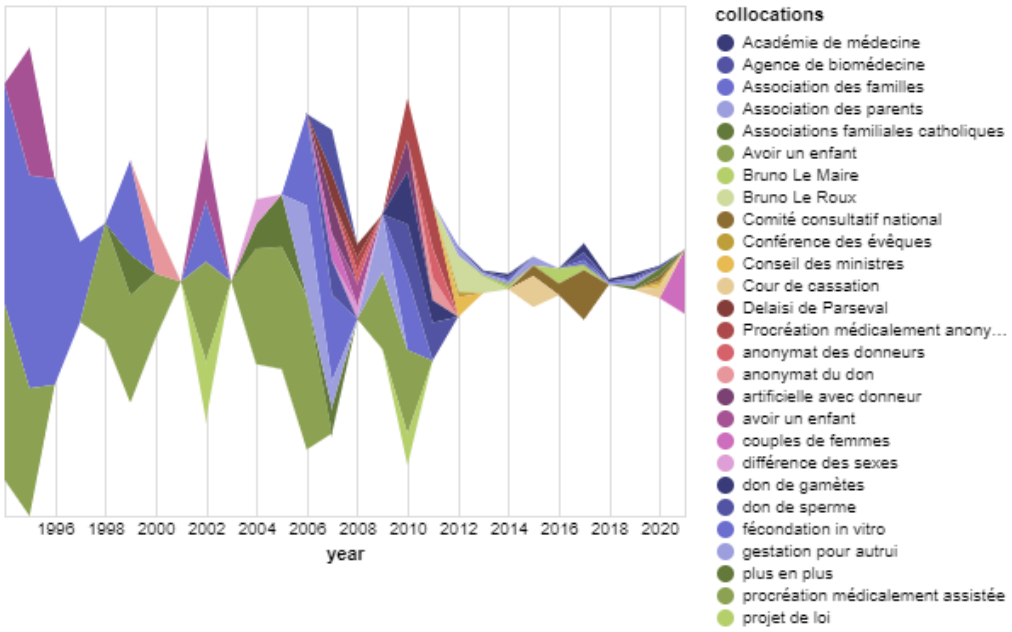

FIGURE B.2 – Trigrammes détectés avec fréquence (corpus PMA)

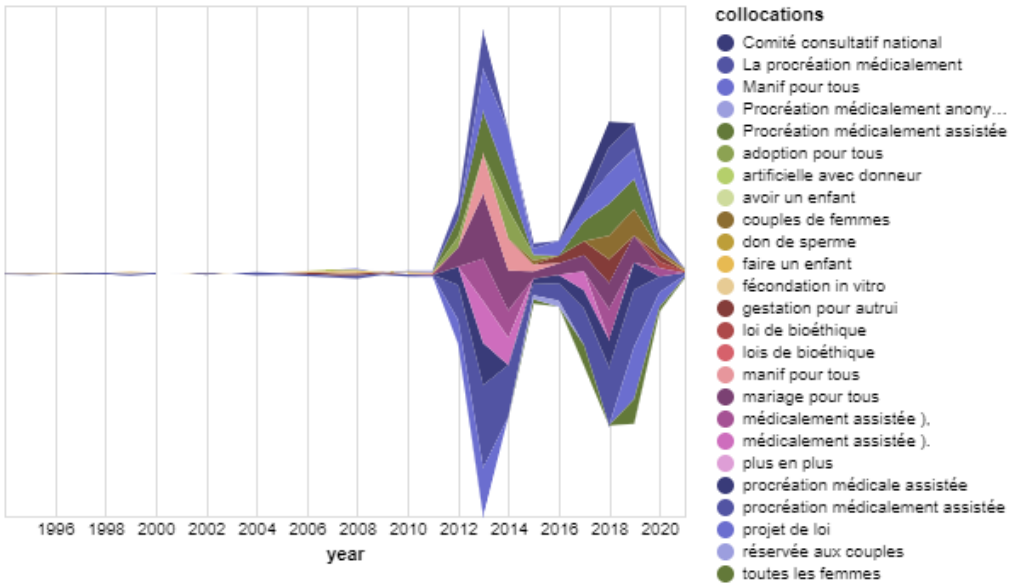

FIGURE B.3 – Trigrammes détectés avec fonction de vraisemblance (corpus PMA)

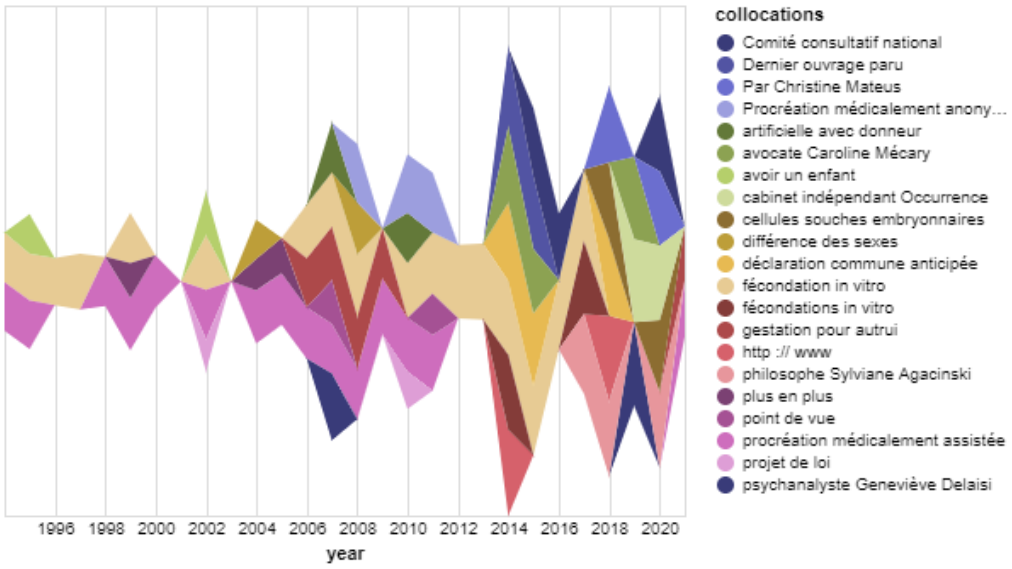

FIGURE B.4 – Trigrammes détectés avec PMI (corpus PMA)

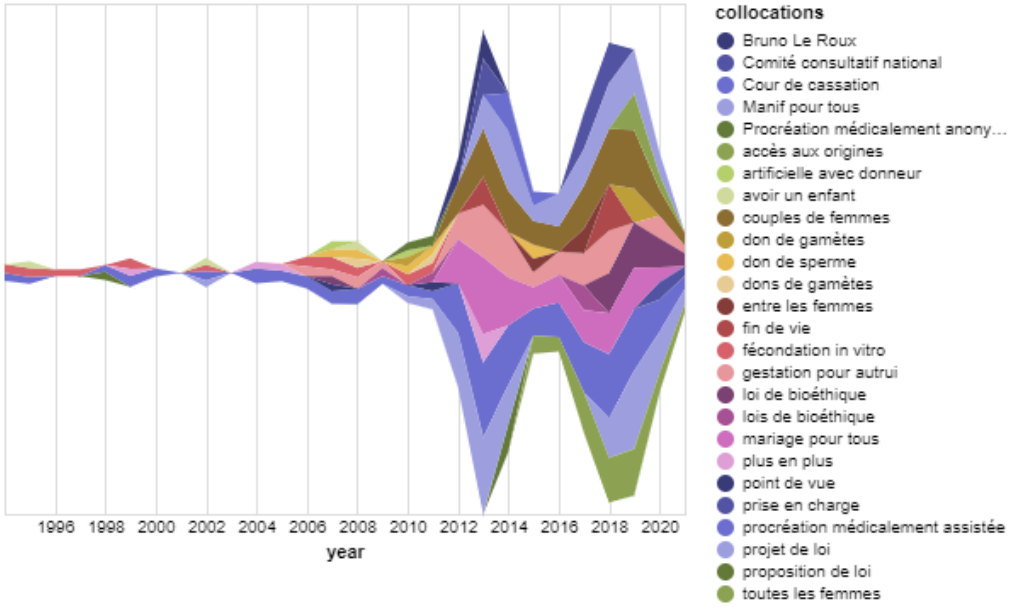

FIGURE B.5 – Trigrammes détectés avec test t (corpus PMA)

B.2 Tableaux de comparaison des résultats, corpus Sackler

Résultats pertinents	Résultats discutables	Résultats non pertinents
	new york	

TABLE B.1 – Résultats communs avec mesures standards (corpus Sackler)

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
crise des opiacés	of art, sackler gallery, purdue pharma, nan goldin, etats unis, johnson johnson, arthur m sackler, metropolitan museum of, museum of art, espérance de vie, metropolitan museum of art, arthur m sackler gallery	antiquités orientales, famille sackler, membres de la famille

TABLE B.2 – Résultats non trouvés par les mesures standards (corpus Sackler)

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
	serpentine sackler gallery, art media agency, gallery of art, freer gallery of, metropolitan museum	millions de dollars, millions de francs, département des antiquités

TABLE B.3 – Résultats trouvés uniquement par les mesures standards (corpus Sackler)

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
crise des opiacés	new york, of art, sackler gallery, purdue pharma, espérance de vie	antiquités orientales

TABLE B.4 – Résultats communs avec la méthode Van de Cruys-Villada (corpus Sackler)

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
	nan goldin, etats unis, johnson johnson	famille sackler

TABLE B.5 – Résultats non trouvés par la méthode Van de Cruys-Villada (corpus Sackler)

Résultats pertinents	**Résultats discutables**	**Résultats non pertinents**
crise des opioïdes	conseil d administration, art of, m. sackler, freer and, metropolitan museum, londres gallery, mettre fin	xixe siècle, ans imc, millions de dollars, diabète type, milliards de dollars, états américains, morts overdose, fortune estimée, loi faillites, loi protection

TABLE B.6 – Résultats trouvés uniquement par la méthode Van de Cruys-Villada (corpus Sackler)