Contributions sur la structure morphosyntaxique des graphies terminologiques et sur l'hybridation entre terminologie et modèles de thèmes

Amaury Delamaire


N°d'ordre NNT : 2020LYSEM016

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Ecole des Mines de Saint-Etienne

Ecole Doctorale N° 488
Sciences, Ingénierie, Santé

Spécialité de doctorat :
Discipline : Informatique

Soutenue publiquement le 12/10/2020, par:
Amaury Claude Silvère DELAMAIRE

Contributions sur la structure morphosyntaxique des graphies terminologiques et sur l'hybridation entre terminologie et modèles de thèmes

Devant le jury composé de :

Calabretto, Sylvie Professeur des universités CNRS et INSA Lyon Présidente
Tannier, Xavier Professeur des universités CNRS et Sorbonne université Rapporteur
Bellot, Patrice Professeurs des universités CNRS et Aix Marseille Université Rapporteur

Beigbeder, Michel Directeur de recherche CNRS et Écoles des Mines de Saint-Étienne Directeur de thèse
JUGANARU-MATHIEU, Mihaela Professeur des universités École des Mines de Saint-Étienne Co-directrice de thèse

Résumé

Nous présentons ici diverses expériences et hypothèses en lien avec l'extraction terminographique automatique et de potentielles hybridations avec des modèles de thèmes. Dans le domaine du TAL, la construction automatique de terminologies n'est que peu consensuelle. Les différents objectifs des chercheurs font poindre des divergences d'opinion quant à ce qui constitue ou non une unité terminologique. Les divergences se situent à différents niveaux de la tâche.

Sur le plan linguistique, les chercheurs sont parvenus à un accord relatif quant à la structure morphosyntaxique des graphies terminologiques. De nouvelles propositions apparaissent régulièrement mais qui complètent le consensus plus qu'elles ne l'invalident. Si la structure des graphies fait consensus, il n'en est pas de même pour leur caractérisation en tant qu'unité terminologique. L'aspect terminologique d'une unité est déterminé à partir de différents facteurs internes – la structure que nous venons d'évoquer – ainsi qu'externes – contexte, contraste.

Notre contribution concerne les deux points évoqués ci-dessus. Dans un premier temps nos expériences portent sur le contexte d'apparition des unités terminologiques à partir de modèles de thèmes. Nous verrons si et comment les unités terminologiques peuvent bénéficier à la construction de modèles de thèmes. Ce bénéfice sera estimé à l'aune de la pertinence des modèles construits et de diverses mesures statistiques. Dans un second temps, nous proposerons une extension de la structure morphosyntaxique consensuelle des graphies terminologiques. Nous montrerons comment notre proposition est adéquate mais difficile à évaluer dans un contexte d'automatisation du processus d'extraction. Enfin nous proposerons une formalisation stricte d'un protocole d'évaluation basé sur celui de la recherche d'information, méthode déjà employée dans la littérature sur la terminologie sous différentes formes mais non détaillée.

Dédicace

À mes parents, mon frère et mon neveu.

Remerciements

Je remercie châleureusement toutes les personnes qui m'ont soutenu dans cette entreprise :

- Michel Beigbeder et Mihaela Mathieu pour leur patience, leur pédagogie, leur compréhension et leur soutien.
- Bruno Léger pour avoir cru en mon profil atypique et pour sa compréhension.
- L'Ecole des Mines pour m'avoir permis de réaliser ma thèse, qui plus est dans de bonnes conditions.
- L'ensemble du jury ; une soutenance à distance n'aurait pas pu mieux se dérouler.
- Jean-François Tchebanoff pour m'avoir permis de travailler tranquillement et pour nos discussions.
- Anaëlle Leroyer pour son soutien indéfectible et le temps passé à corriger mes étourderies linguistiques.
- L'entreprise Storyzy et plus particulièrement son directeur technique Ramòn Ruti qui m'ont permis de concilier thèse et activité professionnelle.
- Toute ma tribu : Florian, Axel, Alexandre, Hugo, Marie & Marie, Margot, Steven, Manon, Jonathan, ... pour leur soutien permanent même pendant les périodes difficiles.
- Mes chats, qui m'ont évité des crises de nerf. 🐾

Merci à tous, je n'y serais pas parvenu seul.

Table des matières

Chapitre 1

Extraction terminologique automatique : un procédé du traitement automatique de la langue

Le *Traitement Automatique de la Langue* ou TAL désigne un ensemble de tâches automatisées qui ont spécifiquement trait à l'analyse de la langue. Les méthodes et objectifs du TAL sont largement ignorés du grand public, qui ne considère que les *emblèmes* que sont les assistants vocaux ou la traduction automatique. Nous tentons dans cette partie d'apporter des éclaircissements sur la nature et la diversité du TAL, ainsi que les motifs qui ont mené à son développement. Une première section introduira le contexte sociétal dans lequel s'insère le TAL : les raisons de son apparition, de son développement, ses perspectives actuelles. Dans une seconde partie, nous nous concentrerons sur la technicité par une présentation de la transversalité du TAL. Nous verrons notamment comment le TAL sollicite sciences sociales, sciences dures et ingénierie informatique. Dans une troisième et dernière partie, nous présenterons des notions et concepts dont la compréhension est nécessaire pour nos discussions sur la construction automatique de terminologie et son exploitation.

1.1 Contexte et enjeux des techniques de traitement automatique de la langue

Le développement du TAL vise à répondre à différentes attentes de la société en lien avec l'avènement de l'informatique. La mondialisation du partage d'informations quasi-instantané a fait émerger diverses problématiques que le TAL tente de résoudre.

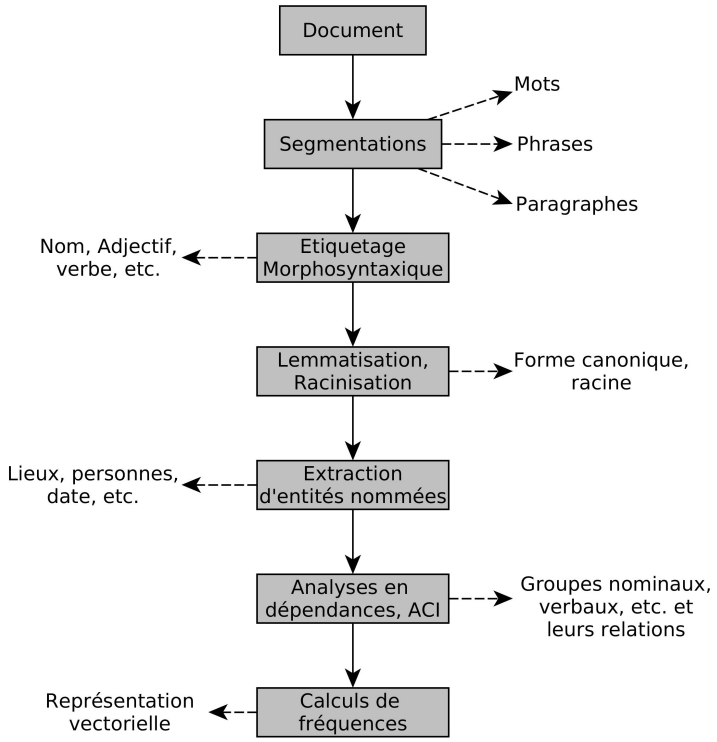
1.1.1 Quantité et variété des données

La quantité de données numériques produites est en constante augmentation. L'entreprise Domo, spécialisée dans la visualisation de données, a réalisé en 2018 une étude sur la quantité de données numériques sauvegardées et a abouti à l'estimation d'un quintillion (10^{30}) octets quotidiennement [1] [2], soit plus de 10 millions d'exaoctets à la milliseconde – 1 exaoctet vaut 1 024 pétaoctets, 1 pétaoctet vaut 1 024 téraoctets. Parmi ces quantités astronomiques de données, le TAL se concentre sur des données spécifiques, en l'occurrence linguistique, soit la parole et le texte.

Avec l'informatique et la surproduction de données, la problématique de l'accès à l'information a été inversée : il y a cinquante ans, avant l'informatique, l'accès à l'information était restreint par l'insuffisance de cette dernière ; il est maintenant difficile de trier parmi toutes les informations à disposition. Les techniques du TAL visent à extraire de la connaissance de ces ensembles de données à diverses fins. Parmi les objectifs concrets du TAL, nous pouvons évoquer la traduction automatique, le résumé automatique de multiples documents, la rédaction automatique d'articles d'actualité, etc., ainsi que des tâches spécifiques à la langue parlée avec la reconnaissance et la synthèse de la parole. Pour fonctionner, ces tâches complexes s'appuient sur une succession de traitements plus succincts : la chaîne de traitements ou *pipeline*.

1. `https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/`
2. `https://www.domo.com/learn/data-never-sleeps-5`

FIGURE 1.1 – Exemple de chaîne de traitements linguistiques appliqués à un document.



1.1.2 La notion de chaîne de traitements

L'ensemble des tâches complexes du TAL s'appuie sur une succession de traitements invisibles pour l'utilisateur final. Certains sont très abordables, comme les segmentations en mots et phrases, d'autres requièrent des connaissances spécialisées. L'automatisation d'un processus implique nécessairement une part d'erreur dans les résultats, chaque étape de la chaîne de traitements produit donc une certaine proportion d'erreurs. Dans le contexte d'une chaîne de traitements, le résultat de l'analyse passe d'étape en étape jusqu'à l'obtention du résultat.

Le graphique 1.1 illustre le concept de chaîne de traitements ainsi que la propagation d'erreurs : des erreurs du dernier traitement peuvent être liées à des erreurs du premier. Il est communément admis qu'il est impossible de développer un module automatisé fiable à 100%, quelle que soit la tâche :

même la segmentation en mots, base de tout autre traitement, est problématique. De ce fait, plus la tâche induit de nombreux sous-traitements plus la proportion finale d'erreurs sera élevée.

Au delà de la nécessité technique d'une chaîne de traitements, la division en modules précis permet de les réutiliser pour d'autres tâches ou de les remplacer aisément. Parmi les traitements de TAL de base consensuellement définis, nous pouvons identifier par ordre croissant de complexité :

Segmentation en mots : En informatique, un texte est considéré comme une suite de caractères : le système ne sait pas ce qu'est un mot. Si considérer les espaces entre les mots permet généralement de les délimiter, il demeure des exceptions relativement fréquentes. Par exemple ***Moi,*** *je* ***n'aime*** *pas ça* avec une virgule accolée et un *ne* contracté. La problématique est par contre tout autre quand le texte est constitué d'idéogrammes ou rédigé dans une langue (semi-)agglutinante : la notion même de *mot* est remise en question.

Segmentation en phrases : Comme pour les mots, la segmentation en phrases vise à délimiter les unités phrastiques dans un texte. Comme pour les mots, une méthode simple recouvre la majorité des cas – l'identification de la ponctuation forte.

Lemmatisation/Racinisation : Méthodes qui consistent à trouver le lemme ou la racine d'un mot. Le lemme est sa forme canonique : la forme sous laquelle il apparaît dans un dictionnaire. Sa racine désigne sa base fixe, la partie du mot qui ne varie pas quel que soit le contexte, malgré la conjugaison et les accords.

Etiquetage morphosyntaxique : Il s'agit d'une méthode nécessaire à de très nombreux traitements du TAL. L'étiquetage morphosyntaxique consiste à déterminer la catégorie grammaticale d'un mot identifié à la première étape. Relativement à la propagation d'erreurs, il est évident qu'une mauvaise segmentation en mots aboutira à de mauvaises étiquettes morphosyntaxiques. La nature de la tâche fait consensus : attribuer une étiquette à chaque mot. En revanche son application varie fréquemment chez les chercheurs, qui utilisent des étiquettes différentes – différences de nom et différences de nombres.

Entités nommées : Les entités nommées désignent initialement les références à des entités du monde réel : les noms de lieux, de personnes et d'organisations – respectivement toponymes, anthroponymes et er-

gonymes [3]. La notion a ensuite été étendue en fonction des besoins des chercheurs, par exemple dans le domaine médical avec l'annotation de maladies et de symptômes. La découverte d'entités nommées se base généralement sur un étiquetage morphosyntaxique préalable – de nombreux mots étiquetés *noms propres* sont des entités nommées.

Analyse en dépendances : L'analyse en dépendances identifie les groupes syntaxiquement et sémantiquement pertinents ainsi que la nature de la relation potentielle qu'ils peuvent entretenir. C'est par exemple l'analyse en dépendances qui va permettre d'identifier le sujet ou l'objet d'un verbe. Il y a consensus sur la définition de la tâche, mais cette dernière étant complexe, des améliorations restent à trouver : les systèmes actuels produisent trop d'erreurs et demandent des machines puissantes.

Analyse en constituants immédiats (ACI) : L'ACI est proche de l'analyse en dépendances en ce qu'elle identifie des groupes pertinents et des relations entre ces groupes. La différence porte notamment sur le fait que les relations détectées sont de nature strictement syntaxique : l'ACI décrit le processus de formation de la phrase à partir de ces groupes.

Analyse et extraction d'arguments : la reconnaissance et l'analyse automatiques des arguments exprimés dans un verbatim est un sujet de recherche actuel. L'automatisation du processus suppose des prétraitements lourds (analyses syntaxiques) mais peut remplir des objectifs multiples [Delobelle et al., 2020] (à paraître).

La liste ci-dessus ne se veut pas exhaustive, elle vise davantage à illustrer les différents points introduits sur les chaînes de traitements et la propagation d'erreurs. Il s'agit néanmoins à notre connaissance des traitements de TAL les plus fréquemment nécessaires à la réalisation de tâches complexes. Les briques intiales – segmentation, étiquetage morphosyntaxique, lemmatisation et racinisation – ont atteint des niveaux satisfaisants pour les langues dotées [4]. A partir de ces briques sont construites les tâches complexes du TAL

3. `https://fr.wikipedia.org/wiki/Entit%C3%A9_nomm%C3%A9e`

4. Une langue dotée, en opposition à une langue peu dotée, dispose d'importantes ressources linguistiques numériques permettant le développement de modules de TAL. La plupart des langues dans le monde sont peu dotées, la mieux dotée est évidemment l'anglais. Le français est relativement bien doté concernant les tâches communes – morphosyntaxe, lemmatisation, entités nommées, etc. – mais moins pour des tâches complexes comme

mentionnées plus haut.

1.1.3 Une absence de consensus entre communautés scientifiques

Une des difficultés que rencontre le TAL émane de la variété des profils de chercheurs concernés par le domaine. Avec des profils variant des sciences sociales aux sciences dures, les solutions proposées pour répondre aux divers objectifs recherchés peuvent être diamétralement opposées. Là où un linguiste va enquêter sur la langue, le statisticien recherchera les lois mathématiques qui la régissent et qui permettent sa description. La situation réelle n'est évidemment pas aussi binaire : comme nous allons le voir, le profil des chercheurs recouvre le plus souvent différents domaines scientifiques concernés par le TAL.

1.2 Le TAL : une discipline transverse

Nous présentons ici les différents domaines scientifiques sollicités par les techniques du TAL. Une analyse du terme TAL met en évidence deux points capitaux : l'automatisation au travers d'algorithmes et de formalisations mathématiques et la linguistique. Sans donner de détails techniques, nous proposons une partition en trois sections : les sciences humaines, les mathématiques et l'implémentation informatique. L'implémentation relève de l'ingénierie davantage que de la recherche ; nous verrons en quoi elle est néanmoins pertinente. Nous développerons dans un premier temps les aspects strictement scientifiques du TAL, à savoir les sciences humaines puis les mathématiques. Dans une troisième et dernière partie, nous détaillerons en quoi l'ingénierie informatique peut conditionner les propositions faites par les chercheurs.

1.2.1 Aspects sciences humaines

Comme son nom l'indique, le TAL traite de la langue, de la linguistique. Le Larousse en propose la définition suivante [5] :

> Science qui a pour objet l'étude du langage et des langues.

l'extraction terminographique voire l'analyse en dépendances

5. https://www.larousse.fr/dictionnaires/francais/linguistique/47271?q=linguistique#47200

La langue est un outil de communication en évolution, qui est conditionné par des éléments externes. Parmi ceux-ci nous en avons identifiés deux principaux qui concernent nos recherches : l'élément sociologique et l'élément psychologique, les deux pouvant être interdépendants. Nous expliciterons leurs liens avec le TAL après avoir présenté le rôle de la linguistique dans l'automatisation de l'analyse de la langue. D'autres éléments externes auraient pu être retenus – anthropologie, éducation, démographie, etc. – mais que nous ne développerons pas.

1.2.1.1 La linguistique

La science linguistique est l'un des fondements du TAL. La linguistique est communément admise comme une science qui étudie la forme de la langue, sa signification et son contexte. Autrement dit son origine, son état et son évolution. Selon Noam Chomsky, la linguistique fait partie des sciences cognitives. Elle comprend également des branches plus spécifiques et parfois éloignées comme la sémiotique (étude des signes et de leur signification), la stylistique, la traduction, la pédagogie de l'apprentissage ou la correction des troubles du langage. Notre cœur de recherche, la terminologie, se situe entre différents sous-domaines de la linguistique : la sémantique pour l'interprétation du sens, l'étymologie pour l'origine du mot, la sociolinguistique pour son contexte d'utilisation et plus généralement la lexicologie, qui porte sur l'analyse du lexique.

Le TAL peut être considéré comme un outil mis à disposition des linguistes pour leur permettre des observations et analyses jusque-là impossibles du fait des limites humaines. Les linguistes travaillaient fréquemment sur des extraits de textes ou des exemples dont ils ne pouvaient estimer la représentativité relativement à l'usage ; la linguistique de corpus, propre à répondre à cette problématique, n'est que difficilement réalisable par un humain sur des quantités de données importantes.

Le logiciel AntConc [6] propose par exemple diverses analyses de manière très intuitive et accessible via une interface graphique : cooccurrences, concordanciers, calculs de fréquences, etc., et ce en un laps de temps relativement court au regard de la taille du corpus analysé. Les analyses produites quasi-instantanément par AntConc auraientt pris un temps considérable à un linguiste ; elles lui permettront ensuite de vérifier ses hypothèses de manière

6. `http://explorationdecorpus.corpusecrits.huma-num.fr/antconc/`

plus large.

1.2.1.2 La sociologie

Dans nos recherches, l'aspect sociologique est particulièrement présent au travers des notions de terminologie et de néologie. Comme nous le verrons dans les sections suivantes, une terminologie consiste en un vocabulaire discriminant entre groupes de personnes spécialisées. Autrement dit un vocabulaire technique propre à une communauté et généralement abscons pour les personnes non spécialistes. La décision de qualifier un terme de terme technique ou non est floue, pour deux raisons principales. Premièrement, la technicité d'un terme peut varier avec le temps, comme nous avons pu le constater avec les termes en lien avec l'informatique, réservés en premier lieu aux initiés. Ensuite, la technicité d'un terme est un jugement humain, donc subjectif : l'accoutumance d'une personne à l'utilisation d'un terme peut la pousser à le considérer comme générique. Le jugement peut aussi être biaisé par les fins de la terminologie construite, à savoir l'utilisation prévue pour la ressource construite.

La néologie concerne quant à elle soit l'apparition de nouveaux mots, soit l'apparition de nouveaux sens pour des mots pré-existants. Elle est donc intrinsèquement liée au temps qui passe et à l'usage des néologismes : la caractérisation de *néologique* n'est valable que sur un laps de temps limité, au terme duquel le néologisme peut disparaître – par manque d'usage – ou entrer dans le language courant. Le néologisme peut parfois être terminologique, comme c'est le cas pour le mot *souris* qui est une occurrence de néologie sémantique terminologique : au sens originel de *petit rongeur* a été ajouté celui plus spécialisé de *périphérique de pointage*. Nous pouvons voir ici comment les phénomènes de néologie et de terminologies peuvent être complémentaires.

De manière plus générale, l'aspect sociologique peut être associé aux étapes d'annotations manuelles de corpora. L'annotation manuelle de corpora consiste à faire annoter par un ou plusieurs *experts* des éléments linguistiques spécifiques. Les annotations peuvent être simples et sans ambiguïté – étiquetage morphosyntaxique, lemmatisation, etc. – ou à l'inverse complexes et potentiellement ambigües – dépendances, sentiments, technicité, subjectivité, omission [Alonso et al., 2017], etc. Le développement des tâches complexes de TAL a développé le besoin pour davantage de fiabilité et de robustesse dans les étapes d'annotation. Pour ce faire, des métriques d'évaluation ont

été proposées afin de comparer les annotations de plusieurs experts sur un même corpus : les mesures d'accord inter-annotateurs [Fort, 2016].

1.2.1.3 La psychologie

Nous avons présenté les différents liens du TAL avec certaines sciences humaines, en l'occurrence la linguistique et la sociologie. Nous abordons ici ses relations avec la psychologie, ou plutôt la psychologie appliquée à la linguistique. La psychologie, au même titre que la sociologie, est à prendre en compte lors de l'établissement des limites de phénomènes linguistiques. C'est par exemple le cas pour la néologie et pour la terminologie qui impliquent de nombreuses considérations extralinguistiques telles que la chronologie, le niveau de consensus ou encore les objectifs des chercheurs. Les notions de néologie et de terminologie – et d'autres : subjectivité, émotion, etc. – sont particulièrement affectées par la psychologie humaine en ce qu'elles reposent sur des jugements particuliers : si la nature d'un adjectif est explicite, la détermination de la subjectivité, de l'émotion, de la néologie, etc., ne l'est pas. Une part d'interprétation est laissée à l'expert qui annote le corpus, d'où un consensus parfois marginal selon les axes de recherche. La multiplication des annotateurs associée à des métriques d'évaluation adaptées a permis de mettre en évidence des domaines de recherches complexes à annoter car la décision de l'annotateur peut varier d'une personne à l'autre. Les campagnes d'annotation terminologiques sont particulièrement sujettes à ce biais psychologique des annotateurs : la notion de qualité terminologique est peu consensuelle, les annotateurs en ont généralement une vague esquisse mentale mais sans davantage de précision.

La psychologie peut également expliquer le comportement de certains annotateurs lors de la construction d'une référence. Nous l'avons vu, les campagnes d'annotation sont inégales dans leur complexité. Afin de pallier l'incertitude d'un annotateur humain seul lorsque l'annotation est complexe, la campagne est étendue à un ensemble d'experts : chaque segment est annoté par plusieurs experts, ce qui permet d'obtenir des mesures de consensus entre annotateurs. Afin de construire des références volumineuses, les chercheurs font fréquemment appel à des annotateurs externes : on parle alors de *crowd sourcing*.

Lors d'une campagne, les annotateurs suivent un guide d'annotation qui spécifie les éléments à identifier ainsi que la manière de procéder. L'annotation d'une référence commence par une étape dite *d'apprentissage* durant

laquelle les annotateurs apprennent à appliquer les prérogatives décrites dans le guide d'annotation. Selon la complexité de la tâche et du guide associé, la longueur de la courbe d'apprentissage peut varier. Une campagne d'annotation de catégories grammaticales peut par exemple présenter des périodes d'apprentissage particulièrement distinctes relativement au jeu d'étiquettes employé : davantage d'étiquettes implique un apprentissage plus long, et réciproquement un jeu d'étiquettes restreint induit un apprentissage plus court.

Nous avons précédemment évoqué la fiabilité des annotateurs, fiabilité qui peut se manifester sous deux aspects : i) la capacité de l'annotateur à apprendre rapidement le guide d'annotation et à l'appliquer, ii) la volonté de l'annotateur d'effectuer un travail de qualité. Le premier aspect est à prendre en compte en fonction de la tâche : pour une petite campagne d'annotation un apprentissage lent n'est pas envisageable, et inversement pour une campagne plus développée. Le second aspect est à considérer lors d'une campagne d'annotation participative – *crowd sourcing*. La multiplication d'annotateurs non experts, potentiellement distants – voir AMAZON MECHANICAL TURK [7] – induit une baisse de la qualité des annotations relativement à celles d'experts *standards*.

Nous pouvons sembler nous éloigner de l'aspect psychologique du TAL en élaborant la notion de campagne d'annotations, ce n'est cependant pas le cas : le guide d'annotations doit être clair et non ambigu afin de minimiser les courbes d'apprentissage des annotateurs ; ces courbes sont directement liées à i) l'éducation de l'annotateur et ii) son implication dans la tâche. De manière générale, que les annotations soient triviales ou fortement spécialisées, la tâche est intrinsèquement liée aux profils psychologiques des chercheurs ainsi que des annotateurs – ce sont parfois les mêmes personnes. Le guide d'annotations tente de décrire les hypothèses des chercheurs avec les biais qui lui sont propres, les annotateurs tentent d'interpréter ce guide afin de l'appliquer. La constitution du guide d'annotation est à prendre en considération avant de démarrer une campagne. Une période d'apprentissage du guide est à prévoir au début de la campagne, période qui peut être plus ou moins longue selon la complexité de la tâche et la qualité du guide d'annotation.

Nous avons choisi d'introduire l'aspect psychologique du TAL à partir des campagnes d'annotation manuelles car elles constituent un socle largement commun à de nombreuses – si ce n'est toutes – tâches. Qu'ils s'agissent de catégories grammaticales, de subjectivité, de terminologie ou encore d'enti-

7. `https://www.mturk.com/`

tés nommées, la constitution d'un corpus qui permettra l'évaluation est un prérequis au développement des différents systèmes. Des méthodes automatisées ont été proposées pour constituer ce corpus de référence sans intervention humaine – du moins de façon marginale – mais sans avoir parvenu à s'affranchir des divers biais qui rendent ces ressources caduques : un automatisme signifie une formalisation de la problématique ; une formalisation fixe implique la non prise en compte des évolutions de cette formalisation. Prenons un exemple trivial : si une terminologie de référence construite automatiquement ne contient que des groupes nominaux, elle ne permettra pas l'évaluation de segments terminologiques verbaux. Notre exemple peut être étendu à d'autres domaines que celui de la terminologie : une annotation limitée a priori – aux groupes nominaux, aux verbes, ou à quoi que ce soit d'autre – ne permettra pas des évaluations exhaustives lors d'applications qui sortent de ces contraintes a priori. La conclusion est qu'à notre connaissance la construction de corpora de référence assistée par des automatismes produit généralement des ressources incomplètes, spécifiques à des applications données. C'est particulièrement vrai pour des tâches complexes et fastidieuses qui ne peuvent être réalisées que par des publics avertis – des experts – dont le temps est une ressource précieuse.

En dehors de l'annotation, fréquente en TAL, certaines tâches sont parfaitement centrées sur l'aspect psychologique et les sciences cognitives en général ; c'est par exemple le cas des jeux dits *sérieux* – ou serious games – qui visent à dissimuler l'objectif d'une application au milieu d'un divertissement interactif. L'utilisateur participe à un processus qui semble s'apparenter au divertissement mais qui vise en réalité à assembler des connaissances diverses, produites par l'utilisateur. Dans certains cas, les jeux sérieux peuvent être considérés comme des alternatives aux campagnes d'annotation participatives : au lieu d'être dans un cadre de travail explicite, l'annotateur partage ses connaissances au travers d'une interface distrayante mais entretenant les mêmes objectifs. Nous nous concentrons ici sur ces jeux sérieux en particulier. Dans le cadre du TAL, le développement des jeux sérieux se justifie par des critères économiques. Les annotateurs qui créent la référence peuvent se diviser en quelques grands groupes :

— Les chercheurs eux-mêmes : solution pragmatique et de facilité, le coût de la tâche est nul excepté en terme de temps alloué. La qualité/pertinence des annotations peut être remise en question : les chercheurs ne sont pas experts de tous les domaines.

— Les élèves des chercheurs : les chercheurs sont souvent également en-

seignants ; il n'est pas rare qu'ils exploitent leurs étudiants pour des annotations manuelles moyennant des crédits – universitaires – supplémentaires.
— Les annotations participatives : tout le monde. N'importe quelle personne prétendant pouvoir répondre au besoin exprimé peut participer. Les campagnes participatives nécessitent le déploiement d'outils de suivi tels que ceux évoqués ci-dessus, ainsi qu'un budget.
— Les experts : les travaux de recherches impliquent régulièrement des partenaires industriels, qui peuvent dès lors faire participer leurs experts aux campagnes d'annotation. Ces annotations sont vraissemblablement les plus fiables.

La solution proposée par les jeux sérieux permet de s'affranchir du coût lié aux campagnes participatives en développant un cadre qui incite les utilisateurs à participer d'eux-mêmes à la campagne. Les jeux sérieux développent donc un double objectif : l'objectif principal – acquérir des connaissance – et l'ergonomie / l'expérience utilisateur – faire en sorte que l'intérêt de l'utilisateur passe outre l'objectif principal. L'objectif de l'utilisateur reste la distraction même s'il peut trouver un intérêt à partager ses connaissances.

Nous avons présenté plusieurs domaines des sciences humaines concernés par le TAL de manière non exhaustive mais pertinente au regard de nos recherches. De la même manière, nous allons développer maintenant les domaines des sciences dures sollicités par le TAL, plus spécifiquement les facettes mathématiques des solutions de TAL.

1.2.2 Aspects mathématiques

Les solutions de TAL proposées dans la littérature s'appuient très largement sur plusieurs sous-domaines des mathématiques intrinsèquement liés : les statistiques, les probabilités ainsi que les espaces de représentation. Nous l'avons brièvement évoqué, dans les analyses du TAL les textes sont représentés à partir de modèles statistiques, le plus commun étant le *sac-de-mots*. Dans un modèle en sac-de-mots, un texte est représenté par les fréquences des mots du vocabulaire – i.e. un vecteur multidimensionnel de valeurs numériques. Plus trivialement, le vecteur créé décrit partiellement le texte d'origine : comme son nom l'indique, la représentation sous forme de sac-de-mots induit – entre autres – la perte de la connaissance de l'ordre des mots à l'origine. Prenons un exemple simple.
Exemple :

Texte : *Le* TAL *est un sous-domaine de l'apprentissage automatique mais l'apprentissage automatique n'est pas un sous-domaine du* TAL
Représentation :

```
le 1
\tal 2
est 2
un 2
sous-domaine 2
de 1
l'2
apprentissage 2
automatique 2
mais 1
n'1
pas 1
du 1
```

Nous pouvons observer ci-dessus une perte de connaissance entre le texte et sa représentation : l'ordre des mots est inconnu, de même que les liens qu'ils entretiennent – grammaticaux comme syntaxiques. Autrement dit, bien que le vecteur ci-dessus représente le texte d'origine, il ne permet en aucun cas de le reconstituer. Bien que triviales, ces représentations se sont avérées efficaces pour divers procédés du TAL. Elles posent cependant certaines questions que nous avons préalablement évoquées :

- La segmentation en mots : la segmentation à partir des espaces est efficace mais ne gère pas tous les cas : *n'est* peut être considéré comme un seul mot ou comme deux. La remarque est valable pour toutes les contractions ainsi que les mots composés : *l'apprentissage*, *sous-domaine*.
- La lemmatisation : la décision de segmenter *n'est* ou *l'apprentissage* en deux mots pose la question de la forme du mot contracté ; c'est généralement sa forme canonique qui est conservée – respectivement *ne* et *le*.
- Une combinaison de segmentation et de lemmatisation : le cas *du*. Il faut choisir entre conserver *du* sous sa forme d'origine – dans le texte – ou comme la somme de *de+le*. La question se pose également pour *au(x)* en français, *cannot* en anglais.

Ce modèle de représentation est largement utilisé car efficace, facile à construire et peu coûteux en termes de ressources informatiques : la taille maximale d'un vecteur est la taille du vocabulaire, ce qui n'est pas le cas pour d'autres modèles proposés.

Un second modèle largement reconnu et utilisé est le modèle dit de n-grammes. Un n-gramme se définit simplement comme une séquence de n mots, où le sens de *mot* est défini à partir de l'étape de segmentation. Ainsi, au lieu de faire correspondre une valeur du vecteur à un élément unique du vocabulaire, le modèle en n-grammes associe des séquences de 1 à n mots à chaque valeur. Reprenons notre exemple ci-dessus pour une représentation sous formes de bigrammes ($n = 2$) – les unigrammes ($n = 1$) ont déjà été extraits, ce sont simplement les mots.

Exemple :
Texte : *Le* TAL *est un sous-domaine de l'apprentissage automatique mais l'apprentissage automatique n'est pas un sous-domaine du* TAL
Représentation :

```
le \tal 1
\tal est 1
estimation un 1
un sous-domaine 2
de l'1
l'apprentissage 2
apprentissage automatique 2
automatique mais 1
mais l'1
automatique n'1
n'est 1
est pas 1
pas un 1
sous-domaine du 2
du \tal 1
```

Nous nous limitons à présenter ici les unigrammes et bigrammes à des fins de concision. Nous pouvons observer deux phénomènes à partir des bigrammes extraits : i) un enrichissement excessif de l'espace de représentation à partir de bigrammes non pertinents, ii) la reconnaissance d'une unité linguistique pertinente ignorée avec les unigrammes : *apprentissage automatique*. Ces deux phénomènes sont caractéristiques des modèles en n-grammes :

ils permettent de reconnaître des unités linguistiques sémantiques de manière triviale – sans analyse linguistique – mais aux dépens de la taille du modèle.

La problématique de la taille du modèle construit nous mène ainsi aux considérations matérielles et techniques en lien avec le TAL.

1.2.3 Considérations techniques

Nous pouvons identifier deux aspects principaux qui ont trait au contexte technique du TAL et de l'apprentissage en général : i) des données en quantité limitée, et ii) une puissance de calcul également limitée. Le manque de données pour l'apprentissage est une problématique récurrente qui aboutit régulièrement à la création ad-hoc de jeux de données propres à une expérience. C'est par exemple le cas pour la construction automatique de terminologies où l'évaluation peut passer par une campage d'annotation de corpus plutôt que par la réutilisation d'une ressource pré-existante. La création de jeux de données spécifiques pose la question de l'évaluation de l'expérience, de sa reproductibilité et de sa représentativité – nous verrons comment nous y avons été confrontés dans les sections qui suivent.

La complexité des algorithmes et la limite de la puissance de calcul des ordinateurs induisent également une limitation de la taille des espaces de représentation. Nous ne pouvons pas accroître la taille d'un modèle sans considération pour les limites pragmatiques qui nous sont imposées : une augmentation de la taille du modèle induit une augmentation du nombre de calculs, donc un traitement plus long. L'augmentation du nombre de calculs relativement à l'espace de représentation est également en lien avec la complexité de l'algorithme choisi : pour un espace de représentation de même taille, les algorithmes nécessitent des laps de temps différents en fonction des tâches réalisées. Il faut donc équilibrer le triptyque taille de l'espace de représentation, complexité de l'algorithme et coût en calculs afin d'obtenir un résultat dans un laps de temps raisonnable.

1.3 Définitions et limites de l'extraction terminologique automatique

Nous avons présenté la variété des tâches du TAL ainsi que les différents domaines scientifiques sollicités. Nous introduisons maintenant le cœur de

nos recherches : l'extraction terminographique automatique et ses applications. L'extraction terminographique automatique est une tâche spécifique du TAL, complexe, qui s'appuie sur de nombreuses analyses que nous avons évoquées. L'objectif d'une extraction automatique de termes est pluriel : elle peut assister la construction d'une ressource terminologique, améliorer les moteurs de recherche, enrichir des bases de connaissances, etc.

Dans un premier temps, nous proposerons une définition préliminaire du concept de *terminologie* et plus précisément de *l'aspect terminologique*. Suivra une présentation généraliste de l'automatisation du processus d'extraction terminographique qui en distinguera les étapes essentielles. Dans un dernier temps, nous évoquerons succinctement – avant d'y revenir en détail plus loin – les différents biais propres à la terminographie automatique.

1.3.1 Conceptualisation de l'aspect terminologique

L'extraction terminographique automatique vise à extraire des séquences de un ou plusieurs mots qui correspondent à des concepts terminologiques. Un concept terminologique peut se définir comme une dénomination spécialisée propre à une communauté : *carte graphique* peut être considéré comme terminologique car propre à la communauté informatique. L'aspect terminologique d'une séquence de mots se définit donc par un contraste d'usage parmi plusieurs communautés : *carte graphique* n'est que peu utilisé dans le *langage courant*, *ciel bleu* est a priori également employé quelle que soit la communauté.

Les définitions et limites de ce qui constitue un concept terminologique ne sont pas consensuelles. Nous avons évoqué plus haut les biais psychologiques et sociaux, notamment dans le cadre d'une campagne d'annotation. Ceci est particulièrement vrai pour l'annotation terminologique. Malgré les désaccords entre chercheurs, l'augmentation de la production de données textuelles techniques induit la nécessité d'une accélération de l'analyse et donc d'automatisation, au moins partielle.

1.3.2 Automatisation du processus d'extraction

Le processus d'automatisation d'une extraction terminographique s'appuie sur une chaîne de traitements de TAL qui aboutit à la production d'un lexique potentiellement organisé – i.e. qui décrit les relations entre les termes.

Si certains traitements sont largement acceptés et appliqués dans la littérature, le consensus reste marginal sur la manière de procéder dans son ensemble. Qu'il s'agisse de l'extraction de termes, de leur validation ou de l'évaluation du lexique construit, aucune solution universelle n'a été mise en évidence.

Nous nous proposons dans nos recherches de définir un cadre formel précis et argumenté pour l'extraction terminographique. Après avoir présenté brièvement les problématiques posées par cette extraction dans ce chapitre, nous présenterons un état de l'art en chapitre 2. Du fait de la transversalité disciplinaire du TAL, nous ne pouvons produire un état de l'art exhaustif, relatif à tous les aspects scientifiques affectés par l'extraction terminographique ; nous présenterons donc des recherches directement en lien avec la terminologie et sa potentialité pluridisciplinaire. Nous nous concentrerons particulièrement sur les méthodes d'extraction et d'organisation de termes ainsi que sur l'évaluation de processus entièrement automatisés.

Pour plus de clarté, nous avons fait le choix de distinguer clairement ce qui relève de nos hypothèses et postulats de ce qui relève de nos expériences à proprement parler. Nous présentons en chapitre 3 les hypothèses qui ont sous-tendu les expériences présentées en chapitre 4. Nous verrons comment la problématique de la polysémie a motivé nos hybridations entre terminologie et modèles de thèmes en section 3.1, comment la structure morphosyntaxique des graphies terminologiques nous a semblé incomplète en section 3.2 et enfin comment le lexique construit nécessite un protocole d'évaluation clair, précis et objectif en section 3.3.4.

De manière logique, nous retrouvons en chapitre 4 des sections qui correspondent à celles présentées en chapitre 3. Nous y présentons diverses expériences qui ont trait à l'hybridation entre terminologie et modèle de thèmes en section 4.1 ainsi qu'une expérience en lien avec une modification de la structure morphosyntaxique des graphies terminologiques en section 4.2.

Suivra une conclusion en chapitre 5 qui évoque les résultats de nos expériences, les perspectives de recherche ainsi que les divers biais et obstacles que nous avons pu rencontrer lors de nos expérimentations.

Chapitre 2

Extraction terminographique automatique

L'extraction terminographique repose sur différents aspects du TAL et de la science des données. Nous présenterons ici les recherches effectuées (in)directement en lien avec la problématique de la construction automatique de terminologies. Comme nous avons commencé à l'évoquer en partie 1.2, l'extraction terminographique est une tâche complexe du TAL en ce qu'elle s'appuie sur de nombreuses autres analyses automatiques, prônes à l'erreur – étiquetage morphosyntaxique, désambiguisation, lemmatisation, etc.

Nous poserons plusieurs définitions et limites de concepts avant de présenter les différentes recherches en lien avec notre problématique (partie 2.1.1). Les concepts définis sont basiques mais la précision de leur définition est nécessaire à la bonne compréhension du procédé d'extraction ; les ambiguités/imprécisions derrière des mots comme *mot*, *terme* ou *syntagme* sont à résoudre. Discutables, ces définitions n'ont pas pour objectif de répondre à des questionnements scientifiques mais à fixer le vocabulaire et les concepts pour la suite ainsi qu'à rendre la synthèse cohérente. A ces définitions préliminaires suivront les prémisses d'une formalisation mathématique (partie 2.1.2) qui sera employée par la suite, notamment pour la présentation des méthodes d'extraction terminographique automatique ainsi que pour détailler nos expériences.

Après avoir fixé le vocabulaire nécessaire, nous introduirons différentes tâches du TAL en lien avec l'extraction terminographique automatique (partie 2.2). Nous avons fait le choix de présenter ces tâches indépendamment de la problématique principale car ce sont pour l'essentiel des tâches optionnelles

destinées à affiner la terminologie construite. L'utilisation de ces procédés du TAL est donc laissée au bon jugement du chercheur/terminologue, en fonction de la qualité attendue et du temps disponible. Ces tâches n'en demeurent pas moins des éléments constitutifs de la chaîne de traitements de construction d'une terminologie idéale et leur présentation est donc nécessaire.

2.1 Concepts de base et formalisation

2.1.1 Définitions et limites de concepts

Nous donnons ici les définitions essentielles à la compréhension de la suite de nos recherches. Les concepts peuvent sembler triviaux, néanmoins leur délimitation est nécessaire à la formalisation – donc à la compréhension – des méthodes d'extraction terminographique automatique.

L'unité de base dans l'analyse de texte est généralement le *mot* ou *token* en anglais, qu'il faut distinguer de la lexie. Alors que le concept de *mot* correspond au stade initial de l'analyse, celui de *lexie* intervient plus tard.

Definition 2.1.1.1. Mot – Désigne une suite finie maximale de caractères non blancs à une position donnée. Dans la phrase « une$_1$ terminologie$_2$ est$_3$ une$_4$ ressource$_5$ linguistique$_6$ » les mots 1 et 4 sont différents. Le mot correspond à la notion de *token* en anglais. Un mot peut également se définir comme une *instance*.

Definition 2.1.1.2. Forme de surface – Nom donné à la seule forme visuelle d'un mot ou d'une séquence de mots – la graphie – sans autre connaissance (lemme, étiquette morphosyntaxique, sémantique, etc.). Dans la phrase « Des$_1$ vers$_2$ creusent$_3$ vers$_4$ les$_5$ racines$_6$ » les mots 2 et 4 sont deux mots différents mais avec la même forme de surface. Leurs lemmes/lexèmes respectifs sont également différents (*ver* vs. *vers*). Nous emploierons *forme de surface* et *graphie* de manière indifférenciée.

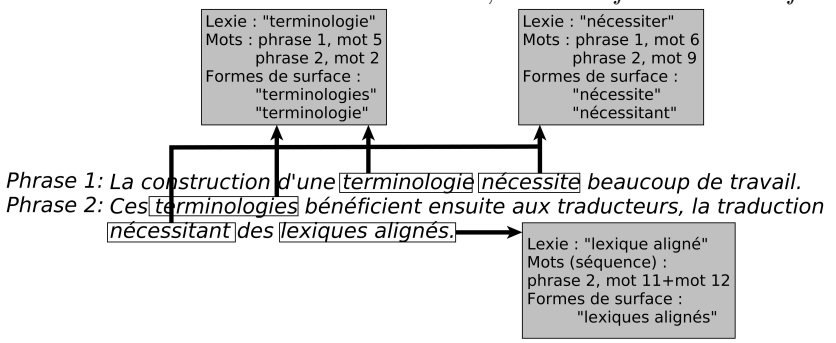
Definition 2.1.1.3. Lexie – Elément du lexique potentiellement normalisé – lemmatisation, racinisation, etc. – auquel peut correspondre un ou plusieurs mots. Dans la phrase « Des$_1$ vers$_2$ creusent$_3$ vers$_4$ les$_5$ racines$_6$ » les mots 2 et 4 ont la même forme de surface mais correspondent à des lexies différentes. Dans la phrase « une$_1$ terminologie$_2$ est$_3$ une$_4$ ressource$_5$ linguistique$_6$ », les mots 1 et 4 correspondent à la même lexie : l'article indéfini *un*. Une lexie

peut se définir sur une séquence de mots, dans l'exemple ci-avant la séquence $ressource_5$ $linguistique_6$ correspond à une lexie.

Une lexie peut être commune à plusieurs mots, un mot étant défini par sa position et sa forme de surface.

D'après ces définitions, nous pouvons voir que le premier critère de regroupement de plusieurs mots sous une même lexie est basé sur la seule graphie. Suivent ensuite les procédés du TAL qui vont permettre de prendre en compte différents phénomènes linguistiques et d'améliorer le regroupement sous différentes lexies. La figure 2.1 illustre la distinction que nous posons ici entre ces trois concepts fondamentaux dans nos recherches.

FIGURE 2.1 – Distinction entre *mot*, *lexie* et *forme de surface*



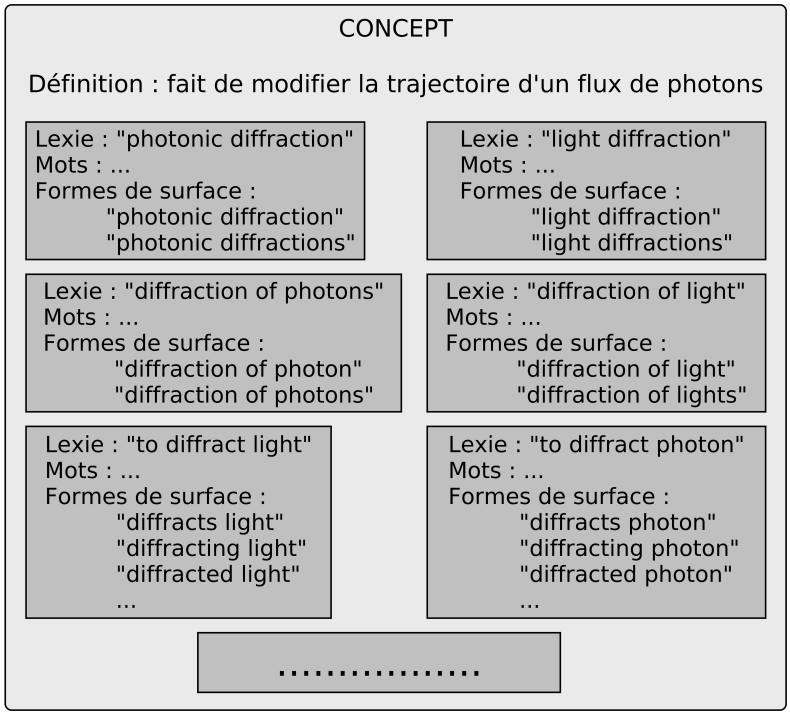
Nous pouvons voir dans la figure 2.1 que la forme canonique d'une lexie n'apparaît pas nécessairement dans les documents analysés : le verbe *nécessiter* n'apparaît pas à l'infinitif dans le texte. De la même manière, *lexique aligné* n'apparaît pas au singulier. Ces définitions situent la construction/reconnaissance de lexies juste avant la construction/reconnaissance de concepts en termes d'abstraction de l'analyse. Le *mot* peut quant à lui être défini comme une occurrence de forme de surface à une position particulière dans le texte.

Definition 2.1.1.4. Concept – Unité sémantique abstraite manifestée au travers d'une ou plusieurs lexies. Un concept fait généralement partie d'un réseau sémantique permettant d'effectuer des raisonnements.

La figure 2.2 illustre la notion de concept relativement à celle de lexie. Le passage au niveau conceptuel nécessite des procédés du TAL plus complexes que pour la construction/reconnaissance de lexies, des procédés comme :

— la reconnaissance de la paraphrase : *photonic diffraction* vs *diffraction of photons*,
— la reconnaissance de synonymes : *light* vs *photon*,
— la dérivation, i.e. le changement de catégorie grammaticale : groupe verbal *to diffract light* vs groupe nominal *light diffraction*.

FIGURE 2.2 – Distinction entre concept et lexie



Nous ne prenons pas en compte le niveau conceptuel dans nos recherches mais uniquement le niveau des lexies, qui correpond au niveau d'analyse de la construction automatique de terminologie – le niveau conceptuel est trop détaché de la graphie. Aux côtés de cette hiérarchie de définitions coexistent des définitions satellites qui peuvent prêter à confusion dans un contexte d'analyse terminographique.

Definition 2.1.1.5. Lemme – Forme canonique – i.e. non fléchie – d'un mot donné issue du processus de lemmatisation. La forme canonique corres-

pond à la forme attendue dans un dictionnaire. Le lemme de *est* dans « le TAL c'est bien » est *être*.

Le lemme semble correspondre à la forme canonique d'une lexie dans nos exemples, ce n'est cependant pas nécessairement le cas : la graphie *trained algorithms* correspond à une lexie dont la forme canonique est *trained algorithm*. Dans cet exemple, seul le nom commun final *algorithm* a été lemmatisé, la lemmatisation du verbe entraînerait une confusion entre l'objet – l'algorithme qui a appris – et l'action – apprendre à un algorithme. Le concept de syntagme peut également prêter à confusion.

Definition 2.1.1.6. Syntagme – Suite finie de mots, constituant syntaxique de la phrase. Dans la phrase « une terminologie est une ressource linguistique », *une terminologie* et *une ressource linguistique* sont des syntagmes nominaux.

La notion de syntagme semble proche de celle de séquence de mots, elle est néanmoins davantage contrainte : un syntagme doit satisfaire un ensemble de règles syntaxiques et correspond à une unité sémantique. L'extraction de séquences de mots est possible à partir d'une simple segmentation en mots ou affinée à partir de procédés du TAL comme l'étiquetage morphosyntaxique. Inversement, l'extraction de syntagmes passe par une analyse en constituants immédiats (ACI) – hiérarchisation des groupes de mots constitutifs d'une phrase donnée – qui est un traitement lourd et relativement hasardeux mais qui permet de reconnaître des structures plus complexes.

A ces définitions s'ajoutent celles qui relèvent strictement du contexte de l'extraction terminographique. Les concepts de *terme candidat*, *terme* et *potentiel terminologique* en sont les concepts principaux :

Definition 2.1.1.7. Potentiel terminologique – Score attribué à une lexie permettant la distinction entre termes et non-termes.

Definition 2.1.1.8. Terme candidat – Lexie dont le potentiel terminologique a été estimé et qui doit être validée. La lexie est *candidate* à l'insertion dans une terminologie.

Definition 2.1.1.9. Terme – Elément d'une terminologie, terme candidat validé. La validation peut être automatique au travers d'un algorithme, ou manuelle par un terminologue.

La figure 2.3 page 65 synthétise l'ensemble de ces définitions dans un exemple simple d'extraction automatique de termes.

2.1.2 Formalisation mathématique

Nous avons vu qu'un texte est traité de manière automatique avec divers outils de traitement automatique de la langue : correcteur ortographique, analyse morphosyntaxique, reconnaissance de motifs (expressions régulières), etc., avant de réaliser des traitements spécifiques. Nous donnons ici un exemple d'execution d'étiquetage morphosyntaxique.

Exemple : *La phrase* "A_1 recursive_2 neural_3 network_4 is_5 a_6 kind_7 of_8 deep_9 neural_{10} $\text{network}_{11\cdot12}$"
est transformée par TreeTagger en une succession de triplets consitutés du mot, de son étiquette et de son lemme :

```
1 A DT a
2 recursive JJ recursive
3 neural JJ neural
4 network NN network
5 is VBZ be
6 a DT a
7 kind NN kind
8 of IN of
9 deep JJ deep
10 neural JJ neural
x network NN network
12 . SENT .
```

2.1.2.1 Document et item

Afin de tenir compte des résultats de tels outils, principalement de l'analyse morphosyntaxique, nous pouvons modéliser un document d par une succession d'items, un item encapsule la graphie d'un mot et les connaissances qui lui sont rattachées.

Exemple : *Si le document d est composé de l'unique phrase de l'exemple précédent et si nous structurons l'item avec graphie, étiquette et lemme, le document t est formé en 12 items :*

```
[ (A, DT, a), (recursive, JJ, recursive), (neural, JJ, neural),
(network, NN, network), (is, VBZ, be), (a, DT, a), (kind, NN, kind),
(of, IN, of), (deep, JJ, deep), (neural, JJ, neural),
(network, NN, network), (., SENT, .) ]
```

Notons $d(i)$ l'item qui se trouve à la position i dans le document d.
Pour un item it, notons :

— $\sigma(it)$ sa graphie
— $\pi(it)$ son étiquette morphosyntaxique
— $\lambda(it)$ son lemme

A tout item peut également être rattaché tout autre élément issu des traitements automatiques : $\gamma(it)$ pour l'orthographe correcte du mot, $\delta(it)$ pour une entrée dans un dictionnaire, etc. Notre travail porte essentiellement sur l'extraction terminologique à partir des textes scientifiques et techniques. Les textes analysés ont un niveau de langue suffisant, nous allons considérer que pour tout item it, garder uniquement $(\sigma(it), \pi(it), \lambda(it))$ est suffisant.
Exemple : pour le court document précédent $\sigma(d(1)) = A$, $\pi(d(3)) = JJ$, $\lambda(d(1)) = a$.

Certains traitements sur texte prennent en compte le nombre de lettres qui composent une graphie. Nous appelons cette valeur **taille** de la graphie et la notons $length(m)$.

Exemples : $length(\texttt{neurone}) = 7$, $length(\texttt{et}) = 2$
et $length(\texttt{de}) = 2$

Deux items sont égaux $it_1 = it_2$, s'il y a égalité entre chaque composant. Nous pouvons introduire aussi une relation d'équivalence $\equiv$ entre les items :

$$item_1 \equiv item_2$$

Cette relation d'équivalence se définira selon le contexte du traitement final visé, de la nature des documents et de la structure des items. Si le texte traité contient des fautes d'orthographe et/ou de l'argot, l'équivalence se définit sur la base de l'égalité des orthographes correctes $\gamma(it_1) = \gamma(it_2)$. Dans le cadre de notre travail sur l'extraction terminographique, l'égalité revient à imposer

que pour deux items la graphie, l'étiquette et le lemme soient les mêmes ; la relation d'équivalence, plus relâchée, impose que l'étiquette et le lemme soient les mêmes.

Exemple : $d(3) = d(10)$ et, si l'égalité stricte des formes surfaciques n'est pas prise en compte, $d(1) \equiv d(6)$.

Pour un document d nous allons noter $|d|$ ou $length(d)$ le nombre d'items qui le composent.
Dans notre cas d'exemple $length(d) = 12$.

2.1.2.2 Termes candidats

Un terme candidat, notons le ct, est aussi composé d'une succession d'items. Pour le moment nous ne faisons aucune hypothèse sur la manière employée pour le construire et/ou le valider.

Exemple : ct_1 et ct_2 sont des termes candidats.

$$ct_1 = [(recursive, JJ, recursive), (neural, JJ, neural), (network, NN, network)]$$

$$ct_2 = [(neural, JJ, neural), (networks, NN, network)]$$

Nous gardons la notation $ct(i)$ pour retrouver l'item qui se trouve à la position i dans ct et la notation $|ct|$ ou $length(ct)$ pour indiquer le nombre d'items qui le composent.

Exemple : $ct_1(1) = (recursive, JJ, recursive)$, $ct_2(1) = (neural, JJ, neural)$, $|ct_1| = 3$ et $|ct_2| = 2$

Nous appellerons inclusion entre termes candidats :

$$ct_1 \prec ct_2$$

Par rapport à la relation d'égalité entre items ou par rapport à une relation d'équivalence, une relation qui indique que ct_1 est d'une taille plus petite que

ct_2 et que tous les items qui composent ct_1 se retrouvent dans ct_2 sur des positions successives.

$$length(ct_1) \leq length(ct_2)$$

$$\exists i_0 \text{ tel que } \forall i = 1, length(ct_1) \ : ct_2(i + i_0 - 1) \equiv ct_1(i)$$

(ct_1 est inclus dans ct_2 à partir de la position i_0). Telle qu'elle est définie cette relation binaire est une relation d'ordre partiel.

Exemple : Considérons les termes candidats suivants :

$$ct_1 = [(recursive, JJ, recursive), (neural, JJ, neural), (network, NN, network)]$$

$$ct_2 = [(neural, JJ, neural), (networks, NN, network)]$$
$$ct_3 = [(recursive, JJ, recursive), (network, NN, network)]$$
$$ct_4 = [(network, NN, network)]$$

Nous avons uniquement les inclusions suivantes :

$$ct_2 \prec ct_1$$

$$ct_4 \prec ct_3,\ ct_4 \prec ct_2 \text{ et } ct_2 \prec ct_1$$

Nous pouvons remarquer que la relation $ct_3 \prec ct_1$ est fausse car les items composants ct_3 ne se retrouvent pas sur des positions successives dans ct_1.

Sur un ensemble de termes candidats $\mathcal{CT}$ nous introduisons l'ensemble de termes candidats qui incluent un terme ct, ou l'ensemble de supra-termes :

$$Supra(ct) = \{x \in \mathcal{CT} | ct \prec x\}$$

Exemple : pour $\mathcal{CT} = \{ct_1, ct_2, ct_3, ct_4\}$*, nous avons* $Supra(ct_4) = \{ct_1, ct_2, ct_3\}$*,* $Supra(ct_2) = \{ct_1\}$ *et* $Supra(ct_1) = Supra(ct_3) = \emptyset$*.*

La forme surfacique d'un terme candidat ct, notée $\sigma(ct)$ est obtenue par concaténation des formes surfaciques des items qui le composent avec des espaces.

Pour ct_1 *de l'exemple précédent,* $\sigma(ct_1) =$ *"recursive neural network".*

2.1.2.3 Fréquences des items et des termes candidats

La fréquence d'un item it dans un document d, notée $F(it, d)$, est le nombre d'items équivalents à l'item it contenus dans d. A savoir :

$$F(it, d) = card(\{j | d(j) \equiv it\})$$

Exemples : $F((neural, JJ, neural), d) = 2$, $F((network, NN, network), d) = 2$ *et* $F((network, VBZ, network), d) = 0$

Ce concept correspond à la notion de fréquence absolue d'un item dans un document, autrement dit le nombre d'occurences d'items équivalents dans le document.

Pour ce qui est d'un terme candidat ct, la fréquence est définie comme le nombre d'occurrences de ce terme candidat dans le document :

$$F(ct, d) = card(\{j | \forall i = 1, length(ct)\ d(j + i - 1) = ct(j)\})$$

Si pour un item il est possible de parler de fréquence relative comme rapport entre la fréquence absolue et la longueur du document, ce concept n'est pas applicable aux termes candidats.

Pour un corpus $\mathcal{D}$ de documents, nous introduisons la notion de fréquence d'un item ou d'un terme candidat x :

$$F(x, \mathcal{D}) = \sum_{d \in \mathcal{D}} F(x, d)$$

2.2 Problématiques linguistiques liées à l'extraction terminologique automatique

La construction de terminologie est une tâche complexe en linguistique, encore davantage lorsqu'il s'agit d'automatiser le processus. Tous les niveaux linguistiques sont sollicités, de la morphologie à la pragmatique. Nous en donnons ici une présentation succincte.

Niveau morphologique La délimitation en morphèmes permet d'isoler les morphèmes grammaticaux flexionnels et dérivationnels, et donc de regrouper différentes formes fléchies sous un même terme. Ces morphèmes correspondent à deux phénomènes linguistiques introduisant

de la variation : la variation grammaticale (conjugaison) et la dérivation (changement du sens). La délimitation en morphèmes peut également permettre de repérer certains morphèmes récurrents significatifs pour la détection automatique de termes. Des morphèmes comme *-logie* (cardiologie, podologie, etc.) ou *-ectomie* (lobectomie, splénectomie, etc.) ont un caractère discriminant dans la détection automatique de termes.

Niveau syntaxique L'analyse morphologique ayant identifié les différents morphèmes d'une phrase, l'analyse syntaxique va permettre d'analyser les relations entretenues entre ces derniers. Le niveau syntaxique sera notamment sollicité pour la gestion des paraphrases et des chaînes anaphoriques (voir partie 2.2.4).

Niveau sémantique La sémantique est le cœur de la construction automatique de terminologie. Comme précédemment évoqué, ce processus passe par un regroupement des différentes formes de surfaces sous un même terme, abstrait, qui fait référence au sens de ce concept. Selon la traditionnelle dichotomie Saussurienne signifiant/signifié, la ressource terminologique les distingue : le signifiant manifesté par les formes de surface, le signifié par la réunion de différents signifiants même si le référent n'est pas explicitement mentionné (définition par extension).

Niveau pragmatique La prise en compte de la pragmatique est inhérente au concepts de *terminologie* et de *terme*. Un terme étant spécifique à une communauté et à un sujet, le contexte d'émission des documents qui contiennent une de ses formes surfaciques est capital tant pour la construction manuelle de terminologie que pour son automatisation. Cela pourra se manifester par une prise en compte humaine dans le cas d'une construction/validation manuelle (partie 2.4.4), ou par des techniques de linguistique de corpus dans le cas d'une automatisation (partie 2.3).

Si les niveaux morphologique et syntaxique sont aisés à appréhender, les niveaux sémantique et pragmatique le sont un peu moins. Considérons par exemple la phrase *Actuellement dans beaucoup de parkings la reconnaissance automatique de la plaque minéralogique est basée sur un réseau de neurones.* Un outil quelconque pourrait détecter *plaque minéralogique* et *réseau de neurones* comme des syntagmes pertinents, i.e. comme de possibles termes candidats. L'un, l'autre ou les deux candidats seront gardés selon la communauté

de travail et selon l'application donnée à l'extraction terminographique effectuée.

Nous présentons maintenant plus en détails différentes problématiques linguistiques spécifiques liées à la construction automatique de terminologie. Du fait de la complexité de la tâche, la liste ne se veut pas exhaustive ; nous noterons par exemple que le niveau d'analyse phonologique est ici complètement ignoré.

2.2.1 Ambiguité et polysémie - un mot, un contexte, un sens

La polysémie peut se définir comme l'existence de plusieurs sens – signifiés – possibles pour une forme de surface/lexie – signifiant – donnée. L'ambiguïté – symptôme de la polysémie – est l'un des phénomènes linguistiques rencontrés lors de la construction de terminologies, automatique ou non. L'ambiguïté désigne l'impossibilité d'attribuer un signifié parmi d'autres à un signifiant indépendamment de son contexte. Il s'agit par exemple d'observer que le terme manifesté par la forme de surface *neural network* a plusieurs sens selon qu'il apparaisse dans un contexte médical ou informatique. La figure 2.4 donne un exemple d'occurrence de cette problématique et illustre le fait que l'identification de termes ne s'appuie pas que sur la morphologie, mais également sur la pragmatique. Dans cet exemple, l'identification des sens distincts de *neural network* passe par une interprétation qui nous amène à considérer les phrases 1 et 2 de la figure 2.4 comme des phrases portant sur des sujets différents. L'exemple ne porte que sur une phrase, mais les mêmes déductions sont possibles sur des unités de longueurs variables – propositions, paragraphes, documents, etc.

La désambiguisation automatique étant en elle-même une tâche complexe du TAL, elle est liée à d'autres problématiques linguistiques comme la néologie et la synchronie/diachronie. Comme nous le verrons dans la partie 2.2.2.1, une des manifestations de la néologie est l'apparition de nouveaux sens – *signifiés* – pour des mots – *signifiants* – pré-existants. Autrement dit, une des formes de la néologie est la création régulière d'ambiguïtés. L'apparition de l'informatique a par exemple été un vecteur important de néologismes, avec de nouveaux sens pour des mots comme *souris*, *navigateur*, *virus*, etc. Si le contexte d'extraction terminographique est statique, i.e. les ressources linguistiques sont fixes, l'analyse est dite *en synchronie*. Si au contraire le

contexte est dynamique, i.e. un flux de documents est analysé en direct, l'analyse est dite en diachronie. Le choix de l'un ou l'autre contexte impactera la gestion de la polysémie : il est possible d'observer l'émergence de nouvelles ambiguïtés en diachronie, pas en synchronie.

La désambiguisation désigne le processus visant à déterminer le(s) sens d'une lexie. Le processus de désambiguisation doit être réalisé en amont de la construction des termes candidats, la polysémie devant idéalement donner lieu à la création d'autant de termes candidats que de sens. L'exemple donné dans la figure 2.4 illustre ce phénomène. Quand elle est manuelle, la fenêtre d'analyse est la plus large ; le terminologue a connaissance de la nature des documents qu'il étudie et de la finalité de la terminologie qu'il construit. A contrario quand elle est automatique, le contexte linguistique est donc plus restreint : la pragmatique en est par exemple absente.

2.2.2 La néologie - apparition de nouveaux mots

La néologie est un phénomène linguistique pouvant généralement se définir comme l'apparition d'une relation entre un signifié et un signifiant. Comme évoqué en partie 2.2.1, la néologie est un problème inhérent à la terminologie qui peut se manifester sous différentes formes. Deux aspects principaux de la néologie sont à distinguer : néologie sémantique (partie 2.2.2.1) et néologie syntaxique/formelle (partie 2.2.2.2). Le premier peut se définir comme l'apparition d'un signifié pour un signifiant préexistant, le second comme l'apparition d'un signifiant pour un signifié potentiellement nouveau.

Dans une analyse en diachronie, à savoir la construction et l'enrichissement automatiques d'une terminologie, la néologie sémantique se manifeste par l'apparition de nouvelles ambiguïtés à un moment donné $\mathcal{T}$ – nouvelles par rapport à l'ensemble des textes analysés avant ce moment $\mathcal{T}$. La néologie syntaxique se manifeste quant à elle par l'apparition de nouveaux signifiants – formes de surface – à un moment $\mathcal{T}$, qui ne sont pas rattachés à un terme préalablement construit à partir des textes analysés avant ce moment $\mathcal{T}$. En synchronie, à savoir la construction automatique d'une terminologie à partir d'un instantané de la langue à un moment $\mathcal{T}$, la notion de *nouveauté* ne peut être perçue que par le biais de ressources externes (par exemple des bases de connaissances) et non plus par comparaison avec une analyse passée.

Nous développons ici plus en détails les liens entre néologie, chronologie et terminologie.

2.2.2.1 La néologie sémantique - un nouveau sens pour un mot pré-existant

La néologie sémantique est une sous-catégorie de la néologie qui se caractérise par l'apparition d'un nouveau sens pour un signifiant préexistant. La figure 2.5 illustre ce phénomène avec le signifiant *virus*, dont le sens a évolué avec l'apparition de l'informatique. La néologie sémantique introduit de l'ambiguïté pour les occurrences futures de la forme de surface *virus*, auparavant monosémique [1].

La problématique de la néologie sémantique relève du contexte d'exécution de l'analyse des documents. Comme évoqué dans la partie 2.2.2, la synchronie/diachronie ainsi que l'utilisation de ressources externes influenceront la manière de détecter les occurrences de ce phénomène linguistique. Théoriquement, une analyse de la néologie sémantique en diachronie – à un moment $\mathcal{T}$ – doit reposer sur des ressources elles-mêmes en diachronie, à savoir des ressources mises à jour selon l'état de la langue à ce moment $\mathcal{T}$. L'utilisation de bases de connaissances comme outils de repérage de la néologie est donc problématique en diachronie, aussi les techniques de détection s'appuient-elles davantage sur une comparaison avec les données traitées avant le moment $\mathcal{T}$. Les figures 2.6 et 2.7 (pages 67 et 67) illustrent la problématique posée par l'alignement entre ressources et contexte d'analyse.

Dans la figure 2.6, il est à noter que le nouveau sens de *virus* n'est pas détecté du fait de la base de connaissances externes qui est statique. Une possibilité serait de mettre à jour ponctuellement la base de connaissances ; cela ne résoudrait cependant pas cette problématique : si une mise à jour a lieu à $\mathcal{T}$ et à $\mathcal{T}+2$, la base de connaissances en $\mathcal{T}+1$ sera lacunaire. La seule solution pour répondre à cette problématique dans un contexte de diachronie consiste à ancrer la base de connaissances dans *l'instant courant*, autrement dit à la rendre diachronique. La figure 2.7 illustre cet ancrage de la base de connaissances dans l'instant courant.

Pour une analyse en synchronie, par définition, l'utilisation de ressources externes ne pose pas de problème d'alignement avec un instant. L'analyse visant à détecter un néologisme sémantique à un instant unique $\mathcal{T}$, il est cependant nécessaire de minimiser l'intervalle effectif entre le moment de création de la ressource et le moment dont les documents sont issus – minimiser voire exploiter une ressource postérieure au moment des documents, avec les conséquences qui s'ensuivent.

1. `https://www.cnrtl.fr/definition/virus`

2.2.2.2 La néologie syntaxique - nouveau mot, nouvelle structure phrastique

La néologie syntaxique est à considérer relativement à la néologie sémantique. Avec la terminologie Saussurienne, la néologie syntaxique peut se définir comme l'apparition d'un nouveau signifiant pour un (nouveau) signifié. De fait, la néologie syntaxique s'appuie sur la morphologie pour la création de formes non attestées par des procédés linguistiques comme les métaplasmes ou l'affixation.

Definition 2.2.2.1. Métaplasme – Ensemble des phénomènes linguistiques qui produisent des formes non attestées à partir d'altérations des séquences de phonèmes. Amphithéâtre → Amphi (apocope)
Américain → ricain (aphérèse)
Monsieur → Msieur (syncope)
etc.

Definition 2.2.2.2. Affixation – Phénomène linguistique morphologique consistant à ajouter/supprimer des affixes à une lexie donnée afin d'en modifier la signification ou la catégorie grammaticale.
Conformiste → Anticonformiste (préfixe)
Conforme → Conformément (suffixe)

Un cas particulier de néologie syntaxique est à distinguer, qui ne repose pas sur les règles de la langue courante : l'emprunt, à savoir l'usage dans une langue donnée d'une lexie d'une autre langue.

2.2.3 Diachronie et synchronie : quel choix pour quelle solution

La langue naturelle est en perpétuelle évolution : des mots apparaissent, d'autres disparaissent, les sens changent, des mutations phonétiques peuvent apparaître, etc. Autant d'événements dont la détection implique la prise en compte de la temporalité, à savoir l'adoption de l'axe diachronique. Analyser le langage naturel en diachronie dans un contexte d'extraction terminographique signifie se concentrer sur les différences de langue entre deux moments donnés, différences issues de phénomènes linguistiques liés à l'extraction terminographique. Comme évoqué dans la partie 2.2.2, l'apparition

de nouveaux mots ou de nouveaux sens pour un mot préexistant – néologie – est un exemple typique de la nécessité de la diachronie : la notion de *nouveauté* est estimée relativement à des connaissances pré-établies, ce qui induit la prise en compte d'au moins deux moments distincts sur l'axe temporel. Dans un contexte de ressource terminologique enrichie dynamiquement, la maintenance de la ressource nécessite d'identifier deux phénomènes terminologiques : i) la *terminologisation* [Calberg-Challot, 2007] d'un mot à partir d'un néologisme sémantique, par exemple la terminologisation du mot *souris* aux débuts de l'informatique ainsi que sa réciproque ii) la *déterminologisation* [Meyer and Mackintosh, 2000] d'un terme à partir d'un glissement de son sens vers le grand public. Il est notable qu'il y a une continuité entre termes et non-termes, de la même manière qu'elle existe entre néologisme et vocabulaire attesté : le sens de *souris* tend à se déterminologiser avec la diffusion de l'informatique, mais l'identification de l'instant précis à partir duquel *souris* n'est plus du tout terminologique dépend de l'objectif du chercheur qui veut modéliser le phénomène. A l'inverse de ces analyses, dans la plupart des cas, c'est l'axe synchronique qui est adopté. L'axe synchronique se définit comme une analyse sur un instantané de la langue à un moment donné ; c'est l'axe adopté tacitement dans la majorité des outils de TAL : la construction d'un modèle de langue statique qui va servir à étiqueter, lemmatiser, raciniser ou quelqu'autre tâche de TAL induit l'axe synchronique dans la mesure où l'évolutivité de la langue est ignorée. Exceptés quelques travaux notamment de veille néologique [Sablayrolles, 2010, Cartier, 2017], la complexité de la diachronie pousse les chercheurs à l'approximer en la discrétisant – seuls quelques points sont sélectionnés sur l'axe temporel au lieu de prendre tout l'axe en considération. C'est par exemple le cas des techniques d'analyse contrastive pour la détection de néologismes où deux corpora distants d'une période donnée permettent l'identification du phénomène. Dans un contexte d'extraction terminographique, malgré l'intérêt de l'axe diachronique, la complexité inhérente à la tâche fait que l'ensemble des travaux réalisés dans ce domaine l'ont été en synchronie.

2.2.4 Paraphrases et variantes surfaciques d'un même concept

La variabilité de la langue accompagnée d'une tendance à éviter les répétitions induit l'existence de nombreuses variantes graphiques pour un même

terme candidat. Sont réunis dans cette partie l'ensemble des phénomènes linguistiques générateurs de variations dans les graphies des termes. Malgré les observations de [Justeson and Katz, 1995] qui indiquent que les termes sont des séquences de mots fortement lexicalisées – à savoir justement peu sujets à la variation – de nombreux phénomènes occurrent qui peuvent nuire aux performances d'outils d'extraction terminographique, notamment via l'augmentation indue de la taille du vocabulaire [Park et al., 2002].

Variantes symboliques : *micro-processeur, microprocesseur*, apparition/disparition/remplac d'un séparateur.

Variantes de compositions : *microprocesseur, micro processeur*, remise en cause de la segmentation du terme.

Variantes flexionnelles : *microprocesseur, microprocesseurs*, variations en genre et en nombre, aisément corrigées avec un lemmatiseur.

Erreurs typographiques : *microprocesseur, mcroprocesseur*, les erreurs typographiques sont très fréquentes et généralement repérées grâce à la distance d'édition (mesure de similarité entre chaînes de caractères).

Acronymes : *Traitement Automatisé du Langage,* TAL, les acronymes sont très fréquents dans les documents techniques et les articles scientifiques ; des solutions à base de patrons ont été proposées pour les repérer, aboutissant à une précision et un rappel supérieurs à 93% [Park and Byrd, 2001].

Paraphrases : *diffraction photonique, diffraction de photons*, la paraphrase est un phénomène linguistique complexe pour lequel les techniques de plongements lexicaux (*word embeddings*, voir [Mikolov et al., 2013a, Mikolov et al., 2013b, Mikolov et al., 2013c]) sont souvent utilisées.

Dérivation : *photonique, photon*, changement de catégorie grammaticale, généralement par adjonction de morphèmes dérivationnels ; le repérage de la dérivation passe généralement par la racinisation.

2.2.5 Les chaînes de coréférences

Les chaînes de coréférences sont un phénomène linguistique qui permet d'éviter les répétitions, grâce par exemple aux pronoms relatifs :

Exemple 1 : *J'ai entraîné* ***un réseau de neurones. Il*** *peut reconnaître les graphies terminologiques.*

Dans l'exemple 1, les deux signifiants *un réseau de neurones* et *il* en gras font référence au même signifié. La reconnaissance des chaînes de coréférence – i.e. l'association entre un signifié et un ou plusieurs signifiants – est une tâche complexe du TAL en ce qu'elle peut solliciter les niveaux les plus abstraits de l'interprétation linguistique – sémantique, pragmatique. Elle permet cependant d'affiner les calculs lors de la reconnaissance de graphies terminologiques en prenant en compte toutes les occurrences de ces dernières. A partir de l'exemple 1, nous pouvons distinguer deux décomptes possibles :

Avec coréférences : $F(\text{réseaux de neurones}) = 2$

Sans coréférence : $F(\text{réseaux de neurones}) = 1$

Le choix d'opter pour un signifiant plutôt qu'un autre pour un même signifié dépend du choix de l'auteur du texte. N'analyser le document qu'au travers des signifiés permettrait donc en théorie de s'affranchir de l'influence du style.

- *J'ai entraîné* ***un réseau de neurones****.* ***Ce réseau de neurones*** *peut reconnaître les graphies terminologiques.*
- *J'ai entraîné* ***un réseau de neurones****.* ***Celui-ci*** *peut reconnaître les graphies terminologiques.*
- *J'ai entraîné* ***un réseau de neurones qui*** *peut reconnaître les graphies terminologiques.*
- *J'ai entraîné* ***un réseau de neurones****.* ***Ce réseau*** *peut reconnaître les graphies terminologiques.*

Ci-dessus trois occurrences du phénomène de coréférence à partir de l'exemple 1. De même que les paraphrases et les variantes surfaciques, nous regroupons les chaînes de coréférences sous une même problématique en extraction terminographique : la conflation de la variation. La conflation de la variation inclut l'ensemble des phénomènes linguistiques induisant des modifications dans le décompte d'une graphie donnée : flexion, dérivation, paraphrase, coréférence, fautes typographiques, etc. Nous illustrons cette notion plus en détail dans la partie 2.3.1.3.

2.3 Techniques d'extraction automatique de termes candidats

Bien que le consensus soit relativement rare sur la terminologie théorique, de nombreuses recherches ont été effectuées sur la construction automatique de terminologies, recherches qui ont abouti à diverses solutions situées entre

les statistiques et la linguistique. Les solutions proposées reposent généralement sur trois étapes principales :

1. L'extraction des termes candidats
2. L'attribution de leur potentiel terminologique
3. La validation des termes candidats

L'ensemble des étapes n'est pas systématiquement sollicité, mais elles sont suffisamment récurrentes pour justifier ce triptyque. Comme nous le verrons dans la partie 2.3.2, l'étape d'attribution du potentiel terminologique est parfois confondue avec la validation des termes candidats – par exemple lorsque le score est booléen. Les différentes solutions proposées par les chercheurs nous poussent à suivre cette partition de la tâche dans notre présentation et à développer les différentes méthodes des trois étapes : nous aborderons en premier lieu les différentes méthodes de reconnaissance de termes candidats (partie 2.3.1) avant de développer les techniques de validation (partie 2.3.2). Les méthodes de calcul de potentiels terminologiques seront abordées dans la partie 2.3.2 du fait de leur frontière parfois intangible avec l'étape de validation.

2.3.1 Identification des termes candidats

La reconnaissance des termes candidats constitue l'étape initiale de la construction automatique de terminologie. Elle consiste en l'identification de séquences de mots qui ont *des chances* d'être terminologiques. Le repérage de ces séquences peut s'appuyer sur une argumentation linguistique ([Justeson and Katz, 1995], autres ?), sur une argumentation plus statistique [Peñas et al., 2001] ou plus généralement sur une approche hybride. La majorité des techniques s'appuie sur des informations syntaxiques et/ou techniques que nous présenterons dans la partie 2.3.1.1. Les techniques indépendantes de la syntaxe seront présentées dans la partie 2.3.1.2.

2.3.1.1 Exploitation de la syntaxe

Le niveau morphosyntaxique est le niveau linguistique le plus simple à analyser, notamment par des processus automatisés. Les recherches se sont naturellement concentrées sur les outils fonctionnels à disposition afin de réaliser des expériences et de proposer des solutions.

2.3.1.1.1 Outils de TAL pour l'extraction terminographique En linguistique, le niveau d'analyse syntaxique correspond au niveau phrastique. Cela signifie que les unités supérieures – paragraphes, documents, corpora – ne sont pas prises en considération. Différentes analyses syntaxiques sont possibles, plus ou moins coûteuses et avec des taux d'erreurs variables. L'étiquetage morphosyntaxique – des langues dotées, i.e. pour lesquelles des corpora de *textes propres* et annotés sont disponibles – est peu coûteux et très précis : 97,3% pour Stanford CORNLP [Manning et al., 2014a], plus de 96% pour TREETAGGER [Manning, 2011], 95% pour BRILL [Brill, 1992], etc. L'analyse en constituants immédiats est une autre analyse syntaxique plus complexe qui vise à délimiter et hiérarchiser dans un arbre les différents syntagmes constituants une phrase. Cette analyse est généralement peu utilisée du fait du temps de traitement requis, de plus les performances des meilleurs systèmes restent modestes pour une automatisation complète avec une précision pouvant par exemple descendre autour de 85% pour Stanford CORENLP [Manning et al., 2014a]. Il n'en demeure pas moins que l'extraction terminographique automatique pourrait bénéficier d'analyses plus complexes que l'étiquetage morphosyntaxique ou l'analyse en dépendances, retenus dans une majorité de travaux de recherches.

2.3.1.1.2 Observations terminographiques empiriques Comme évoqué plus haut, la majorité des méthodes de reconnaissance des termes candidats exploitent la syntaxe à différents niveaux d'analyse – dépendances, étiquettes morphosyntaxiques, analyse syntagmatique, etc. Les travaux de [Justeson and Katz, 1995] sont fondateurs dans l'automatisation du processus d'extraction en posant des bases linguistiques empiriques simples, efficaces et qui ne nécessitent qu'un étiquetage morphosyntaxique. Les propriétés observées sont propres à l'anglais, avec une extension possible aux langues exploitant les racines greco-latines – cf. l'extraction de termes médicaux.

- Les termes sont très majoritairement des groupes nominaux, potentiellement prépositionnels, éventuellement composés d'un nom commun unique – bien que ces derniers ne soient pas pris en compte dans les travaux de [Justeson and Katz, 1995] ;
- *[...] 97% of multi-word terminological NPs in these sources consist of nouns and adjectives only, and more than 99% consist only of nouns, adjectives, and the preposition of.*, [Justeson and Katz, 1995]
- La probabilité qu'une séquence de mots soit effectivement terminolo-

gique décroît relativement à sa taille – plus la séquence est longue, moins elle a de chances d'être terminologique (pour une séquence de trois mots et plus). Les chercheurs placent la limite autour de cinq à six mots pour l'extraction.

— Ils précisent également que les résultats sont à nuancer en fonction du domaine des documents analysés : le domaine médical est particulier en ce que ses termes sont souvent composés d'un mot unique. Ce phénomène peut s'expliquer par les processus de formation des termes qui s'appuient sur la composition de bases gréco-latines.

Dans leurs recherches, [Justeson and Katz, 1995] comparent leurs mesures sur des corpora de domaines différents – fibre optique, médecine, physique et mathématiques, psychologie, ce qui permet de généraliser leurs observations, excepté pour la médecine, où la proportion de termes composés d'un mot unique est la plus importante. Les proportions observées sont les suivantes, pour 200 termes validés manuellement par catégorie :

thème	unigrammes	bigrammes	trigrammes	4+ grammes
médecine	44%	40%	11%	5%
psychologie	32%	60%	6%	2%
fibre optique	21,5%	54,5%	18%	6%
ph. & math.	20,5%	62,5%	14,5%	2,5%

Pour les différentes longueurs de graphies, nous pouvons par exemple observer chez [Justeson and Katz, 1995] :

Unigrammes : *melanuria*/NN, *synarthrophysis*/NN

Bigrammes : *linear*/JJ *function*/NN, *lexical*/JJ *ambiguity*/NN, *word*/NN *sense*/NN, *surface*/NN *area*/NN,

Trigrammes : *Gaussian*/JJ *random*/JJ *variable*/NN, *lexical*/JJ *conceptual*/JJ *paradigm*/NN, *cumulative*/JJ *distribution*/NN *function*/NN, *degree*/NN *of*/IN *freedom*/NN, *energy*/NN *of*/IN *adsorption*/NN,

4+ grammes : pas d'exemple fourni.

Voici quelques exemples de formation de termes à partir de racines gréco-latines dans le domaine médical :

hypercholestérolémie : Préfixe *hyper-* signifiant *excès*, suffixe *-émie* désignant une substance dans le sang, le mot complet désignant un excès de cholestérol dans le sang,

rhinopharyngite : Préfixe *rhino-* se rapportant au nez, *pharyng-* étant un dérivé de pharynx après mutation consonnantique du *x* du fait de la voyelle initiale du suffixe *-ite*, qui désigne une inflammation. L'ensemble désigne une inflammation du nez et du pharynx.

antiarythmique : Préfixe *anti-* signifiant contre, préfixe ici interfixe *-a-* signifiant sans, le tout désignant un médicament contre l'arythmie cardiaque.

Le processus de formation sur base gréco-latines est également observable dans d'autres langues, notamment l'anglais. [Justeson and Katz, 1995] détaillent spécifiquement les deux unigrammes mentionnés ci-dessus :

***melanuria* :** *melan-* pour *noir* et *uria* pour *urine*, le tout désignant une couleur anormalement foncée des urines,

***synarthrophysis* :** avec *syn-* pour *ensemble*, *arthro* pour *articulation* et *physis* pour croissance, le tout désignant la calcification d'une articulation aboutissant à sa disparition/dégradation.

2.3.1.1.3 Patrons morphosyntaxiques Les travaux de [Justeson and Katz, 1995] ont été repris et adaptés à de nombreuses occasions, permettant de généraliser leurs résultats mais également de remettre en question certaines observations empiriques probablement biaisées par la nature du matériel linguistique analysé. La performance de leur méthode de repérage des termes candidats nominaux anglais a été corroborée dans plusieurs travaux (par exemple [Frantzi et al., 1998], [Park et al., 2002]), mais certains postulats linguistiques sont fréquemment remis en question – la limitation aux formes verbales, l'exclusion des termes composés d'un mot unique, l'intérêt des structures prépositionnelles, etc. A partir de leurs observations, la méthode d'extraction terminographique proposée par [Justeson and Katz, 1995] repose donc sur l'identification de groupes nominaux potentiellement prépositionnels, identification réalisée à l'aide de *patrons morphosyntaxiques*.

Definition 2.3.1.1. Patron morphosyntaxique – Description d'une séquence d'étiquettes morphosyntaxiques. Un patron syntaxique peut être sous forme brute – NN NN pour une séquence de deux noms – ou sous forme d'expression rationnelle – NN^+ pour une séquence de noms de taille indéfinie.

A défaut de consensus sur la nature des éléments à extraire, les patrons syntaxiques se sont rapidement imposés comme une méthode de référence en

la matière. Ils ne nécessitent pas de prétraitements complexes, uniquement un étiquetage morphosyntaxique. Les faibles prérequis ont sans doute participé à leur succès. Cette simplicité est notamment à mettre en regard avec d'autres méthodes d'extraction, souvent appuyées sur des prétraitements plus lourds – par exemple l'ACI.

Le patron employé par [Justeson and Katz, 1995] est le suivant :

Patron 1 : $((\text{JJ}|\text{NN})^+((\text{JJ}|\text{NN})^*(\text{NN IN})^?)(\text{JJ}|\text{NN})^*)\text{NN}$

Le jeu d'étiquettes utilisé est celui du PENN TREEBANK [2] : JJ pour adjectif, NN pour nom commun, IN pour préposition. Quelques exemples de structures repérées par ce patron : *linear function, lexical ambiguity resolution, degree of freedom,* etc. De très nombreuses recherches en construction automatique de terminologies que nous présentons ici s'appuient sur un ou plusieurs patrons morphosyntaxiques [Peñas et al., 2001, Park et al., 2002, Frantzi et al., 1998, Saneifar et al., 2009, Condamines and Rebeyrolle, 2000, Navigli and Velardi, 2002] etc. Une consultation du jeu d'étiquettes du PENN TREEBANK nous informe que des formes adjectivales et nominales ne sont pas prises en compte dans le Patron 1 : les étiquettes JJR, JJS, NNS correspondant respectivement aux adjectifs comparatifs, superlatifs et aux noms communs au pluriel. Nous supposons néanmoins que [Justeson and Katz, 1995] ont inclus ces étiquettes au moins partiellement dans leur extraction, sans les faire apparaître dans le Patron 1 – certains éléments extraits présentés dans leur appendice sont au pluriel.

Le patron 1 a été comparé à d'autres patrons à des fins d'évaluation, notamment par [Frantzi et al., 1998] :

Patron 2 : NN^+NN

Patron 3 : $(\text{JJ}|\text{NN})^+\text{NN}$

L'observation des différents patrons nous indique une interdépendance entre ces derniers : les structures reconnues par le patron 2 sont également reconnues par le patron 3 et celles du patron 3 sont également reconnues par le patron 1. Cette observation combinée à celles de [Justeson and Katz, 1995] concernant la nature des structures terminographiques permet de dégager des tendances en fonction du patron : le patron 2 favorise la précision en n'acceptant que les noms (prédominants en terminologie) ; le patron 1 favorise le rappel en étant plus lâche avec des adjectifs et des groupes prépositionnels et

2. `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

le patron 3 se situe quelque part entre les deux. Ces observations sont corroborées par les résultats de [Frantzi et al., 1998], qui concluent à une supériorité du patron 3 relativement à leurs données. Il est notable que la gestion des groupes prépositionnels pose plusieurs problèmes pour l'extraction terminographique : i) ils sont peu fréquents, ce qui implique ii) une augmentation du bruit ; de plus iii) les structures prépositionnelles ne sont parfois pas annotées par les terminologues, comme dans le jeu de données d'évaluation terminographique ACL RD-TEC [3] [Handschuh and QasemiZadeh, 2014]. Ces limitations se retrouvent dans les travaux de [Park et al., 2002] au travers des conjonctions :

MODIFIER : (JJ(CC JJ)*)|NN|NNP|NNPS

Patron 4 : (DT (VBG|VBN))* MODIFIER* (NN | NNP | NNPS)

Où : CC est une conjonction de coordination, VBN un participe passé, VBG un gérondif et NNP un nom propre potentiellement au pluriel (NNPS). Les auteurs donnent des exemples de structures avec conjonction de coordination – *unpaved or dirty road*, *certain frontal or near-frontal collision*, etc. – mais aucune n'apparaît dans les 45 *meilleurs termes* listés dans leurs résultats. Les résultats de leurs recherches posent la question de l'intérêt des conjonctions de coordination dans la graphie des termes candidats : la compositionnalité sémantique induite par les conjonctions est-elle compatible avec le processus de formation de graphies terminologiques ? [Justeson and Katz, 1995] évoquent une forte lexicalisation des graphies terminologiques comme discriminant entre termes et non termes, lexicalisation qui semble difficile sur des graphies comprenant une articulation autour d'une conjonction. Dans leurs recherches, [Roche et al., 2004] reprennent l'idée de considérer les groupes prépositionnels dans les graphies, mais dans un contexte d'annotation manuelle. Les résultats présentés ne permettent pas de distinguer l'impact des groupes prépositionnels.

Avec TERMOSTAT [Drouin, 2003], des chercheurs ont proposé une évolution du patron 3 :

Patron 5 : (JJ|NN){0,5}NN

Les auteurs ont ajouté deux contraintes au patron initial : i) le nombre de mots maximum est fixé à six, ii) chaque mot doit être un *pivot lexical spécialisé*, à savoir un mot signalé comme discriminant après une analyse contrastive entre corpus de langue générique et corpus spécialisé. Les deux

3. `http://pars.ie/lr/acl-rd-tec-terminology`

contraintes découlent des observations de [Justeson and Katz, 1995] relativement à la taille maximale des graphies et à la surreprésentation de certaines graphies potentiellement terminologiques dans les documents techniques. Les auteurs concluent à un gain de précision/rappel avec l'utilisation des *pivots lexicaux spécialisés* pour les termes candidats avec une faible fréquence ($F(\text{CT}) < 3$) et à une perte de précision/rappel autrement.

Bien que s'appuyant aussi sur des patrons, les recherches de [Vergne, 2003] de celles précédemment mentionnées en ce qu'elles n'exploitent aucun outil de TAL mais des méthodes statistiques indépendantes de la langue analysée. Ces méthodes visent à discriminer entre mots vides et mots pleins, les séquences desquels vont permettre l'application de nouveaux patrons :

P : Mot plein ou non-vide,

V : Mot vide,

Patron 6 : P^+

Patron 7 : $\text{P}^+\text{V}^+\text{P}^+$

Patron 8 : $\text{P}^+\text{V}^+\text{P}^+\text{V}^+\text{P}^+$

La définition de *mot vide* choisie est la même que celle généralement employée :
« Nous avons choisi dans cet article les termes : « mot vide »-« mot plein », synonymes de « mot grammatical »-« mot lexical » et de « function word »-« content word » » [Vergne, 2003] [...].

L'automatisation de leur détection dans un contexte multilingue passe par une combinaison de deux critères qui reposent sur deux propriétés d'un mot donné m_1 : sa fréquence $F(m_1)$ et sa taille $length(m_1)$. A partir de ces deux mesures, [Vergne, 2003] appliquent la logique suivante : est qualifié de *vide* tout mot m_1 précédé par m_0 et suivi par m_2 si et seulement si

$F(m_1) > F(m_0)$, $F(m_1) > F(m_2)$
et $length(m_1) < length(m_0)$, $length(m_1) < length(m_2)$.

Exemple : *J'entraîne un réseau de neurones.*
Généralement, *F(réseau) < F(de)* et *F(neurones) < F(de).*
De plus, *length(de) < length(réseau)* et *length(neurones) > length(de).*
Donc *de* est qualifié de *vide.*

[Vergne, 2003] effectuent leurs expériences sur deux corpora :

Corpus 1 : "22 sites de la presse française nationale et régionale, 17 sites de la presse européenne (Suisse, Belgique, Allemagne, Italie, Espagne, UK, Irlande), et 4 sites de presse nord-américaine, chaque langue étant représentée par au moins deux sites."

Corpus 2 : "une centaine de sites publiés par Google News, environ la moitié étant des sites nord-américains, le reste du monde entier."

Bien qu'intéressante de par son indépendance vis-à-vis du langage analysé, les résultats présentés ne semblent pas probants comparativement aux recherches qui elles exploitent des outils de TAL. Les termes candidats suivants sont présentés comme étant les plus fréquents : *article, guerre, Jean-Luc Lagardère, monde, Açores*. De la même manière après validation des termes candidats, les termes suivants sont présentés comme étant les plus fréquents : *guerre, Lagardère, Jean-Luc Lagardère, monde, 15, 16, Aznar, Açores, empire*. Parmi les *meilleurs* termes candidats et termes proposés, a priori aucun n'est effectivement un terme. En revanche, la détection des mots vides indépendamment de la langue semble prometteuse, malgré la taille raisonnable de leurs corpora – 15 000 mots pour le Corpus 1 et 28 500 pour le Corpus 2. Aucun bruit n'est observable parmi les mots vides les plus fréquents (24 premiers) extraits sur les deux corpora :

Corpus 1 : *de, la, l', le, d', à, du, et, des, en, les, a, un, Le, La, L', in, une, Les, 's, to, pour, au, sur*

Corpus 2 : *to, in, of, the, 's, de, for, on, and, a, The, en, la, by, Al, with, is, A, from, at, i, 't, un, à*

Les auteurs signalent également un silence et un bruit notables dans l'extraction des graphies terminologiques liés à l'automatisation de la reconnaissance des mots vides : la non reconnaissance d'un mot vide mène à l'extraction de graphie non terminologique, la reconnaissance à tort d'un mot vide mène à la non reconnaissance d'une graphie potentiellement terminologique. Les auteurs ont ainsi pu observer un taux de bruit entre 2,7% (Corpus 2) et 4,3% (Corpus 1) et un taux de silence entre 2,6% (Corpus 1) et 10,7% (Corpus 2) – ces taux sont directement liés à la (non) reconnaissance des mots vides.

Graphies extraites à tort du fait de la non reconnaissance d'un mot vide :

Corpus 1 : *Was, Tutti, vous, About, Alors, Ein, Have, If, Mais, Qu',*

Wie, Wo, avant, contra, could, depuis, encore, faut, mieux, nous, now, plusieurs, that, tout, tutto

Corpus 2 : *This, How, Don', It, Most, contra, won', Alla, My, auf, One, Wer, Where, Why, après, down, einer, enough, only, they, when, which*

Pour rappel, l'extraction est multilingue. Aussi pouvons-nous observer du français (*vous, Alors, après*, etc.), de l'anglais (*Was, About, How, My,* etc.), de l'espagnol (*contra, Alla*), de l'italien (*Tutti, tutto*) et de l'allemand (*Ein, Wie, Wo, auf, Wer*, etc.).

Graphies non extraites du fait de la reconnaissance à tort d'un mot vide :

Corpus 1 : *War, paix, soir, war, aide, dimanche, Photo, baisse, Aide, Groupe, attendu, home, turn, voie, world*

Corpus 2 : *News, New, news, killed, Home, Help, Free, Global, Air, help, make, First, Get, get, groups*

Bien qu'il soit difficile de dégager des généralités à partir d'une expérience sur deux corpora de tailles modérées, la comparaison des résultats sur un corpus de 15 000 mots et sur un corpus de 28 500 – quasiment deux fois plus grand – met en évidence quelques tendances. Nous pouvons observer une augmentation du silence et une diminution du bruit relativement à l'augmentation de la taille du corpus :

$silence(\text{Corpus 1}) < silence(\text{Corpus 2})$
et $bruit(\text{Corpus 1}) > bruit(\text{Corpus 2})$

Ce phénomène est comparable aux performances des différents patrons morphosyntaxiques, que nous détaillerons en section 4.2 : les taux de bruit et de silence des différents patrons sont liés à leur niveau de contrainte. Un patron qui décrit un grand nombre de structures implique un silence faible (rappel important) mais un bruit plus élevé (précision moindre) ; réciproquement un patron qui ne décrit que peu de structures implique une précision plus élevée mais plus de silence. Ces observations ne sont évidemment valides que si le niveau de contrainte est recentré sur les éléments les plus pertinents sur le plan terminologique. Pour les mêmes raisons, la solution proposée par [Vergne, 2003] devrait voir son taux de silence augmenter et celui de bruit diminuer pour des corpora de tailles supérieures : davantage de données implique davantage de mots vides reconnus et davantage de mots

vides implique plus de précision (moins de mauvaises extractions de graphies terminologiques) et plus de silence (augmentation du nombre d'erreurs dans la liste de mots vides).

2.3.1.2 Autres méthodes

Comme présenté dans la partie 2.3.1.1, de nombreuses recherches en extraction terminographique s'appuient sur une forme de patron morphosyntaxique. D'autres méthodes du TAL permettent cependant d'identifier des lexies pertinentes de plusieurs mots, méthodes qui ont été expérimentées dans un contexte terminologique, sans exploiter de patron.

Nous pouvons identifier une autre manière de construire des lexies complexes : l'exploitation des cooccurrences et/ou des collocations. Les collocations se distinguent des cooccurrences en ce qu'elles induisent une connaissance lexicale, alors que les cooccurrences ne sont qu'une représentation statistique du vocabulaire d'un texte. Dans leurs recherches, [Enguehard and Pantera, 1995, Roche et al., 2004] exploitent les cooccurrences pour construire des lexies complexes. Leurs résultats démontrent la capacité des systèmes d'analyse en cooccurrences pour l'extraction terminographique, mais les méthodes de construction des lexies complexes peuvent être coûteuses : [Roche et al., 2004] procèdent à plusieurs phases d'extractions binaires ou ternaires pour reconnaître les lexies de plus de trois mots :
Exemple : *I am implementing a recursive neural network.*
Première analyse : *I am implementing a recursive*[neural_network].
Deuxième analyse : *I am implementing a* [recursive_neural_network.

A ces itérations s'ajoute l'étiquetage morphosyntaxique préalable, qui en fait une méthode d'extraction coûteuse relativement à l'application de patrons. Leurs évaluations, effectuées manuellement par un expert, montrent une précision dépassant les 94%. Il est nécessaire de préciser qu'à l'extraction des cooccurrences s'appliquent des contraintes morphosyntaxiques relatives à la formation des syntagmes pertinents, en l'occurrence terminologiques : une cooccurrence – avec un nom commun – ne sera considérée comme *valide* que si l'étiquette de l'élément qui cooccurre peut apparaître dans un syntagme nominal.

Les travaux de [Bourigault, 1992, Bourigault and Jacquemin, 1999, Bourigault et al., 1996], qui ont abouti à la création des outils LEXTER et FASTR, s'appuient quant à eux sur une méthode entièrement indépendante des patrons et des analyses en cooccurrences et collocations. Nous nous concentrons

ici sur LEXTER, qui effectue l'extraction de graphies terminologiques sans autre traitement. Nous présenterons FASTR dans la section 2.3.1.3, relative à la conflation de la variation et à la gestion des variantes de graphies.

Comme évoqué, LEXTER ne s'appuie pas sur des cooccurrences mais sur des règles de *découpage*. Le système s'appuie sur un raisonnement inverse : il cherche à identifier les limites des séquences de mots terminologiques, autrement dit les mots qui ne sont pas terminologiques. Cette identification est permise par un étiquetage morphosyntaxique et par les constats préalablement établis sur la nature des lexies terminologiques (section 2.3.1.1), notamment qu'il s'agit très majoritairement de groupes nominaux. Considérant ce fait, il reste à déterminer les éléments qui ne peuvent en faire partie. Les chercheurs combinent une grammaire permettant l'identification de groupes nominaux de taille maximale avec les frontières préalablement déterminées sur le plan morphosyntaxique – catégories grammaticales acceptées ou non. La grammaire – assimilable à un patron morphosyntaxique – nécessite ici d'être combinée avec le traitement des frontières pour en augmenter la précision : les lexies extraites à cette étape seront proposées comme termes candidats plus tard pour validation, la précision de cette étape de découpage doit donc être maximale – contrairement aux méthodes d'extraction avec tri vues en section 2.3.2. Les groupes nominaux extraits sont ensuite analysés pour extraire des groupes de taille plus restreinte à partir de l'identification de leur *tête* et de leur extension (méthode d'enrichissement qui a été introduite par [David and Plante, 1990] dans le logiciel TERMINO) : *bronchial cell* dans *cylindrical bronchial cell* et *cell* dans *bronchial cell*[Bourigault and Jacquemin, 1999].

La méthode d'extraction de graphies terminologiques proposée par ACABIT[Boulaknadel et al., 2008] (outil d'extraction terminographique) est similaire à celle de LEXTER, excepté pour le traitement des frontières. Les syntagmes nominaux sont extraits à partir d'un ensemble de règles réduit :

- NN JJ : *instruction publique*
- NN IN NN : *principe d'égalité*
- NN IN DT NN : *apprentissage de la lecture*
- NN NN : *apprenti lecteur*
- NN *à* VB : *savoir à enseigner*[Boulaknadel et al., 2008]

Bien que présentée sous form ed'une énumération de règles, les méthodes proposées dans ACABIT et dans LEXTER sont assimilables à des patrons morphosyntaxiques. Les règles énumérées ci-dessus, mentionnées dans [Boulaknadel et al., 2008], peuvent être transformées en un patron unique :

NN ((IN DT?)|à)? NN.

Nous avons pu identifier deux principaux groupes de méthodes d'extraction dans cette partie : les cooccurrences/collocations et la définition d'ensembles de règles ad-hoc. Nous avons également montré l'équivalence entre règles d'extraction et patrons morphosyntaxiques. Une comparaison entre cooccurrences et patrons reste pertinente : l'analyse en cooccurrences applique également des contraintes linguistiques permettant de définir ce qui constitue une cooccurrence *valide* dans le contexte d'une extraction terminographique. Autrement dit, l'analyse en cooccurrences applique de manière progressive les contraintes décrites dans un patron. En conclusion, quel que soit le système d'extraction automatique considéré, il repose nécessairement sur des contraintes morphosyntaxiques. Ces dernières peuvent être présentées sous forme de patrons, de règles, de contraintes progressives, mais elles tentent dans tous les cas de repérer des syntagmes spécifiques – nominaux ou verbaux.

2.3.1.3 Conflation de la variation

Ce que nous appelons *conflation de la variation* désigne l'ensemble des méthodes du TAL permettant la détection des variantes graphiques pour un concept donné. La lemmatisation et la racinisation sont des étapes de conflation simples : l'extraction du lemme ou de la racine permet de diminuer les variations inhérentes à la langue. La conflation de la variation permet d'identifier des graphies terminologiques qui seraient considérées comme trop peu fréquentes, de diminuer l'espace de représentation et d'améliorer la qualité des extractions terminologiques en général. Différentes formes de variations sont identifiables et pertinentes, formes que nous avons énumérées en section 2.2.4. De même que les limites de ce qui est terminologique sont floues, les limites de ce qui constitue une variante graphique le sont également, comme nous allons le voir.

Une autre forme de variation à été évoquée dans [Bourigault and Jacquemin, 1999] avec FASTR : les variations dites *syntaxiques*. Bien que leurs résultats soient intéressants, leur définition de *variantes* ne nous convient pas ici pour la première métarègle proposée : l'insertion d'adjectifs et d'adverbes. FASTR produit un réseau de termes candidats reliés par des relations de *variances* : *cylindrical bronchial cell* est considéré comme une variante de *bronchial cell*, *bronchial cell* comme une variante de *cell*. Nous considérons pour notre part qu'il s'agit davantage de relations de spécialisation plutôt que

de variations, en accord avec les constats effectués par [Justeson and Katz, 1995] sur la lexicalisation des graphies terminologiques. Nous avons défini la variation comme une relation d'équivalence entre deux graphies, aussi le phénomène de spécialisation n'y correspond-il pas : il n'y a pas d'équivalence entre *bronchial cell* et *cell*.

La deuxième métarègle proposée dans [Bourigault and Jacquemin, 1999] et dans [Boulaknadel et al., 2008] correspond à notre définition, métarègle intitulée *preposition switch & determiner insertion* chez [Bourigault and Jacquemin, 1999] qui permet l'insertion d'un déterminant et/ou le changement de la préposition : *distribution of words, distribution over words, distribution over the words, etc.* Contrairement à l'insertion d'adjectifs ou d'adverbes, l'insertion d'un déterminant ou un changement de la préposition ne spécialise pas le sens du nom, il y a bien une relation d'équivalence entre les graphies. Intéressante sur le plan théorique, l'impact de cette conflation n'est cependant pas évalué. Les erreur potentielles générées par une modification de la préposition ne sont pas abordées : le sens du nom n'est pas affecté par la préposition, mais le sens du syntagme peut l'être.

Les recherches de [Park et al., 2002] sont particulièrement intéressantes en ce qu'elles abordent plusieurs phénomènes générateurs de variance correspondant à la définition que nous avons donnée en début de cette section : i) la segmentation en mots, ii) les erreurs typographiques, iii) les abréviations et iv) les variantes flexionnelles.

i) Sont considérées comme équivalentes les graphies qui ne diffèrent que par leurs séparateurs : *treebank, tree-bank, tree bank.*

ii) Sont considérées comme équivalentes les graphies dont la distance d'édition est inférieure à 2. La distance d'édition correspond au nombre d'opérations (insertion, suppression, remplacement) nécessaire à la transformation d'une chaîne en une autre. La distance d'édition est fréquemment utilisée à ces fins, mais peut produire des erreurs pour des lexies distinctes ne différant que de peu de lettres. La distance d'édition tend également à ignorer certains morphèmes sémantiques courts qui marquent des modifications de sens : *trained, untrained, accurate, inaccurate,* etc.

iii) Sont considérées comme équivalentes les graphies et leurs acronymes/abréviations respectifs. Les acronymes/abréviations sont extraits à partir d'un procédé décrit dans [Park and Byrd, 2001], article dans lequel les auteurs formalisent les règles de construction des acronymes/abréviations à partir d'une étude sur corpus.

iv) Sont considérées comme équivalentes les graphies et leurs formes lem-

matisées. Dans [Park et al., 2002], seul le mot final est lemmatisé. Nous reviendrons sur ce point plus en détail dans nos expériences en section 4. Nous verrons notamment comment nous avons également appliqué une lemmatisation contrainte, mais pas autant que celle proposée dans [Park et al., 2002].

La conflation de la variation est une étape qui bénéficie à toutes les tâches de TAL. Nous avons présenté ici des solutions qui ont été proposées à cet effet dans notre contexte. Nous avons pu délimiter le champ des variations qui concernent spécifiquement l'extraction terminographique à partir des observations de [Justeson and Katz, 1995] : les graphies terminologiques sont fortement lexicalisées, elles tendent à n'accepter aucun nouvel élément en leur sein. La reconnaissance de ces variantes graphiques est une tâche sans fin d'une complexité croissante.

2.3.2 Validation des termes candidats

Nous présentons ici des solutions proposées dans le cadre de la validation de termes candidats. Pour rappel, un terme candidat désigne un ensemble de graphies extraites correspondant à un concept supposément terminologique. Ce que nous appelons ici validation consiste en l'étape de sélection des termes candidats pertinents parmi l'ensemble extrait à partir de méthodes de calculs de potentiels terminologiques. Afin de resituer l'étape décrite ici, rappelons les différentes étapes de l'automatisation du procédé d'extraction :

- Extraction des graphies correspondant à un modèle donné,
- Conflation de la variation et création de termes candidats,
- Validation des termes candidats selon un critère donné – poids, occurrence dans un dictionnaire, etc.
- Evaluation des termes candidats validés.

Nous avons présenté les méthodes d'extraction de graphies en section 2.3.1 et la conflation de la variation en section 2.3.1.3, nous présentons maintenant l'étape que nous avons intitulée *validation*. Au-delà d'une présentation et d'une typologie des solutions proposées, nous verrons comment certains systèmes ont pu s'affranchir de cette étape. Des chercheurs ont comparé diverses mesures dans [Pazienza et al., 2005], dont certaines dont nous allons parler. A partir d'une terminologie de référence construite manuellement par des experts, les chercheurs ont pu observer une prévalence des mesures fortement corrélées à la fréquence relativement aux mesures dites d'association. Autrement dit, les comparaisons effectuées ont permis de déterminer que la

fréquence d'une graphie est un marqueur terminologique plus pertinent que le degré d'association – ou niveau de lexicalisation – des mots qui la composent.

Deux types de mesures sont à distinguer dans l'étape de validation : les mesures de lexicalisation ou d'association – *unithood* – et les mesures de l'aspect terminologique – *termhood*. Les mesures de lexicalisation sont généralement utilisées en combinaison avec une autre métrique. Chez [Park et al., 2002], les chercheurs combinent linéairement une mesure de lexicalisation (*term cohesion*) à une mesure basée sur les fréquences (*domain specificity*). Nous retrouvons également une combinaison linéaire chez [Navigli and Velardi, 2002], qui associe également une mesure basée sur les fréquences (*domain relevance*) à une mesure de lexicalisation (*domain consensus*). Chez [Frantzi et al., 1998], la mesure de la lexicalisation n'est pas combinée linéairement mais est en quelque sorte incluse dans le calcul de la *C-valeur*. Cette dichotomie entre aspect terminologique et mesure de lexicalisation a été initialement identifiée par [Kageura and Umino, 1996]. Parmi les mesures basées sur les fréquences proposées par la suite nous pouvons identifier une hypothèse récurrente : celle de l'analyse contrastive.

L'analyse contrastive met en application la définition même d'*aspect terminologique*, à savoir la spécificité à un domaine. Largement utilisée [Park et al., 2002, Navigli and Velardi, 2002, Drouin, 2003, Peñas et al., 2001], l'analyse contrastive repose sur une postulat simple : un terme donné apparaît davantage dans un corpus de son domaine relativement à un corpus de *langue générale*. Le corpus de langue générale est utilisé comme point de levier pour discriminer les graphies terminologiques/spécifiques des autres. Au delà même de la formule exploitée pour associer un poids à un terme candidat, la définition et la constitution d'un corpus dit *de langue générale* est problématique.

Comme nous l'avons vu, la terminologie relève du vocabulaire spécifique à un domaine. Hors, il est difficile de concevoir un document qui n'ait trait à aucun domaine. La problématique revient alors à identifier des documents les moins spécialisés pour constituer le corpus ou à identifier des documents de domaines particulièrement distincts – solution ad-hoc particulière à la terminologie d'un domaine donné [Velardi et al., 2001]. Dans leurs recherches, [Velardi et al., 2001] construisent par exemple le corpus de langue générale à partir d'une combinaison de corpora : le corpus Brown [Francis and Kucera, 1964], un extrait du Wall Street Journal, des textes médicaux, sportifs, de tourisme, de description d'hôtels ainsi que des nouvelles de Wells. Le tout formant un corpus de $3,2$ millions de mots, ce qui semble très raisonnable pour

un corpus visant à représenter la langue générale. Nous retrouvons approximativement les mêmes proportions chez [Peñas et al., 2001] : ils construisent le corpus à partir de 7400 articles provenant d'un site d'actualités internationales unique (`https://elpais.com/`) et aboutissent à un corpus de $2,9$ millions de mots. Chez [Park et al., 2002], l'exploitation du corpus de langue générale est détaillée mais pas sa constitution ni ses dimensions.

Un postulat fréquent consiste à considérer les textes de journalisme généralistes comme suffisamment neutres pour permettre l'analyse contrastive [Drouin and Doll, 2008, Velardi et al., 2001, Peñas et al., 2001]. Le postulat se justifie par plusieurs raisons : les textes sont nombreux et facile d'accès – argument pragmatique – et les thèmes abordés dans les articles sont généralement vulgarisés, induisant une faible proportion de lexies terminologiques. Bien qu'efficace, la vulgarisation ne permet pas d'identifier toutes les lexies non terminologiques qui peuvent être sur-représentées dans un corpus spécialisé. La problématique a été identifiée dans les recherches précédemment mentionnées et concerne particulièrement deux points :

— Les lexies terminologiques à faibles fréquences seront ignorées : une faible fréquence dans le corpus spécialisé induit une faible différence avec le corpus de langue générale, donc un faible aspect terminologique. Toutes les mesures d'aspect terminologique basées sur les fréquences – i.e. les plus performantes [Pazienza et al., 2005] – sont concernés par cet écueil.

— Des lexies fréquentes seront considérées à tort comme terminologiques. Certains éléments de structures inhérents à la littérature scientifique pourront par exemple être concernés : *tableau ci-dessus*, *graphique ci-dessous*, etc. Ce genre de lexies hautement fréquentes dans les articles scientifiques n'apparaissent généralement pas dans les articles d'actualité.

La méthode de la *NC-valeur* [Frantzi et al., 1998] se distingue de celles évoquées ci-dessus en ce qu'elle ne correspond pas à la typologie préalablement établie. La *NC-valeur* se définit comme une méthode de tri de termes candidats à partir d'une combinaison linéaire de deux poids : la *C-valeur* et ce que nous appellerons le facteur contexte, sans dénomination à l'origine. Présenté simplement – nous y reviendrons en détail dans la section 4 – la *C-valeur* s'appuie sur une simple analyse des sous-chaînes de caractères d'une graphie terminologique donnée : elle associe à la fréquence d'une graphie des informations sur les sous-chaînes qui y sont incluses pour lui attribuer un poids. De fait, la *C-valeur* est une mesure basée sur les fréquences

qui induit une forme de mesure de lexicalisation au travers des graphies terminologiques plus courtes contenues dans la graphie initiale.
Exemple : La *C-valeur* de *recurrent neural network* est affectée par sa propre fréquence et celle de *neural network*.

Le facteur contexte consiste quant à lui en une analyse en cooccurrences d'une graphie terminologique avec des marqueurs externes non terminologiques, à savoir des noms, verbes et adjectifs. A chaque élément du vocabulaire du contexte – noms, adjectifs, verbes – est associé son taux de cooccurrences avec une graphie terminologique, i.e. sa proportion d'apparition contiguë à une graphie terminologique par rapport à son nombre total d'occurrences. Après combinaison linéaire avec la *C-valeur*, la *NC-valeur* obtenue se trouve affectée par des éléments relatifs au contexte d'utilisation de la graphie terminologique. Une problématique rencontrée – qui sera détaillée dans la section 4 – a trait à la définition du *cadre* autour de la graphie terminologique lors du calcul du facteur contexte : la taille du cadre n'est pas explicitement mentionnée chez [Frantzi et al., 1998]. Il n'y également pas de mention d'un traitement des mots vides à ce niveau d'analyse. Par taille du cadre nous entendons le nombre de mots pris en compte autour d'une graphie terminologique lors du calcul du facteur contexte : le cadre peut être un paragraphe, une phrase, un nombre de mots spécifique, etc. Le choix de la taille du contexte peut également être lié à la taille du corpus en nombre de mots ; la première est théoriquement inversement proportionnelle à la seconde.

La *NC-valeur*, ou plus spécifiquement la *C-valeur* est l'une des mesures d'aspect terminologique les plus usitées et dont la performance a été observée à plusieurs reprises. Nous détaillerons particulièrement cette méthode en sections 3 et 4 car nous l'avons retenue pour nos expériences.

Afin de constituer une véritable étape de validation, la liste triée de termes candidats obtenue est ensuite scindée en deux à partir d'un seuil défini ad-hoc : les termes candidats dont le poids est supérieur à ce seuil sont validés, les autres sont rejetés. Les termes retenus à cette étape sont ensuite évalués, autant que faire se peut comme nous allons le voir.

2.4 Techniques d'évaluation d'extraction automatique de termes

Le processus d'extraction terminologique vise à construire une liste de termes candidats à partir d'un ou plusieurs documents. L'appellation *terme candidat* illustre le fait que la liste construite doit ensuite être filtrée, manuellement ou (semi)automatiquement avant de constituer des termes à part entière. Les termes candidats sont fréquemment extraits à partir de patrons morphosyntaxiques [Daille, 1994, Frantzi et al., 1998, Justeson and Katz, 1995, Dagan and Church, 1994] correspondant à des séquences de catégories grammaticales. Une fois les termes candidats extraits, un ou plusieurs filtres sont appliqués pour ne conserver que les termes. Ces filtres peuvent fournir une réponse binaire quant au potentiel terminologique du candidat, ou fournir une méthode de tri des candidats en fonction de ce même potentiel.

Se pose alors la question de l'évaluation des lexiques construits par ces méthodes automatiques. Théoriquement, il serait nécessaire de comparer une extraction terminologique de référence à une extraction terminologique automatique sur un même corpus. Cela permettrait d'obtenir les mesures qualitatives classiques que sont le rappel, la précision et la *F-mesure*. L'annotation manuelle des syntagmes terminologiques est une tâche fastidieuse qui ne peut être faite que par des connaisseurs du domaine du corpus analysé, d'où l'indisponibilité de ces ressources. Différentes méthodes d'évaluation ont été proposées à partir de ces constats.

2.4.1 Bases de connaissances externes

Des chercheurs ont proposé d'exploiter des ressources externes afin d'évaluer le lexique construit. Ces ressources peuvent être sous la forme de dictionnaires spécialisés [Velardi et al., 2001] ou d'ontologies (WORDNET et autres). L'utilisation de ces ressources présente cependant un biais, elles rendent impossible le calcul du rappel. Ces bases de connaissances permettent de déterminer si un élément du lexique est effectivement terminologique, mais elles ne permettent pas de mesurer le silence. Seul un rappel relatif – $card(lexique \cap dictionnaire)/card(dictionnaire)$ – peut être calculé afin de comparer différentes méthodes d'extraction sur un même corpus. L'exploitation de ces ressources externes pose également la question de leur évolution relativement à l'émergence de nouveaux termes, i.e. les néologismes.

2.4.2 Moteur de recherche

L'exploitation par un moteur a été proposée notamment par [Paryzek, 2008]. Le concept consiste à corréler le niveau de lexicalisation d'un candidat néologique au nombre de réponses fourni par le moteur de recherche. Bien que [Paryzek, 2008] utilise le moteur de recherche pour construire sa terminologie, une extension vers une méthode d'évaluation peut être envisagée. Cette proposition est intéressante en ce qu'elle s'appuie sur un principe linguistique fort : *l'usage définit la norme*. Un moteur de recherche externe (Google, Bing, DuckDuckGo, etc.) présente l'avantage de ne pas être statique : l'apparition/l'usage d'un nouveau terme est reflété dans le nombre de correspondances renvoyé à partir d'un terme.

2.4.3 Comparaisons avec un tri sur les fréquences

Comme c'est fréquemment le cas pour le développement de modules de TAL, une *baseline* peut être définie pour estimer l'intérêt de la méthode expérimentée. Une *baseline* peut être vue comme la solution la plus simple à mettre en place pour répondre à une problématique exprimée ; une qualité inférieure à la *baseline* indique une déperdition d'information avec la nouvelle solution. La *baseline* en extraction terminographique automatique consiste généralement en un tri du vocabulaire à partir de ses fréquences. Associée à une terminologie de référence, cette liste triée permet d'obtenir une qualité minimale à dépasser lors de l'expérience.

2.4.4 Annotation manuelle

Comme évoqué précédemment, l'annotation terminographique est une tâche fastidieuse et qui est réservée aux spécialistes du domaine du corpus analysé. Elle nécessite un protocole d'annotation pointu et la prise en compte de l'accord inter-annotateurs afin de décider de l'aspect terminologique d'un syntagme. Nous avons par exemple calculé des coefficients dits *Kappa de Cohen* [Cohen, 1960] pour un projet de développement d'un étiqueteur morphosyntaxique pour le macédonien [Fort, 2014].

L'annotation manuelle permet d'obtenir le lexique le plus proche de la *vérité*, mais le faible consensus sur les définitions et limites des graphies terminologiques explique le peu de ressources annotées disponibles. La complexité de l'évaluation manuelle d'une extraction terminographique induit

généralement une restriction à un nombre de termes candidats plus restreint [Park et al., 2002, Benavent and Parrilla, 2006, Pazienza et al., 2005].

[Benavent and Parrilla, 2006] évaluent manuellement de manière très fine la sortie du logiciel commercial Extraterm de Trados [4]. Trados permet d'enrichir une terminologie automatiquement. L'extraction automatique a été appliquée sur un corpus espagnol hautement spécialisé de 144 documents et environ 2,5 millions de mots. Les documents du corpus traitent de matériaux céramiques. L'extraction réalisée vise à assister l'enrichissement d'une ontologie. L'évaluation s'est quant à elle limitée à certains aspects de la chaîne de traitements : les prétraitements linguistiques et l'évaluation manuelle des candidats extraits. L'évaluation consiste non pas en une réponse binaire mais en une validation de liens sémantiques entre candidats.

Parmi les 66 000 termes candidats extraits, seuls les 1 850 plus fréquents sont évalués manuellement. De nombreux défauts et biais sont constatés :

- Des erreurs typographiques dues à la transformation avec OCR (*Optical Character Recognition*) perturbent les résultats.
- Sur les 1 850 termes candidats analysés, uniquement 197 sont réellement des termes. L'extraction produit donc une importante proportion de bruit (faux positifs). Son taux de rappel ne peut être estimé sans référence annotée.
- De nombreux candidats ne correspondent pas à des éléments sémantiques.
- Certains candidats extraits ne sont constitués que de mots de la langue courante.
- Certains candidats extraits n'ont pas trait aux matériaux céramiques mais à d'autres domaines comme l'optique, la physique ou l'architecture.
- TreeTagger[Manning, 2011] peine à étiqueter correctement des mots qui ne font pas partie du vocabulaire initial notamment lorsqu'ils sont fléchis.
- Certains candidats extraits sont des parties de termes valides plus longs, mais des parties qui ne constituent pas des graphies terminologiques à part entière.

Les biais identifiés par [Benavent and Parrilla, 2006] sont représentatifs des problématiques rencontrées pour l'évaluation d'une extraction termino-

4. Actuellement SDL MultiTerm de SDL Trados (`https://www.sdltrados.com/products/multiterm-extract/`).

graphique, au-delà même du cadre d'une évaluation manuelle.

FIGURE 2.3 – Illustration des différents niveaux d'analyse

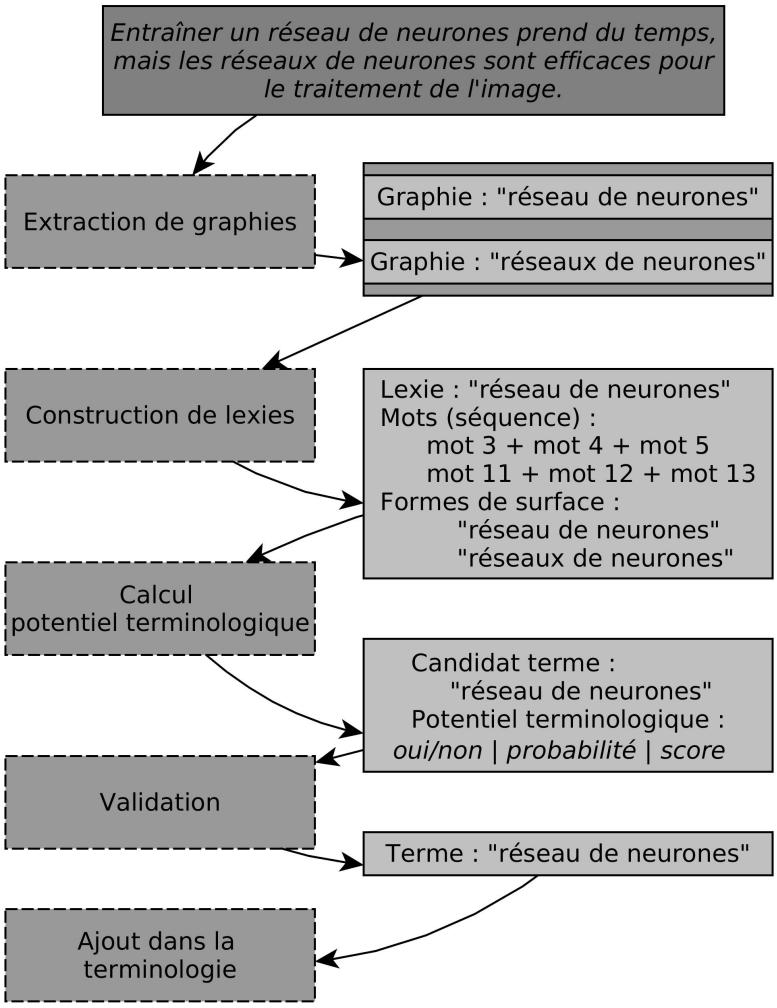


FIGURE 2.4 – Exemple de manifestation d'une ambiguité du fait de la polysémie de *neural network*

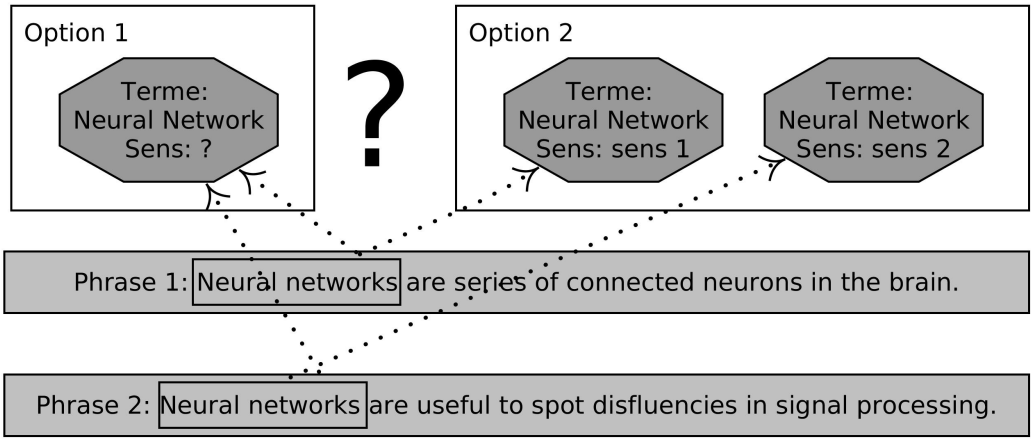


FIGURE 2.5 – Exemple de manifestation de la néologie sémantique

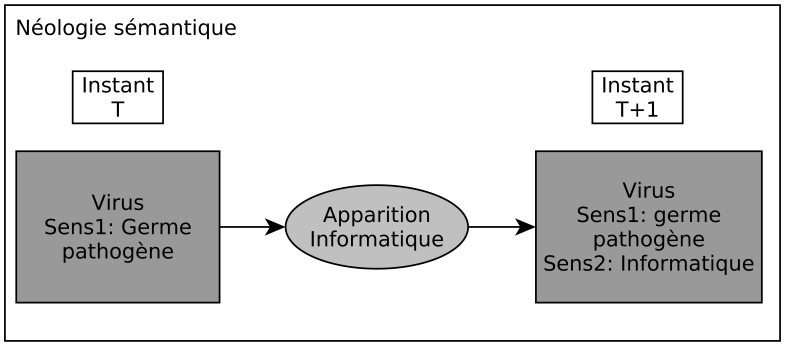


FIGURE 2.6 – Manifestation de la problématique posée par la détection de la néologie sémantique dans le cadre d'une analyse en diachronie à partir de ressources externes statiques. KB : K*nowledge* B*ase*. Notez que cette dernière est ancrée dans l'instant T, mais qu'elle est indifférement utilisée à T et à T+1.

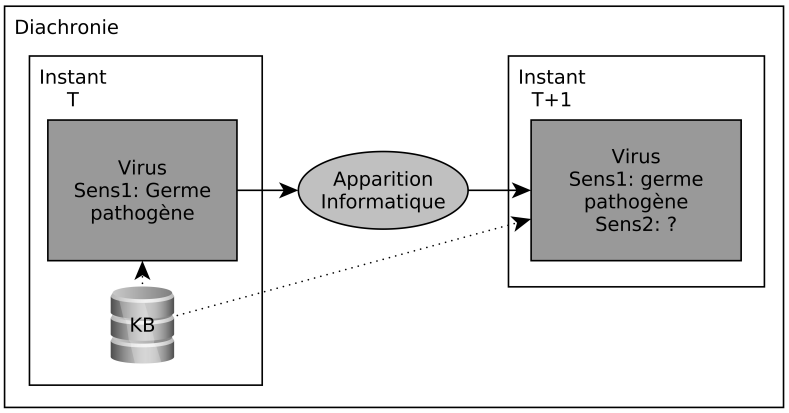
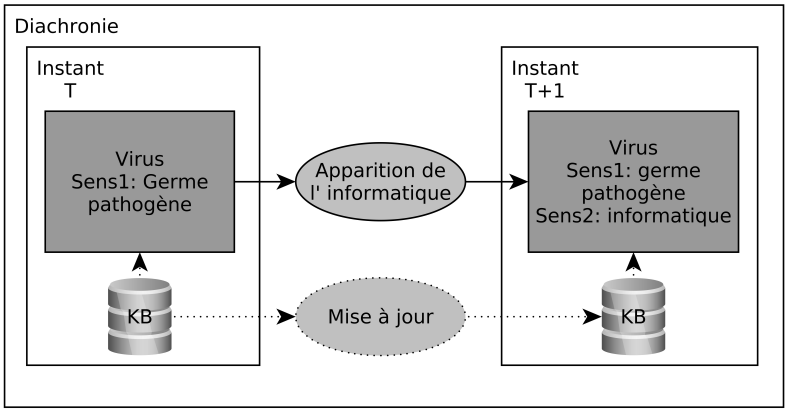


FIGURE 2.7 – Détection de la néologie sémantique dans le cadre d'une analyse en diachronie à partir de ressources externes également en diachronie. KB : K*nowledge* B*ase*. Notez que contrairement à la figure 2.6, cette dernière n'est plus ancrée dans l'instant T.



Chapitre 3

Notre modèle - Notre apport

L'extraction terminographique automatique est une tâche du TAL complexe tant au niveau de l'exécution qu'au niveau de l'évaluation. Nous l'avons évoqué en section 2.2, différents phénomènes, entre autres linguistiques, constituent des obstacles à la bonne qualité de l'extraction. Nous présenterons dans une première partie nos propositions relatives à la prise en compte de la polysémie (section 3.1), notamment sur l'exploitation des modèles de thèmes à des fins de désambiguisation i.e. de résolution de polysémie. Ces propositions passeront par une élaboration du lien théorique entre les modèles de thèmes et l'extraction de graphies terminologiques. Nous porterons notre attention sur l'ambivalence du lien entre les deux domaines : les bénéfices des modèles de thèmes pour l'extraction terminographique et les bénéfices de l'extraction terminographique pour les modèles de thèmes. Afin d'analyser ce lien, nous définirons le concept de modèle de thèmes dans la section 3.1.1 en nous concentrant sur la pragmatique. La pragmatique incidente aux modèles de thèmes sera notre fondement théorique quant à l'existence d'un lien entre modèle de thèmes et extraction terminographique, nous le verrons en section **??**. Subséquemment à cette analyse, nous présenterons trois axes de travail validés relatifs à ce lien :

1. Les modèles de thèmes peuvent bénéficier de connaissances issues d'une extraction terminographique : nous verrons comment *hybrider* une modélisation thématique avec une extraction terminographique, ainsi que les conséquences afférentes à cette hybridation.
2. Il y a une corrélation entre i) la qualité d'un modèle de thèmes sur un corpus donné et ii) la similarité textuelle des documents du corpus. De

plus, à partir de notre première hypothèse, nous pouvons également espérer observer une corrélation plus forte sur un système hybride – avec les graphies terminologiques – que sur un système standard.

3. Le modèle de thèmes que nous utilisons dans nos expériences est *flou* et sa construction *non-déterministe*. Nous verrons comment exploiter ses résultats pour obtenir une classification stricte et approximer les résultats les plus représentatifs pour un corpus donné. Nous verrons pour cela une adaptation d'un algorithme d'alignement de la littérature : l'algorithme dit *hongrois* dans la section 3.1.3.

Dans un second temps, nous argumenterons dans le sens d'un élargissement des patrons morphosyntaxiques proposés dans la littérature. Plus précisément, nous détaillerons le bien-fondé théorique de l'extension de certaines étiquettes morphosyntaxiques. Pour ce faire, nous développerons l'intérêt des structures nominales ainsi que des limites les concernant dans le cadre d'une extraction terminographique. Nous préciserons les distinctions entre les notions de groupes nominaux et les graphies reconnues par les patrons morphosyntaxiques en section 3.2.1.1 avant d'aborder les problématiques liées aux entités nommées à partir des patrons en section 3.2.1.2. Nous verrons en quoi la littérature peut induire une certaine confusion entre les rôles syntaxiques – groupes nominaux, verbaux, etc. – et les graphies potentiellement terminologiques reconnues par les patrons proposés. La même confusion règne quant à la distinction entre graphie terminologique et entités nommées, confusion que nous tenterons d'éclaircir.

Toujours relativement aux structures syntaxiques des graphies terminologiques, nous verrons en sections 3.2.2 et 3.2.3 le rôle des autres groupes syntaxiques – verbaux, adjectivaux, prépositionnels, etc. Nous développerons la complexité de l'automatisation de l'extraction de terminologie verbale : la terminologie verbale s'appuie sur des règles syntaxiques et sémantiques particulièrement complexes à automatiser, nous y reviendrons. Nous investiguerons ensuite l'implication des autres groupes syntaxiques dans la formation de graphies terminologiques :

— **Groupes nominaux :** Peuvent constituer une graphie terminologique à lui seul,
— **Groupes verbaux :** Peuvent constituer une graphie terminologique mais nécessite des spécifications supplémentaires,
— **Groupes prépositionnels, adjectivaux, compléments du verbe/nom, propositions subordonnées et autres satellites :** Ne peuvent pas

constituer une graphie terminologique à lui seul, ne peut exister que comme satellite d'un noyau – un groupe nominal ou verbal.

Dans un troisième et dernier temps, nous introduirons une méthode d'évaluation issue d'un domaine connexe au TAL, celui de la recherche d'information. Nous verrons en quoi les évaluations proposées dans la littérature sont souvent complexes à mettre en œuvre et/ou incomplètes, l'un justifiant souvent l'autre : la complexité de la tâche d'évaluation limite le nombre de graphies évaluées, rendant de fait l'évaluation incomplète. Nous prolongerons la méthodologie d'évaluation des sytèmes de recherche d'information pour affiner les mesures d'évaluation des sytèmes d'extraction terminographique en section 3.3.4. Nous développerons également l'intérêt des mesures d'évaluation de systèmes de recherche d'information dans le cadre d'une automatisation complète d'une extraction terminographique : ne pas limiter la liste de graphies à évaluer permet davantage de rappel aux détriments de la précision, mais l'analyse distributionnelle permet d'observer la répartition des graphies effectivement terminologiques parmi l'ensemble de graphies extraites.

A des fins de clarté, nous avons fait le choix de distinguer clairement sur le plan rédactionnel les axes de recherche des expériences : l'ensemble des expériences relatives aux axes de recherche précédemment mentionnés et développées ci-dessous peuvent être consultées en section 4.

3.1 Problématique de la polysémie : intérêt des modèles de thèmes

Un modèle de thèmes désigne une partition d'un ensemble de documents relativement aux domaines auxquels ils appartiennent. Nous précisons ici les problématiques afférentes à l'automatisation de la construction de ces modèles ainsi que le lien pragmatique entre modèle de thèmes et extraction terminographique, spécifiquement dans une perspective de désambiguisation.

3.1.1 Modèle de thèmes : définitions et intérêts

Definition 3.1.1.1. Modèle de thèmes – Un modèle de thèmes est le résultat de la découverte automatique des thèmes/domaines/sujets dans un corpus de documents. Le modèle consiste en des représentations abstraites de

thèmes construites à partir du vocabulaire des documents. Nous distinguerons le modèle de thèmes, qui est le résultat, de l'algorithme qui l'a construit.

La construction de modèles de thèmes est une tâche classique du TAL depuis le début des années 1990 avec des travaux comme ceux de [Deerwester et al., 1990], de [Hofmann, 1999], ou encore de [Schultz and Liberman, 1999] : [Deerwester et al., 1990] ont posé les bases de la construction de modèles à partir de manipulations matricielles, bases notamment reprises par [Hofmann, 1999] qui en proposent une adaptation probabiliste. Les travaux de [Schultz and Liberman, 1999] s'insèrent davantage dans le cadre des calculs de similarités textuelles traditionnels au travers de cosinus de vecteurs de $tf.idf$. Ces trois travaux illustrent l'idée prédominante derrière l'automatisation de la construction de modèles de thèmes : la distribution du vocabulaire dans un document doit permettre l'identification du domaine dont il est issu, indépendamment de toute connaissance linguistique – hors segmentations en phrases et mots. Ce fonctionnement nous mène à envisager la construction de modèles de thèmes comme une spécialisation des algorithmes de *clustering*, ce qui nous a incités à introduire la méthode d'évaluation que nous proposons plus loin en section 3.1.3 : les *clusters* peuvent être construits à partir de données qui peuvent être extra-linguistiques ; les modèles de thèmes sont construits uniquement sur les fréquences des mots. Quel que soit l'algorithme de construction de modèles de thèmes ou de clustering choisi, certaines problématiques sont à considérer.

— Spécifier le nombre de thèmes a priori dans un contexte non supervisé : un nombre de thèmes spécifié a priori peut induire des partitions qui n'ont pas lieu d'être ou qui devraient être. Pour un corpus de *Médecine*, *Automobile* et *Finance*, demander la construction de deux thèmes seulement impliquerait de *mauvaises* associations entre thèmes et documents. Similairement, demander la construction de quatre thèmes – un de plus que nécessaire – aboutirait à des résultats non pertinents relativement à l'objectif attendu. La figure 3.2 illustre l'impact d'une sous-estimation du nombre de thèmes à identifier relativement au corpus. Si l'impact d'un sur-partitionnement peut être nul, celui d'un sous-partionnement ne peut être que négatif : les partitions supplémentaires peuvent être vides, une partition manquante ne peut qu'être source de bruit/silence.
— Ne pas spécifier le nombre de thèmes : certains algorithmes de construction de modèles de thèmes ne nécessitent pas de spécification a priori

du nombre de thèmes à identifier. Diverses analyses mathématiques sont appliquées afin de déterminer ce nombre, mais qui peut de fait varier lorsque la méthode est non-déterministe. De plus, ces estimations sont généralement issues de processus itératifs relativement lourds : qu'il s'agisse du *Chinese Restaurant Process*[Aldous, 1985] ou de l'échantillonnage de Gibbs pour LDA, la détermination du nombre de thèmes est le résultat de la convergence de nombreuses exécutions. Enfin, l'automatisation de cette détermination pose clairement la question du degré de finesse de l'analyse : dans un corpus de *Médecine*, *Automobile* et *Finance* peuvent co-exister des sous-thèmes qui aboutiront à accroître le nombre de thèmes potentiellement identifiés. Les algorithmes de construction de modèles de thèmes hiérarchiques répondent partiellement à cette problématique.

— Supervision pour les modèles de thèmes : la supervision de la construction des modèles implique la constitution d'un corpus d'entraînement, comme pour la classification automatique. Cependant, avec une liste de thèmes possibles fixée, l'utilisation d'une méthode supevisée pose la question de la souplesse des modèles relativement à l'analyse de thèmes absents du corpus d'entraînement. Si le corpus d'entraînement contient les domaines *Médecine*, *Automobile* et *Finance*, l'identification du domaine *Littérature* ne pourra pas se faire directement. Des solutions ont été proposées dans ce sens mais la problématique reste ouverte.

— Identification d'un thème : nous avons jusqu'alors identifié les thèmes à partir d'étiquettes ou *labels* – *Médecine, Automobile, Finance, Littérature* – mais celles-ci sont absentes des modèles construits. Des méthodes ont été proposées afin de déterminer l'étiquette la plus cohérente pour un thème donné à partir de sa distribution de vocabulaire [Mei et al., 2007, Lau et al., 2011, Mehrotra et al., 2013], mais il s'agit d'une problématique distincte. Il est à noter que nous nommerons les thèmes le cas échéant à des fins de clarté, malgré le fait qu'un thème ne soit défini que par sa distribution sur un vocabulaire.

— Evaluation des thèmes construits : nous verrons comment cette évaluation passe par une transformation d'un modèle flou vers un modèle strict : un document est associé à un unique thème. Ce passage du flou au strict ne répond cependant pas à toutes les problématiques rencontrées lors d'une évaluation à partir d'un corpus de référence : si l'association entre thème et document n'est jamais ambiguë pour

FIGURE 3.1 – Mauvais partionnement suite à la spécification a priori du nombre de thèmes : un thème supplémentaire est contruit.

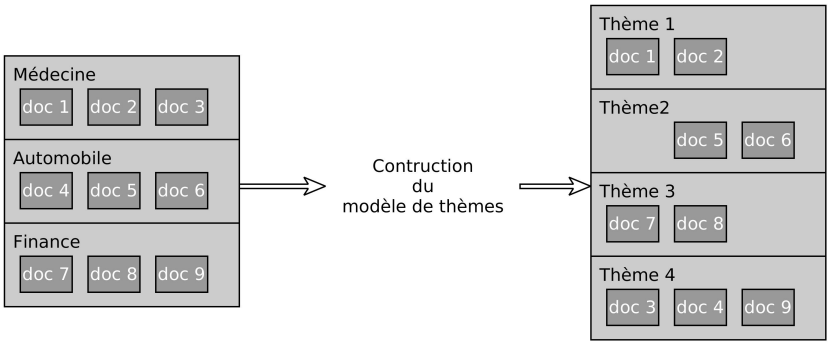
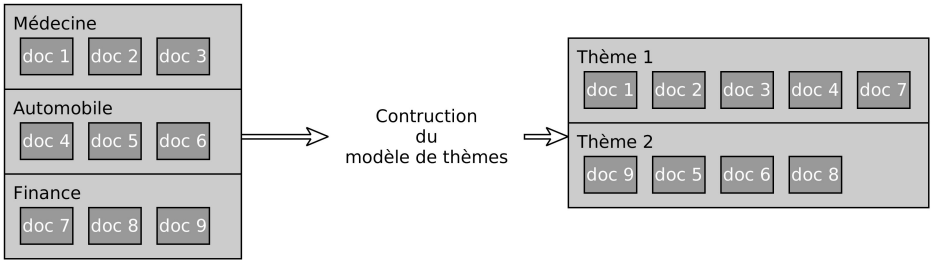


FIGURE 3.2 – Mauvais partionnement suite à la spécification a priori du nombre de thèmes : un thème à identifier est manquant.



une partition en deux sous-ensembles dans un modèle flou, ce n'est plus le cas pour trois ou plus partitions. Nous reviendrons en détail sur l'évaluation en section 3.1.3.

La figure 3.1 illustre l'identification d'un thème supplémentaire par l'algorithme de construction évoqué dans le premier point ci-dessus.

Nous venons de présenter certaines problématiques inhérentes à la construction des modèles de thèmes, nous poursuivons maintenant avec une justification du bien-fondé d'une hybridation entre construction de modèles de thèmes et extraction terminographique.

3.1.2 Apport des termes candidats dans les modèles de thèmes

L'extraction terminographique peut se définir comme l'extraction de graphies ayant un sens spécifique relatif à un groupe de personnes, autrement dit à un *domaine* donné. Cette définition se rapproche de celle des modèles de thèmes, qui tentent eux de découvrir les thèmes inhérents à une collection de documents. Le même contexte pragmatique pour la terminologie et pour les modèles de thèmes nous mène à penser que ces deux tâches peuvent bénificier l'une de l'autre :

Les modèles de thèmes pour la terminographie : Si le sens d'un terme est rattaché à un domaine précis, déterminer le domaine du document dans lequel sa graphie apparaît doit permettre de la désambiguïser s'il y a polysémie. Si le document est reconnu comme appartenant à la catégorie *médecine*, il faut distinguer la graphie *réseau de neurones* de la même graphie dans un document de *sciences de l'information*. Pour ces raisons, les modèles de thèmes peuvent améliorer la qualité et l'organisation des termes extraits par un système automatique. Nous nous concentrerons cependant exclusivement sur le lien entre terminographie et modèles de thèmes.

La terminographie pour les modèles des thèmes : Les expériences présentées dans la section 4 qui portent sur les modèles de thèmes ne traitent que de cet aspect. Nous avons vu en section 3.1.1 que les modèles de thèmes s'appuient sur des calculs de fréquences du vocabulaire pour représenter les documents. Il est probable que compléter cette représentation avec des lexies plus complexes issues d'une extraction terminographique pourrait bénéficier à la qualité du modèle construit. Si une terminologie est vue comme un facteur discriminant entre des domaines, il y a fort à penser que le processus d'identification de ces domaines puisse tirer parti de ces graphies terminologiques.

Nous postulons que la construction des modèles de thèmes peut bénéficier de connaissances terminologiques exprimées sous forme de calculs de fréquences, comme pour les mots simples. L'ajout de graphie de plusieurs mots pose cependant la question de la méthode de calcul des fréquences, comme l'illustre la table 3.1.

TABLE 3.1 – Trois méthodes de calcul de fréquences pour une phrase donnée : *J'entraîne un réseau de neurones.* La colonne *Standard* correspond à la représentation classique de la phrase. La colonne *Cas 1* correspond aux fréquences minorées par celles des graphies plus complexes. La colonne *Cas 2* correspond à une forme de surcharge de fréquences : les mots composant les graphies complexes sont en quelque sorte pris en compte à deux reprises.

Graphie	Standard	Cas 1	Cas 2
j'	1	1	1
entraîne	1	1	1
un	1	1	1
réseau	1	0	1
de	1	0	1
neurones	1	0	1
réseau de neurones	0	1	1

Pour la colonne *Standard*, *F(réseau de neurones)=0* car aucun traitement n'a permis de l'identifier comme une unité complexe pertinente. La fréquence de ses composants est en revanche de 1. Le tableau 3.1 illustre la différence obtenue par l'incorporation de graphies de plusieurs mots si l'on compare les colonnes *Cas 1* et *Cas 2*. Nous pouvons observer que dans la colonne *Cas 1* la fréquence des mots composant la graphie complexe est de 0. A contrario, dans la colonne *Cas 2* la fréquence de ses composants est de 1, rapprochant cette méthode de calcul de la méthode standard. Pour les expériences que nous présentons dans la section 4 nous avons préféré la méthode illustrée dans *Cas 2* : elle nous permettra simplement d'*augmenter* la méthode standard avec des graphies plus complexes, nous permettant une comparaison directe entre les deux propositions – entre *Standard* et *Cas 2*. L'introduction de connaissances terminographiques pour la construction de modèles de thèmes peut se faire directement ; nous pouvons identifier les étapes qui suivent pour une méthode d'extraction terminographique donnée :

1. Construire le vocabulaire – graphies simples – utilisé dans le corpus,
2. Extraire les graphies terminologiques de deux mots et plus,
3. Représenter chaque document du corpus comme une combinaison de graphies simples et complexes,

4. Construire le modèle de thèmes.

Nous postulons qu'une comparaison entre la méthode *Standard* et la méthode *Cas 2* aboutira à des modèles de meilleure qualité pour la seconde, pour les raisons évoquées ci-dessus. Le déroulement de cette expérience ainsi que ses résultats sont présentés en section 4.1. Nous verrons également des possibles étapes supplémentaires à celles énumérées ci-dessus, notamment certaines qui se sont avérées non-concluantes. Nous avons justifié du lien entre la construction de modèles de thèmes et analyse terminologique en général ; nous avons également présenté la méthode d'hybridation entre terminologie et construction de modèles. Nous présentons maintenant les problématiques liées à l'évaluation des modèles de thèmes ainsi que la méthode que nous emploierons en section 4.1.

3.1.3 Evaluation des modèles de thèmes

Nous l'avons brièvement évoqué dans l'introduction, l'évaluation des modèles de thèmes est une problématique en soit : dans un contexte non supervisé et flou, les évaluations passent généralement par des métriques qui estiment la cohérence du modèle construit indépendamment de toute vérité terrain. L'implémentation de la mesure de la *cohérence* varie selon les chercheurs ; elle peut être autonome [Lau et al., 2014, Wallach et al., 2009], liée à des ressources externes [Newman et al., 2009, Newman et al., 2010], mais elle est généralement probabiliste. Les recherches de [Wallach et al., 2009] sont intéressantes en ce que les chercheurs appréhendent l'évaluation d'une manière similaire à la notre, à savoir comme une tâche de classification : leur évaluation passe par l'estimation de la probabilité d'une bonne assignation pour un nouveau document. Ces métriques sont affranchies de vérité terrain, comme induit par le contexte de modèle de thèmes.

Definition 3.1.3.1. Mesure externe – Méthode de calcul de la qualité d'un modèle basée sur des ressources externes : dictionnaires, corpora annotés, bases de connaissances, etc. Les mesures externes sont systématiquement employées dans les contextes supervisés : *F-mesure*, précision, rappel, Indice de Rand [Rand, 1971], Information Mutuelle et ses successeurs, etc.

Definition 3.1.3.2. Mesure interne – Méthode de calcul de la qualité d'un modèle basée uniquement sur les données placées en entrée. Les mesures internes sont généralement employées pour l'évaluation des algorithmes

non supervisés [Liu et al., 2010] : indice de Dunn [Dunn, 1973], silhouette [Rousseeuw, 1987], etc.

Les mesures de qualité internes des modèles de thèmes sont pertinentes dans la mesure où elles permettent de discriminer les algorithmes de construction sans corpus de référence :

Soient alg_1 et alg_2 deux algorithmes distincts de construction de modèles de thèmes,
Soit un corpus non-annoté $\mathcal{C}$,
Et soit la métrique interne de qualité/cohérence m,
La comparaison de $m(algo_1)$ avec $m(algo_2)$ sur $\mathcal{C}$ permet de mettre en évidence le meilleur algorithme de construction relativement à la métrique et au corpus – les résultats varient en fonction de ces deux paramètres.

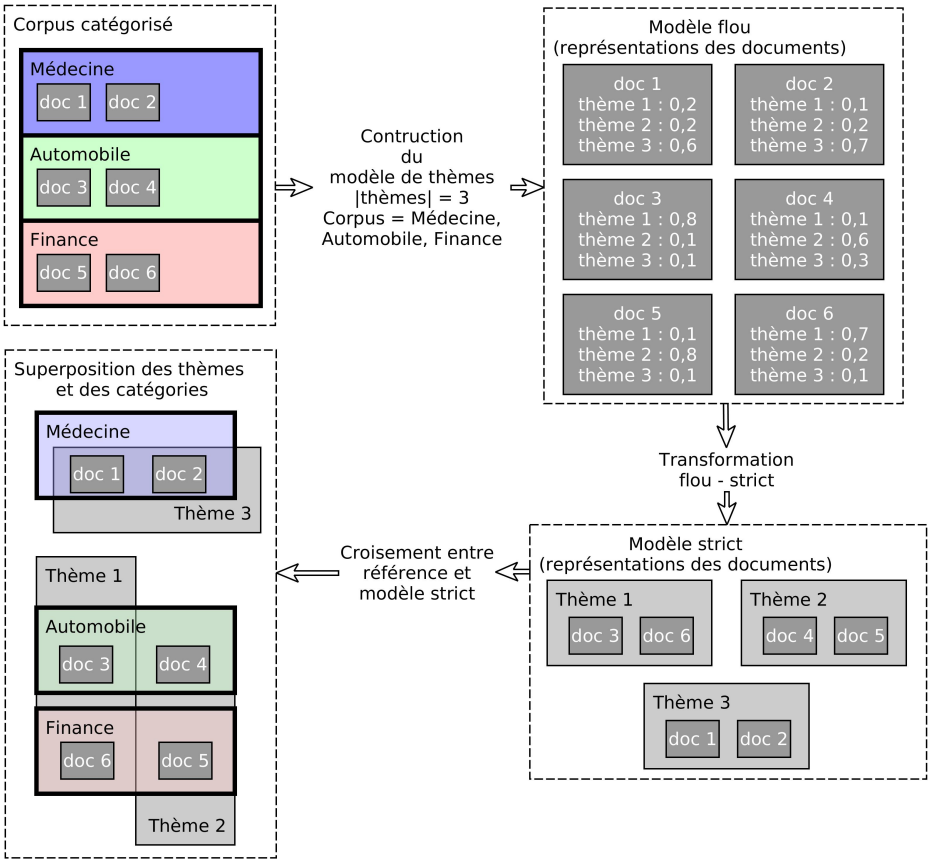
Elles sont en revanche moins pertinentes lorsque comparées aux résultats attendus par un expert humain. Par définition, les mesures internes s'affranchissent de la notion de vérité terrain ; c'est pourtant cette vérité terrain qui permet d'évaluer la qualité d'un algorithme relativement à un objectif attendu. Autrement dit, une mesure interne ne peut apporter une garantie de la qualité des résultats ou même de leur pertinence, c'est pouquoi nous parlons essentiellement de *cohérence*. A contrario, les mesures externes permettent de comparer un modèle au résultat attendu selon une référence, ce qui correspond davantage à l'idée d'une évaluation d'une détection automatique de thèmes.

Evaluer la qualité d'un modèle de thèmes par un algorithme probabiliste non-déterministe

Evaluer la qualité d'un modèle de thèmes calculé par un algorithme par rapport à une référence catégorisée strictement se rapproche d'une évaluation de classification. La transformation d'un modèle de thèmes flou vers un modèle strict passe par la sélection.

se rapproche d'une tâche de classification : déterminer automatiquement le thème *majoritaire* correspondant à un document – *majoritaire* car les documents sont décrits sous forme d'une distribution de poids associés aux différents thèmes. La transformation du modèle de thèmes construit vers un modèle *strict* – i.e. des associations strictes *document/thème* – présente cependant certains écueils afin de procéder à une évaluation complète. La problématique principale porte sur l'alignement entre les thèmes construits et les thèmes/catégories de la référence, comme nous allons maintenant le

FIGURE 3.3 – Illustration de la problématique posée par l'évaluation d'un modèle de thèmes flou relativement à une référence.



développer.

La figure 3.3 illustre la problématique que nous rencontrons lorsque nous tentons d'évaluer un modèle de thèmes comme un résultat de classification. Nous donnons quelques précisions sur la nature des éléments du diagramme :

Corpus catégorisé : Le corpus placé en entrée. Seul le nombre de catégories – en plus des documents – est exploité par l'algorithme de construction du modèle de thèmes. Les catégories en tant que partitions seront utilisées ultérieurement. La notation *Corpus = Médecine, Automobile, Finance* signifie que l'ensemble des documents de chaque catégorie est passé à l'algorithme de construction de modèle de

thèmes ; la connaissance de la nature de la partition n'est bien entendu pas transmise à l'algorithme.

Modèle flou : Nous l'avons vu, les modèles de thèmes consistent en des distributions de probabilités sur des thèmes, autrement dit des modèles de classification flous. Nous ne donnons que la représentation des documents dans les thèmes construits car la représentation des thèmes ne nous est pas utile ici – les thèmes sont représentés sous forme de distributions de probabilités sur un vocabulaire limité. Cette représentation nous permet ensuite de dériver des association strictes entre document et thème.

Modèle strict : Contrairement au *modèle flou*, les documents ne sont plus décrits par une distribution de thèmes ; ils sont à la place associés à un thème unique. Le résultat de cette transformation doit ensuite être aligné avec les catégories de la référence à des fins d'évaluation.

Superposition des thèmes et des catégories : Comme évoqué plus haut, l'alignement entre les *thèmes stricts* construits et les catégories de la référence est la principale problématique. Comme nous pouvons le voir dans la figure 3.3, l'alignement entre le *thème* 3 et la catégorie *médecine* n'est pas ambigu et peut se faire simplement : *médecine* contient les documents *doc* 1 et *doc* 2 et le *thème* 3 ne contient que ces deux documents. En revanche, il n'en est pas de même pour les autres thèmes construits : le *thème* 1 et le *thème* 2 se partagent de manière équilibrée les documents des catégories *automobile* et $finance$.

L'exemple de la figure 3.3 se veut simple, n'est pas représentatif de la problématique rencontrée mais l'illustre convenablement : le partage équilibré entre le *thème* 1 et le *thème* 2 avec les catégories *automobile* et $finance$ n'aurait a priori pas d'influence sur la métrique d'évaluation. Ce cas illustre cependant un phénomène que nous rencontrons avec davantage de catégories/thèmes et de documents : l'équilibre parfait présenté dans la figure 3.3 n'arrive quasiment jamais en pratique, notamment du fait de l'augmentation du nombre de documents/thèmes, mais l'alignement doit répondre à un impératif simple : maximiser la qualité du modèle construit en associant les bons thèmes aux bonnes catégories. La problématique revient à maximiser la qualité moyenne à partir d'un alignement catégorie/thème optimal.

3.1.3.1 Alignement entre thèmes et catégories d'un corpus de référence

Pour un corpus $\mathcal{C}$ composé de k catégories cat_1 à cat_k,
Pour k thèmes construits et transformés (stricts) th_1 à th_k,
Pour une mesure de qualité ensembliste m,
Pour $m(cat_i, th_{\sigma(i)})$ où σ est une permutation sur $[1, i]$,
Le choix des associations entre un thème th_i et une catégorie cat_i doit maximiser la moyenne des scores de l'ensemble des associations :

$$max_\sigma \left(\frac{\sum_{i=0}^{k} m\left(cat_i, th_i\right)}{k} \right) \tag{3.1}$$

Nous présentons ci-dessous un algorithme d'alignement déterministe qui vise à retenir la permutation σ qui permet de trouver le maximum. L'algorithme est voisin de l'algorithme dit *hongrois* : alors que l'algorithme hongrois élimine les valeurs faibles pour remonter vers les valeurs fortes, nous procédons inversement. La complexité de l'algorithme proposé est également moindre relativement à celle de l'algorithme Hongrois : $O(k^3)$ au lieu de $O(k^4)$, bien que la complexité des algorithmes d'alignement n'entre que peu en compte avec des valeurs de k aussi petites. Le pseudocode 1 détaille notre manière de procéder.

La logique de l'algorithme peut être développée comme suit :

1. Pour chaque thème, repérer de quelle catégorie il est le plus proche (*F-mesure* maximum),
2. S'il n'y a pas d'autre thème avec une similarité supérieure, l'association thème-catégorie est validée et est ajoutée au résultat – sinon le thème est ignoré jusqu'à la prochaine itération,
3. S'il reste des éléments à aligner (i.e. la taille de *result* est inférieure au nombre de catégories), alors itération.

3.1.3.2 Evaluation de la pertinence d'un modèle de thèmes relativement à une vérité terrain

Une fois les thèmes construits alignés avec leurs catégories respectives, nous pouvons attribuer un score global au modèle de thèmes construit à partir de la moyenne des scores de chaque thème.

Algorithm 1 Algorithme d'alignement

Input : `cats` les `k` catégories de la référence, `ths` les `k` thèmes transformés calculés, m(ths[i],cats[j]) une mesure de similarité ensembliste

Output : `result` un ensemble de `k` couples catégorie-thème

```
1: cm : matrix[k][k]
2: for i in 1, k do
3:     for j in 1, k do
4:         cm[i][j] ← m(ths[i],cats[j])
5:     end for
6: end for
7: result ← ∅
8: while size(result) < k do
9:     for i in 1, size(ths) do
10:        jmax ← getArgMax(cm[i][])
11:        if getMaxSim(cm[][jmax]) == cm[i][jmax] then
12:            result.add(<ths[i], cats[jmax]>)
13:            ths.remove(clusters[i])
14:            cats.remove(cats[jmax])
15:            cm[i][].remove()
16:            cm[][jmax].remove()
17:        end if
18:    end for
19: end while
```

Soit un modèle de thèmes $\mathcal{M}$ constitué des thèmes $th_1 \dots th_k$
Soient $cat_1 \dots cat_k$ les catégories du corpus analysé,
Soit $align(th_i)$ une fonction retournant une catégorie avec une *F-mesure* maximisée à partir d'un thème,
Soit F_1 une fonction retournant la *F-mesure* calculée entre un thème et une catégorie
Le score global du modèle se définit comme :

$$score(\mathcal{M}) = \frac{\sum_{i=1}^{k} F_1(th_i, align(th_i))}{k}$$

Autrement dit, le score du modèle consiste en la moyenne des meilleures *F-mesures* après alignement strict, i.e. une catégorie n'est associée qu'à un seul thème et réciproquement. Avant que ce score ne nous permette de comparer différents modèles, il nous reste cependant à répondre à une dernière problématique : le non-déterminisme de l'algorithme de construction.

Definition 3.1.3.3. Déterminisme [1] – Principe scientifique d'après lequel tout phénomène est régi par une (ou plusieurs) loi(s) nécessaire(s) telle(s) que les mêmes causes entraînent dans les mêmes conditions ou circonstances, les mêmes effets.

Dans le cadre de la construction de modèle de thèmes, le non-déterminisme signifie que plusieurs exécutions d'un même algorithme sur un corpus peuvent produire des résultats différents. Le non-déterminisme est d'autant plus observable avec l'augmentation de la complexité de la tâche : plus le nombre de thèmes augmente, plus les écarts entre les différentes exécutions peuvent être importants. Il en va de même selon la nature des documents à analyser, nous y reviendrons en détail dans la section 4.1. Le non-déterminisme est particulièrement problématique en ce qu'il induit qu'aucune exécution n'est réprésentative des paramètres de l'expérience : les variations importantes pour des paramètres donnés – corpus, k, etc. – suggèrent qu'aucune exécution prise isolément ne peut permettre d'estimer la qualité d'un modèle. Afin de parvenir à comparer les différents modèles construits, nous proposons de répéter plusieurs fois chaque construction de modèle de thèmes afin d'obtenir non pas une seule *F-mesure*, mais un ensemble de *F-mesures* qui nous permettra de visualiser clairement les résultats, notamment au travers de boîtes à moustaches/diagrammes de Tukey. Ces diagrammes devraient nous

1. `https://cnrtl.fr/definition/d%C3%A9terminisme`

permettre d'observer des différences de variations notables relativement à la complexité de la tâche, en plus de nous permettre de mieux appréhender la qualité du modèle construit avec les écarts-types, médianes et autres mesures statistiques.

Nous venons de décrire le protocole permettant d'évaluer un modèle de thèmes construit relativement à un corpus de référence et en palliant le non-déterminisme de l'algorithme de construction. Nous avons montré que l'évaluation passera davantage par une analyse de diagrammes de type boîtes à moustaches, qui nous permettrons des analyses plus fines que de simples scores. Nous exploiterons néanmoins les scores en eux-mêmes – et non plus les diagrammes – pour l'axe de recherche que nous présentons ensuite.

3.1.4 Corrélations entre modèles de thèmes et vocabulaires

Nous venons de présenter notre méthode d'évaluation d'un modèle de thèmes relativement à un corpus catégorisé de référence. Lors de manipulations avec des algorithmes de construction de modèles de thèmes sur différents corpora, nous avons pu constater des différences notables selon les exécutions : selon la nature des documents, selon les thèmes à identifier, selon le nombre de thèmes, selon la méthode de représentation des documents, etc. Nous l'avons vu précédemment, la construction des modèles de thèmes s'appuie sur le vocabulaire utilisé dans les documents du corpus, or il existe de nombreuses mesures de similarité textuelle qui permettent de comparer des documents. L'axe de recherche présenté ici porte sur la nature du lien entre qualité du modèle et mesures de similarité.

Definition 3.1.4.1. Mesure de similarité/distance textuelle – Une mesure de similarité est une fonction symétrique prenant deux chaînes de caractères en paramètres et produisant un score dit de similarité. Par extension, une mesure de similarité peut également être appliquée à un ensemble (catégorie, thème) : l'ensemble est alors représenté comme un seul document issu de la concaténation des documents contenus dans cet ensemble.

Le calcul de similarités inter-catégorielles – entre deux catégories du corpus – devrait permettre de mettre en évidence un lien direct entre similarité et qualité des modèles : la similarité entre la catégorie *Médecine* et la catégorie *Automobile* pourrait être corrélée à la qualité du modèle construit.

Observer une telle corrélation entre qualité et similarité pourrait servir différents objectifs :

- Dans un contexte supervisé, i.e. à partir d'un corpus de référence, le calcul de la similarité entre les catégories du corpus à analyser permettrait d'estimer a priori la qualité du modèle construit : si la similarité entre les deux catégories/thèmes à identifier est élevée, la qualité du modèle sera probablement mauvaise, et réciproquement. Un rapide calcul de similarité devrait donc permettre d'estimer la pertinence d'un calcul plus complexe – celui de la construction du modèle de thèmes correspondant.
- Dans un contexte non supervisé, le calcul de la similarité entre les thèmes construits constituerait une mesure d'évaluation en elle-même : des similarités basses entre les thèmes indiqueraient des thèmes effectivement distincts, des similarités hautes seraient indicatrices d'une frontière floue entre thèmes. En non supervisé, la similarité pourrait également être utilisée comme une mesure de convergence – par exemple fixer un seuil comme critère d'arrêt.

Notre axe de recherche porte également sur une comparaison avec et sans connaissance terminographique : de même que nous avons modifié les représentations utilisées par les modèles de thèmes, nous comparerons les différentes corrélations avec et sans connaissance terminographique. Pour les mêmes raisons qui nous poussent à penser que les modèles avec connaissance terminographique seront de meilleure qualité, nous tentons de montrer l'intérêt de ces connaissances au niveau de la similarité textuelle. De fait, nous comptons observer des corrélations plus élevées à partir de représentations de documents plus complètes/complexes.

De nombreuses mesures ont été proposées dans la littérature, basée sur de nombreux postulats, nous les verrons plus en détail dans la section 3.1.4 ; nous en présenterons notamment une typologie. Nous y développerons également le protocole expérimental relatif aux calculs de corrélations entre similarité et qualité ainsi que nos résultats sur de nombreuses mesures de similarité. Nous avons vu dans la section précédente (section 3.1.3) que l'évaluation des modèles construits passe par l'analyse de diagrammes. Les mesures de corrélations seront quant à elles calculées à partir de *F-mesures* moyennes.

Soit $\mathcal{M}$ un modèle de thèmes,
Soit $score(\mathcal{M})$ la moyenne des *F-mesures* entre catégories et thèmes issus de l'alignement,

$score_{avg} = \frac{\sum_{x=1}^{y} score(\mathcal{M})}{y}$ avec $\mathcal{M}$ reconstruit à chaque itération.

La valeur de y est à fixer en fonction de la nature des documents utilisés et des paramètres de l'expérience ; nous avons pour notre part pu constater que des valeurs comprises entre 10 et 20 sont suffisantes pour obtenir une visualisation pertinente de la qualité du modèle. Pour rappel, ces itérations sont nécessaires pour pallier le non-déterminisme de l'algorithme de construction. C'est donc $score_{avg}$ qui nous permettra de mesurer les différentes corrélations entre similarité textuelle et performance lors de l'identification des thèmes.

3.2 Structure des termes candidats : de l'intérêt d'un élargissement des motifs

De nombreuses études portent sur la nature des éléments constitutifs des graphies des termes. Par *éléments constitutifs* nous entendons les mots qui constituent les graphies, de même que les informations – linguistiques, syntaxiques, lexicologiques – s'y rapportant. Nous avons décrit en section 2 quelques tendances relatives à la détection automatique de graphies terminologiques que nous résumons ici :

— Une terminologie consiste en une structure organisée de *termes*, i.e. de concepts spécifiques à un domaine manifestés par des graphies terminologiques.

— Une limitation générale aux groupes pseudo-nominaux – *pseudo* car l'absence de déterminants parmi les éléments constitutifs ne permet pas de satisfaire les contraintes structurelles des groupes nominaux. Les séquences d'étiquettes morphosyntaxiques concernées varient fortement dans la littérature, se détachant plus ou moins du concept central de groupe nominal. Ainsi certains chercheurs ont tenté de décrire exhaustivement les séquences constitutives de groupes nominaux [Frantzi et al., 1998, Justeson and Katz, 1995] à partir de combinaisons de noms communs, adjectifs et groupes prépositionnels. L'exhaustivité est néanmoins relative : alors que le patron morphosyntaxique de [Justeson and Katz, 1995] semble le plus complet, certains aspects des groupes nominaux sont omis tels que les propositions subordonnées relatives ou certaines formes verbales fléchies que nous verrons plus loin. D'autres chercheurs ont quant à eux proposé des frontières

plus souples, autorisant par exemple la reconnaissance d'une graphie à partir de la conjonction de plusieurs groupes nominaux, comme nous l'avons vu dans la section 2.

— Une forme de consensus sur la nécessité et la complexité du traitement des structures verbales. Peu de recherches ont été effectuées sur la détection/l'intérêt des structures verbales en extraction terminographique malgré l'intérêt indéniable des groupes verbaux pour la terminologie. Le peu de travaux sur le sujet s'explique par deux facteurs : premièrement, l'extraction de groupes nominaux présente encore des écueils ; ensuite la gestion des structures verbales est plus complexe que celle des structures nominales. Cette difficulté accrue entre la gestion des groupes nominaux et verbaux s'explique par le niveau d'analyse nécessaire à la description d'une graphie : alors qu'un groupe nominal est sémantiquement autonome, l'interpétation d'un groupe verbal peut s'appuyer sur sa valence, ses actants, ainsi que la nature des liens qu'il entretient avec ses actants. Les structures adjectivales – de même que prépositionnelles – sont quant à elle considérées comme des satellites de groupes nominaux, et ne nécessitent donc pas de traitement à part entière.

— Une évaluation complexe et coûteuse. Nous l'avons vu, l'évaluation d'une extraction terminographique est généralement effectuée par des *experts* du même domaine que le corpus dont est issu l'extraction. L'évaluation manuelle présente divers écueils inhérents à l'analyse terminologique et aux frontières floues de ses concepts. Outre le fait qu'il soit nécessaire de disposer de plusieurs experts *du domaine* pour évaluer l'extraction, les limites de ce qui constitue (ou pas) une graphie terminologique diffèrent d'un expert à l'autre. Nous pouvons par exemple observer une absence totale de structures prépositionnelles dans un corpus annoté pour la terminographie (cf. section 4.2.2), alors que ces dernières ont au préalable été retenues comme *pertinentes* à ces fins [Justeson and Katz, 1995]. L'évaluation d'une extraction terminographique pose donc la nécessité d'une formalisation des strutures acceptables en amont, formalisation qui se retrouve tant chez les experts – pour la constitution de ressources annotées et/ou l'évaluation d'extractions – que dans les systèmes automatisés (voir section 4.2.5). Les méthodes d'évaluation utilisées aujourd'hui ne permettent pas d'analyser de manière exhaustive la distribution des graphies effectivement terminologiques dans la liste triée produite, comme nous

y reviendrons en section 3.3.4.
Nous présentons dans cette partie nos hypothèses expérimentales basées sur ces constats ainsi que sur nos observations préliminaires sur différents corpora non-annotés. Nous présentons dans un premier temps nos postulats sur les groupes nominaux, avant de poursuivre avec les autres structures. Les *autres structures* – verbales, adjectivales, prépositionnelles, etc.– seront discutées mais peu développées ; nos expériences portent toutes sur les structures nominales.

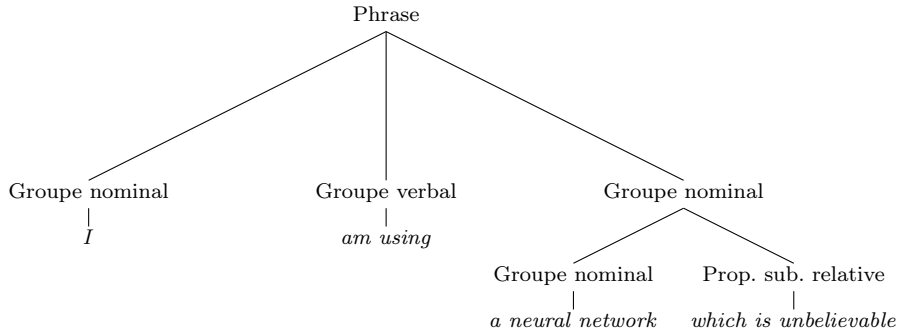
3.2.1 Structures nominales

3.2.1.1 Distinction entre termes candidats et groupes nominaux

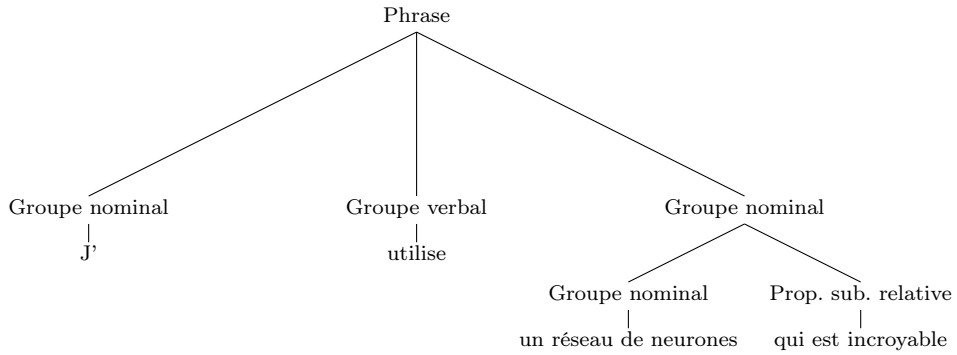
Les travaux de [Justeson and Katz, 1995] et ceux qui ont suivi corroborent l'importance des groupes nominaux pour l'extraction terminographique. S'il y a consensus sur le rôle essentiel occupé par les groupes nominaux dans l'extraction terminographique, il n'en est pas de même concernant les propriétés de ces groupes. La qualification *nominal* de la fonction d'un groupe syntaxique n'est pas ambiguë sur le plan linguistique : la qualification n'est pas issue d'un jugement ou d'une interprétation, mais de faits linguistiques concrets. Dans un contexte d'extraction terminographique, le substantif *groupe nominal* est entendu dans un sens plus restreint, certaines structures ne pouvant apparaître dans un terme candidat. C'est par exemple le cas des propositions subordonnées relatives, satellites d'un groupe nominal, ou des pronoms, groupes nominaux en eux-mêmes, mais sans perspective de lexicalisation – en anglais comme en français. De ce fait, aucune recherche portant sur l'extraction automatique de termes n'inclut ces éléments.

Exemple :
I'm using a neural network which is unbelievable.
J'utilise un réseau de neurones qui est incroyable.

L'exemple ci-dessus illustre la frontière entre les notions de groupes nominaux et de termes candidats. Une analyse en groupes syntaxiques produit :



Et :



Nous n'identifions ici que les groupes de *haut niveau*, les analyses plus fines seront évoquées et exploitées en section 4.1.3.1. Cet exemple nous permet néanmoins d'observer dans les faits la frontière évoquée. Bien que groupes nominaux, les mots *I* et *Je* ne sont pas reconnus comme correspondant à de potentiels termes : nous avons pu voir en section 2 qu'aucun des patrons morphosyntaxiques proposés dans la littérature ne permet l'extraction de pronoms, non significatifs en dehors de la connaissance de la chaîne de coréférence. Leur utilisation dans un texte pose cependant une problématique stylistique relative à la propension de son auteur à les utiliser : la modification du décompte des fréquences de graphies induite par l'utilisation de pronoms impacte les calculs de potentiels terminologiques (voir section 2.2.5).

La non reconnaissance des propositions subordonnées relatives se justifie par la variabilité de la langue et la tendance de l'auteur à paraphraser afin d'éviter les répétitions. Nous l'avons vu en section 2.2.4, la paraphrase est un phénomène linguistique problématique et récurrent qui concerne tous les décomptes de fréquences, dont ceux effectués à des fins d'extraction terminographique. Les formes canoniques des deux exemples ci-dessus seraient *I am using an unbelievable neural network* et *J'utilise un incroyable réseau de neu-*

rones. Contrairement aux pronoms, les propositions subordonnées relatives véhiculent du sens compréhensible hors contexte ; néanmoins les nombreuses variantes correspondant à une même graphie en font des phénomènes générateurs de bruit dans les systèmes automatisés.

3.2.1.2 Rôle des entitées nommées dans la construction de syntagmes terminologiques

La majorité des travaux en extraction terminographique automatique exclut les entités nommées des graphies acceptées. Les travaux – sur l'anglais – de [Park et al., 2002] sont notables en ce que les chercheurs proposent un patron morphosyntaxique qui inclut des éléments jusque là ignorés : les conjonctions de coordinations et les entités nommées. L'introduction des entités nommées pose la question de la délimitation entre graphies terminologiques et entités nommées, notamment lorsque la dite introduction est faite à plusieurs niveaux : en tant que satellite du mot final et/ou en tant que mot final. La permissivité du Patron 4 induit la reconnaissance de l'ensemble des entités nommées comme potentiellement terminologiques : les premiers éléments du Patron 4 sont optionnels, il peut donc reconnaître une entité nommée unique (NNP ou NNPS), ainsi qu'une succession d'entités nommées. La question posée peut donc se résumer à : est-ce qu'une entité nommée seule peut être terminologique ?

Après quelques observations effectuées dans [**Delamaire, Amaury** et al., 2019a] sur des corpora en anglais, nous suggérons que la reconnaissance des entités nommées dans les graphies terminologiques est effectivement nécessaire, bien que dans une moindre mesure. Autoriser la reconnaissance d'unigrammes et de séquences d'entités nommées comme graphies pertinentes nous semble trop permissif, aussi proposons-nous de n'en permettre la reconnaissance que comme satellite d'un nom commun final, induisant de fait une limitation aux graphies de deux mots et plus. L'importance des entités nommées dépend fortement de la thématique dont le corpus est issu ; comme nous le verrons plus en détail en section 4.2, très peu de graphies contenant une entité nommée sont extraites à partir d'un corpus de sciences des données. A contrario, dans les mathématiques, de nombreuses entités – fonctions, méthodes, représentations, formules, etc. – portent le nom de leur découvreur. Ainsi pouvons nous voir apparaître les graphies suivantes parmi les graphies extraites :

— *Ising model*,

— *Gibbs measure*,
— *Schrödinger equation*,
— *Euler equation*,
— *Cauchy problem*,
— *nonlinear Schrodinger equation*,
— *Navier-Stokes equation*

Le choix de l'une ou l'autre solution est également lié au choix de méthode de validation des graphies extraites. Dans le cas d'une validation par analyse contrastive, il y a tout lieu de penser que l'extraction permise par le Patron 4 induise un bruit dont l'importance est relative à la thématique (cf. *mathématiques* vs *sciences des données* ci-dessus). Le Patron 4 ajouterait les graphies suivantes à la l'énumération ci-dessus : *Ising, Gibbs, Schrödinger, Euler, Cauchy* et *Navier-Stokes*. Leur fréquence d'apparition dans le corpus mathématique sont i) importance relativement à un corpus de langue générique, et ii) supérieures aux fréquences des graphies composées qui les contiennent. De ce fait, une analyse contrastive considérerait les entités *Ising, Gibbs, Schrödinger, Euler* et *Cauchy* comme terminologiques et leur donnerait l'ascendant sur les graphies composées véritablement terminologiques qui ont été énumérées.

La *NC-valeur* [Frantzi et al., 1998] ne peut pas non plus être appliquée sur une extraction issue du Patron 4. Comme nous le verrons en détail en section 4.2.1, le calcul repose sur une estimation de la lexicalisation d'une graphie, estimation qui passe par une analyse des termes composés reprenant la dite graphie. L'inclusion d'unigrammes entités nommées produirait les mêmes effets que pour l'analyse contrastive, à savoir un ascendant des graphies simples sur les graphies composées.

Pour pallier ce problème, [Park et al., 2002] s'appuient sur une combinaison de mesures permettant de calculer le potentiel terminologique d'une graphie (voir section 2.3.1.1) au travers d'estimations de deux aspects linguistiques – analyse constrastive de graphies composées, cohésion de graphies. Ils rencontrent cependant le même problème évoqué ci-dessus :

> The equation 3 produces much higher values for single-word terms than multi-words terms because the association of a single-word term only depends on its frequency.

[Park et al., 2002].

La solution qu'ils proposent semble ad-hoc et difficilement généralisable,

au vu des différences de langage selon les thématiques :

> Thus, we reduce the scale of association of single-words terms by taking only a fraction of the value (for example, 10%.)

.

Aucune des 45 graphies avec le potentiel terminologique le plus élevé ne contient d'entité nommée, et de nombreuses graphies composées d'un mot unique ne semblent pas particulièrement terminologiques : *vehicle, equip, lamp, collision, engine, mode, etc.*

Ces observations nous amènent à ne considérer que les graphies de plusieurs mots dans nos expériences à venir, ainsi qu'à ne considérer les entités nommées que comme des satellites de noms communs.

3.2.2 Structures verbales et adjectivales

L'étiquetage morphosyntaxique est une étape nécessaire à la très vaste majorité des systèmes automatiques de reconnaissance de termes, notamment pour l'application des patrons éponymes. Comme tout système automatique, et plus particulièrement appliqué au langage, les étiqueteurs ne sont pas infaillibles – voir section 2.3.1.1.1. Leurs performances dépendent grandement de la nature du langage analysé : des textes issus de courriels, SMS ou de commentaires sur les réseaux sociaux seront particulièrement difficiles à analyser correctement pour des outils entraînés sur le langage *académique*. De nombreux mots seront considérés OOV (*Out Of Vocabulary*, *hors vocabulaire connu*), de nouveaux éléments extra-linguistiques font leur apparition (*smileys*, émoticônes) ; le rôle habituel de la ponctuation est remis en question.

Il en va de même pour les documents issus d'OCR (Optical Character Recognition) ou de transformations depuis un PDF, deux procédés qui opposent des écueils aux outils de TAL : confusion entre les caractères, non-reconnaissance d'espace, mauvaise gestion des tableaux et formules, etc. L'exploitation de corpora issus d'un de ces procédés nécessite la mise en place de filtres en amont de tout traitement.

Ces écueils sont spécifiques en ce qu'ils concernent des anomalies dans la matière analysée. Nous avons pu observer plusieurs phénomènes de *confusion morphosyntaxique* lors de nos expériences sur un corpus anglais dans [**Delamaire, Amaury** et al., 2019a], plus spécifiquement sur l'étiquetage des formes verbales et adjectivales. La confusion porte sur la distinction entre

deux paires d'étiquettes :

VBG & JJ : entre gérondifs et adjectifs,

VBN & JJ : entre participes passés et adjectifs.

Prenons l'exemple : *The algorithm has symmetrised these measures*. Dans ce contexte, il n'y a pas d'ambiguïté sur l'étiquette à apposer à *symmetrised* : c'est un participe passé (VBN) qui suit l'auxiliaire *to have*. A contrario, dans *The algorithm has symmetrised measures*, la détermination de l'étiquette est plus complexe. Un humain avec quelques connaissances de grammaires identifie *symmetrised* comme un adjectif ; les étiqueteurs peuvent quant à eux être induits en erreur par la présence de l'auxiliaire *to have* juste avant *symmetrised*. La même confusion apparaît entre les gérondifs et les adjectifs : *The algorithm was filtering equations*, confusion ici induite par l'ambiguïté de *was*, pouvant être l'auxiliaire du gérondif *filtering* ou bien signifier "consists in".

A partir de ces observations sur l'anglais, nous posons l'hypothèse qu'étendre le concept de qualificatif aux participes passés et formes gérondives pourrait bénéficier aux systèmes d'extraction terminographique, ces derniers exploitant de manière quasi-systématique les adjectifs qualificatifs dans les graphies extraites.

3.2.3 Autres structures

Parmi les structures syntaxiques sur lesquelles nous n'expérimenterons pas figurent les groupes verbaux. Si de précédentes recherches les ont inclus dans leurs analyses [Roche et al., 2004, L'Homme, 2012], le traitement des groupes verbaux reste particulièrement complexe. Nous l'avons évoqué en introduction de la section 3.2, l'analyse d'un terme verbal passe nécessairement par l'analyse de sa valence et de ses actants.

Definition 3.2.3.1. Actant – Groupe nominal dont la présence est obligatoire qui entretient une relation sémantique et syntaxique avec un mot donné.
Ils entraînent un réseau de neurones
<actant$_1$ = Ils> entraînent <actant$_2$ = un réseau de neurones>

Definition 3.2.3.2. Valence – Nombre d'actants nécessaire à la création d'un syntagme grammaticalement correct à partir d'un mot donné.
Ils entraînent un réseau de neurones
*<actant$_1$ = Ils> entraînent <actant$_2$ = ?>.
*<actant$_1$ = ?> entraînent <actant$_2$ = un réseau de neurones>

<actant$_1$ = Ils> entraînent <actant$_2$ = un réseau de neurones>
valence(entraîner)= 2

Dans l'exemple utilisé dans les définitions ci-dessus, l'extraction du terme *entraîner* et l'estimation de son potentiel terminologique passent par la description de ses actants : l'aspect terminologique du verbe repose sur un changement de nature de ses arguments canoniques ; le complément d'objet passe d'un être animé à un inanimé – plus précisément d'un humain/animal à un algorithme. Cette distinction de nature entre les actants ((in-)animé, (in-)dénombrable, etc.) correspond à un niveau d'analyse du TAL assez élevé, source de nombreuses erreurs – au même titre que l'ACI, cf. section 2.3.1.1.1. Nous avons vu précedemment que les quelques erreurs des étiqueteurs morphosyntaxiques peuvent être problématiques pour la reconnaissance de certaines graphies, problématique accrue avec l'ajout de traitements produisant davantage d'erreurs dans la chaîne – l'automatisation de l'abstraction des actants est hasardeuse relativement aux performances des étiqueteurs morphosyntaxiques. A la complexité que nous venons de décrire peuvent également s'ajouter certains écueils linguistiques qui génèrent des variantes syntaxiques.

La pronominalisation des verbes peut poser des problèmes de reconnaissance de variantes graphiques pour un verbe donné : la pronominalisation réduit dans les faits la valence du verbe de 1 en induisant que le sujet joue également le rôle d'objet, même si le rôle du deuxième actant est toujours satisfait.

Exemple : *Le réseau de neurones s'entraîne facilement.*
<actant$_1$ = le réseau de neurones> <actant$_2$ = s'> entraîne facilement.
valence(s'entraîner)= 2

Malgré le fait que le sens véhiculé est le même pour les deux occurrences d'*entraîner*, elles ne peuvent être reconnues comme similaires du fait de la différence de valence. La polysémie (2.2.1), et plus particulièrement la terminologisation, s'identifie parfois à partir de valences différentes pour un terme donné ; aussi faire le choix de considérer les deux occurrences d'*entraîner* comme équivalentes provoquerait une perte de finesse de l'analyse. Nous l'avons vu dans l'exemple, la pronominalisation ne réduit pas la valence effective du verbe avec le *s'* remplissant le rôle de premier actant, mais reconnaître le *s'* comme un actant ne fait que décaler le problème vers la résolution de chaînes de coréférences, dont l'automatisation n'est pas encore satisfaisante

pour qu'une extraction terminographique automatique s'y appuie.

Enfin, l'extraction des verbes terminologiques associés à leurs actants respectifs pose la question de l'évaluation. Nous avons déjà pu constater qu'il n'y a que peu de consensus relatif à l'évaluation des terminologies nominales (cf. 2.4) alors que les groupes nominaux terminologiques ne nécessitent pas d'informations en dehors de leur(s) graphie(s). La prise en compte de nouveaux facteurs dans l'évaluation ne rend la tâche que plus complexe.

3.3 Méthode d'évaluation d'une extraction terminographique

Nous avons pu voir en section 2 que de nombreuses méthodes d'évaluation d'une extraction terminographique ont été proposées. La méthode d'évaluation d'une terminologie est une problématique ouverte. La faible quantité de ressources dédiées à disposition constitue un frein important à l'automatisation de l'évaluation. Bien que de nombreuses recherches s'affranchissent d'une évaluation mathématique stricte, des solutions ont été proposées, notamment via l'annotation d'experts. A partir de ces annotations, [Frantzi et al., 1998] estiment les ratios des véritables termes en tête d'une liste triée de termes candidats relativement à la queue de la liste. De manière similaire, les auteurs du corpus ACL-RD-TEC [Handschuh and QasemiZadeh, 2014] estiment la proportion de termes pertinents en tête de liste. Ces mesures donnent une estimation de la capacité du système à mettre en avant de "vrais" termes relativement aux non-termes.

Bien que permettant une comparaison objective entre différents systèmes, cette méthode d'évaluation ne permet pas de tirer de conclusions plus précises que des tendances en début et fin de liste triée. L'exploitation de ressources dédiées va nous permettre de développer cette méthode. L'exploitation de terminologies préconstruites pose cependant le même problème que l'évaluation par des experts, à savoir la subjectivité de ses auteurs et leur propension à accepter certaines séquences morphosyntaxiques plutôt que d'autres. Pour le corpus ACL-RD-TEC, le biais morphosyntaxique émane essentiellement du choix du patron par les chercheurs, patron qui semble trop contraint – nous le développerons dans le chapitre 4.

Ce problème d'évaluation des résultats est important lors de l'élaboration des méthodes, du choix des paramètres et de leur implémentation. Des

mesures d'évaluation très simples et bien connues peuvent directement s'appliquer. Nous présenterons les mesures de précision et de rappel et surtout nous transposerons d'autres mesures connues dans le domaine de la Recherche d'Information (RI) pour le cadre de l'extraction des termes.

Nous proposons une méthode d'évaluation inspirée de la recherche d'informations. La structure du corpus ACL-RD-TEC nous permet d'appliquer les métriques d'évaluation classiques en recherche d'information. Une extraction terminographique automatique est généralement évaluée par sa propension à faire remonter des éléments effectivement terminologiques dans une liste triée selon une certaine métrique. Cette propension est généralement estimée au travers de ratios de termes effectifs calculés sur des échantillons de la liste produite. Cette restriction s'explique par l'évaluation manuelle, contrainte par le besoin d'experts et de temps. Cette probabilité, bien que permettant de comparer les systèmes, ne permet pas d'analyser en intégralité la distribution des vrais termes dans la liste triée proposée. Pour cette raison, et grâce au corpus ACL-RD-TEC, nous avons pu appliquer les métriques de recherche d'information proposées dans les conférences TREC, pionnières dans la systématisation de l'évaluation des systèmes de recherche d'information [Schütze et al., 2008]. Celles-ci nous ont permis de compléter les informations préliminaires fournies par l'analyse d'échantillons.

Cette transition d'une évaluation pour systèmes de RI à une évaluation d'extraction terminographique est rendue possible par une analogie entre le triptyque *documents, besoins d'informations, pertinence* considéré en RI et celui de l'extraction terminographique *terme candidat, sous-corpus, réalité terminologique*.

1. *Documents/terme candidat* désigne les éléments à évaluer. En RI, il s'agit de documents. En extraction terminographique, de termes candidats.
2. En RI, les besoins d'information sont exprimés sous forme de requêtes ; la pertinence des documents est ensuite estimée relativement à cette requête. En extraction terminographique, nous n'exploitons pas des requêtes mais directement des corpora : nous évaluons une extraction qui se veut la plus exhaustive possible : la requête consiste donc en un corpus.
3. La pertinence consiste en l'évaluation d'un élément relativement aux besoins d'information exprimés : en RI, elle revient à déterminer si le document considéré est pertinent relativement aux besoins d'informa-

tions. En extraction terminograpique, elle revient à signifier l'aspect terminographique d'un terme candidat relativement à un corpus – i.e. le besoin d'informations exprimé.

Parmi les métriques proposées dans le logiciel TREC_EVAL [2], nous exploitons particulièrement les mesures de précision relatives au rappel. Ces dernières nous permettent de construire des représentations visuelles explicites de la distribution des vrais termes dans la liste construite – relativement au corpus ACL-RD-TEC. La *F-mesure*, théoriquement pertinente, ne peut être exploitée ici : l'absence de besoin d'experts nous permet de ne pas réduire la taille de la liste de sortie, qui est de fait bien supérieure au nombre de termes effectifs à extraire. En conséquence, la précision de même que la *F-mesure* sont extrêmement faibles. Les mesures de précision relatives au rappel sont cependant particulièrement explicites, comme nous le verrons en section 4.

3.3.1 Précision et rappel

Soit $\mathcal{C}$ un corpus de documents dont tous les termes ont été annotés manuellement et soit $T(\mathcal{C})$ l'ensemble de ces termes. Un système d'extraction terminographique automatique produit un ensemble de termes candidats, notons le $CT(\mathcal{C})$. Idéalement un système d'extraction automatique produirait tous les termes présents et uniquement ceux-ci, à savoir :

$$CT(\mathcal{C}) = T(\mathcal{C})$$

Si nous traitons uniquement l'ensemble $CT(\mathcal{C})$, deux mesures de qualité du résultat peuvent se mettre en œuvre, la **précision** et le **rappel** [3] :

$$Precision(CT) = \frac{card(CT \cap T)}{card(CT)}$$

$$Rappel(CT) = \frac{card(CT \cap T)}{card(T)}$$

La précision mesure la capacité du système à extraire uniquement des termes et le rappel la capacité à retrouver la globalité de termes.

Idéalement, $Precision(CT) = 1$ et $Rappel(CT) = 1$ lorsque $CT = T(\mathcal{C})$.

Deux cas extrêmes :

2. `https://trec.nist.gov/trec_eval/`

3. Afin de simplifier l'écriture on renonce à mettre en évidence le corpus de travail et on utilise simplement les notations T et CT.

— si $CT_0 = \emptyset$, alors, par extension de la définition, $Precision(CT_0) = 1$ et, par calcul, $Rappel(CT_0) = 0$
— si CT_∞ est constitué de tous les mots et toutes les successions de 2, 3, 4 ou 5 mots du corpus, alors $Rappel(CT_\infty) = 1$, mais $Precision(CT_\infty) \approx 0$.

Toutefois, ces deux mesures manquent de finesse et ne tiennent pas compte du fait que le système d'extraction automatique fournit souvent un candidat terme *ct* avec un score $score(ct)$ de potentiel terminologique. Si un candidat avec un faible score n'est pas un terme, l'impact est moindre relativement à un faux candidat avec un score élevé.

3.3.2 Précision relative au rappel

Une mesure qui vient du domaine de la recherche d'information peut être appliquée aussi dans l'évaluation plus fine des performances. Elle vise à mesurer les performances de la réponse à une requête pour un système de recherche d'information, sous l'hypothèse que les résultats – documents – rendus ont aussi un score – *rank* – par rapport à la requête posée.

La mesure de précision relative au rappel va pallier ce défaut d'évaluation de résultat sur la base de deux valeurs synthétiques et elle prend donc les valeurs du score de potentiel terminologique exprimé sous la forme de couples $(ct, score(ct))$.

Considérons que l'ensemble CT est trié selon les scores des candidats en ordre décroissant. Soit $CT|\, i$ l'ensemble formé par les i premiers termes de CT avec i entre 0 et $card(CT)$. Il est aisé de calculer $p_i = Precision(CT/i)$ et $r_i = Rappel(CT/i)$. Nous pouvons remarquer que $p_0 = 1$, $r_0 = 0$ et aussi que, pour toute valeur de i, p_i et r_i sont des valeurs comprises entre 0 et 1. Nous pouvons également prouver facilement par récurrence selon i que : $r_i \leq r_{i+1}$. Il n'y a en revanche pas de monotonie pour p_i.

La valeur p_i est aussi appelée précision au rang i.

La **précision relative au rappel** est définie comme la courbe (r_i, p_i), pour $i = 0, card(CT)$. La représentation graphique de cette courbe est souvent en dents de scie car p_i n'est pas toujours décroissante et sa représentation par ses valeurs interpolées est préférée, avec :

$$Precision_{interpolee}(\rho) = \max_{\{i; r_i \geq \rho\}} (p_i)$$

L'analyse de la performance de systèmes différents sur un même corpus mène à la considération des précisions interpolées à 11 points de rappels

$((\rho_j)_{0\leq j\leq 10}$ avec $\rho_j = j/10)$ et la moyenne des précisions interpolées est calculée.

Idéalement, si $CT = T$, alors $p_i = 1$ et $r_i = \frac{i}{card(T)}$ pour tout i entre 0 et $card(T)$ et la représentation graphique est un segment parallèle à l'axe des abscisses.

3.3.3 Précision moyenne

Si la représentation graphique pour deux cas différents est très ressemblante, nous pouvons mettre en place l'indicateur synthétique de précision moyenne comme la moyenne des précisions relatives au niveau des termes correctement extraits. Si nous considèrons pour chaque terme candidat ct ayant un $score(ct)$ sa valeur de vérité de l'extraction $Verite(ct) \in \{0, 1\}$, nous pouvons définir la **précision moyenne** comme la moyenne arithmétique des précisions aux indices ayant des termes candidats correctement extraits :

$$AP = \frac{1}{card(\{i|Verite(ct_i) = 1\})} \sum_{i;Verite(ct_i)=1} p_i$$

Cette valeur peut également se calculer sur des valeurs de la précision interpolée :

$$AP_{interpollee} = moyenne(Precision_{interpollee}(\rho))$$

avec ρ les valeurs de points d'interpolation.

3.3.4 Précision moyenne globale

En recherche d'information, la performance d'un système est estimée sur plusieurs requêtes et la mesure de précision moyenne globale *(*MAP - Mean Avarage Precision*)* est définie comme la moyenne des précisions moyennes de chaque requête.

Ce concept est transposable pour un système d'extraction automatique de terminologie dans le cas où le système est évalué sur k corpus disjoints $\mathcal{C}_j$, $j = 1, k$. Pour chaque corpus $\mathcal{C}_j$ nous disposons de $T_j = T(\mathcal{C}_j)$ l'ensemble de ses termes, et le système d'extraction obtient les termes candidats $CT_j = CT(\mathcal{C}_j)$; donc nous avons k courbes précision rappel (r_{ij}, p_{ij}), $j = 1, k$ et k valeurs de la précision moyenne.

La précision moyenne globale se définit comme la moyenne des précisions moyennes sur chaque corpus :

$$MAP = \frac{1}{k} \sum_{j} AP(\mathcal{C}_j)$$

$$MAP = \frac{1}{k} \sum_{j=1}^{j=k} \frac{1}{card(\{i|Verite(ct_{ij}) = 1\})} \sum_{i=1, Verite(ct_{ij})=1}^{card(CT_j)} p_{ij}$$

Cette unique valeur synthétique permet de comparer des systèmes dont les courbes de précision/rappel se superposent visuellement, tout comme la valeur plus simple de précision moyenne quand elle est utilisé pour l'évaluation des systèmes différents sur un même corpus.

Chapitre 4

Nos expériences

Après avoir introduit l'extraction terminographique (chapitre 1), présenté ses problématiques (chapitre 2) et introduit nos hypothèses d'analyse (chapitre 3), nous présentons ici les expériences entreprises. La construction automatique de terminologies étant un domaine de recherche récent et complexe sur le plan linguistique, de nombreuses expériences sont possibles à différents niveaux d'abstraction et de traitement. Par niveau d'abstraction, nous signifions le niveau de linguistique théorique incorporée au processus de construction ; par niveaux de traitement nous désignons les différentes étapes de la construction automatique préalablement présentées – extraction, score, sélection. Nos expériences portant exclusivement sur l'automatisation du processus, nous avons choisi un angle d'évaluation proche de la *recherche d'information*. La complexité et la longueur des tâches de construction de terminologies automatique justifient également ce choix – constructions de corpus annoté et évaluations des résultats. Ce choix s'illustre par l'adoption de méthodes d'évaluation très directes et détachées de la nature des données, comme nous le verrons en partie 4.2. Cette partie est divisée en trois sous-parties qui présentent chacune une expérience entreprise en lien avec nos observations dans la partie 3. Nous présentons dans un premier temps nos tentatives d'hybridation entre modèle de thèmes et extraction terminographique en partie 4.1. Nous développons ensuite une méthode d'évaluation d'un algorithme de construction de modèle de thèmes non supervisé à partir d'un corpus de référence. Enfin, nous présentons nos résultats et méthodes d'évaluation sur une extraction terminographique automatique complète en partie 4.2.

4.1 Croisement entre extraction terminologique et modèles de thèmes

Comme précédemment évoqué en partie 2.2, l'extraction et l'organisation automatiques de syntagmes terminologiques sollicitent diverses problématiques linguistiques, parmi lesquelles la polysémie. C'est à ce phénomène que nous nous intéressons dans cette partie. D'après l'hypothèse exposée en partie 3.1, nous présentons une expérience portant sur une hybridation entre contruction de modèles de thèmes et extraction terminographique. L'objectif de cette hybridation peut être double :

1. Améliorer la qualité des modèles construits grâce à l'incorporation d'informations terminographiques,
2. Désambiguiser des termes polysémiques.

Notre expérience porte sur le premier élément mentionné ci-dessus, à savoir estimer l'intérêt de l'exploitation des termes candidats dans la construction de modèles de thèmes. Nous présentons les résultats de deux méthodes d'extraction terminographiques – ACI et patrons – et de deux modifications de l'espace de représentation – fréquences et *NC-valeur*. Nous développons également les écueils pratiques qui ont limité nos expériences.

4.1.1 Un corpus commun

Ces expériences ont été menées sur un unique corpus anglais catégorisé que nous avons construit. Nous avons sélectionné un sous-ensemble du corpus Isearch [Lykke et al., 2010], que nous avons restreint selon les paramètres suivants :

Nombre de documents par catégorie : De 750 à 1 000 documents par catégorie, selon les disponibilités du corpus,

Nombre de catégories : 4 catégories divisées en 4 à 5 sous-catégories,

Nature : Nous n'avons exploité que les résumés des documents, nous nous en expliquons ci-après,

Contraintes : Un document doit satisfaire une contrainte linguistique important : contenir a minima une phrase grammaticalement correcte. Nous avons identifié les *phrases correctes* à partir de l'ACI, en détectant l'étiquette correspondante dans l'arbre construit – P.

Nous avons retenu les catégories et sous catégories suivantes pour nos expériences :

Mathematics : Probability, analysis of PDEs, quantum algebra, differential geometry, algebraic geometry,

Physics : General physics, optics, atomic physics, chemical physics, fluid dynamics,

High Energy Physics : Experiment, lattice, phenomenology, theory,

Condensed Matter : Statistical mechanics, strongly correlated electrons, materials science, superconductivity, mesoscale and nanoscale physics.

Les catégories énumérées ci-dessus correspondent aux étiquettes apposées manuellement sur les articles scientifiques du corpus Isearch. Nous avons retenu cet extrait d'Isearch pour nos expériences après avoir rencontré différents écueils avec d'autres ressources. L'ACI est une technique du TAL complexe et coûteuse en temps de traitement, qui plus est si elle est appliquée à du texte informel – fora, SMS, commentaires, etc. Du fait de l'exploitation que nous faisons de l'ACI, nous n'avons pas pu compléter nos expériences sur le corpus 20 NEWSGROUP sans appliquer de filtres en amont pour éliminer les segments de textes problématiques. Ce problème s'est également manifesté dans le corps des articles du corpus Isearch, mais pour une autre raison : de nombreux documents sont issus d'OCR (*Optical Character Recognition*), or cette reconnaissance produit de nombreuses erreurs. Qu'il s'agisse d'erreurs de segmentation (mauvaise reconnaissance des limites des mots) ou d'erreurs de reconnaissance de caractères, l'application d'une ACI automatique devient impossible. Le corpus de résumés que nous avons construit à partir d'Isearch n'est cependant pas exempt d'erreurs de transcription, même si elles sont moins fréquentes. Nous avons identifié deux éléments récurrents sources d'erreurs : les tableaux et les formules mathématiques. Nous avons alors imposé une contrainte de taille minimale qui nous a permis de limiter les exécutions de l'ACI. Un minimum de cinq mots – cinq séquences de caractères non blancs – nous a permis de diminuer drastiquement le temps de traitement et de limiter les erreurs – informatiques – de l'ACI automatique liées aux segments asyntaxiques. Il s'agit de la seule contrainte en amont de l'ACI dans nos expériences.

Exemple de document de la catégorie mathématiques : *Discrete stationary classical processes as well as quantum lattice states are asymptotically confined to their respective typical support, the exponential growth rate of*

which is given by the (maximal ergodic) entropy. In the iid case the distinguishability of typical supports can be asymptotically specified by means of the relative entropy, according to Sanov's theorem. We give an extension to the correlated case, referring to the newly introduced class of HP-states.

Après filtres, seules 3 des 19 sous-catégories de notre corpus ont moins de 1 000 documents :

1. *algebraic geometry* avec 791 documents,
2. *analysis of PDEs* avec 923 documents,
3. *strongly correlated electrons* avec 955 documents.

Le corpus comprend donc 18 669 documents et contient un peu plus de deux millions de mots, ce qui donne une moyenne de 110 mots par document.

4.1.2 Le modèle de thèmes choisi

Nous avons dans un premier temps expérimenté avec CRP – *Chinese Restaurant Process* [Aldous, 1985] – qui est un modèle de thèmes strict non-déterministe de la famille de LDA [Blei et al., 2003]. CRP développe une métaphore filée qui associe l'image d'un restaurant à la détection de thèmes. Considérons une salle de restaurant de taille infinie et avec un nombre de tables potentiellement infini où, lorsqu'un client entre, ce dernier peut choisir entre s'asseoir seul – sur une *nouvelle* table – ou à une table occupée par d'autres personnes. Le choix de la personne dépend de ses préférences pour les plats qui sont sur les différentes tables. Dans le contexte de la détection de thèmes, les clients sont des documents, les tables des thèmes, les plats le vocabulaire. L'affectation d'un document à un thème se fait via un système de multinomiales basé sur le vocabulaire du thème – le vocabulaire du thème est défini par ses documents, i.e. les clients assis à la table. Nous rappelons que nous disposons d'un corpus strictement catégorisé décrit en section 4.1.1 qui doit nous permettre d'effectuer une évaluation du modèle construit. Si CRP est adéquat de par l'assignation stricte d'un document à un thème, les deux formes de son non-déterminisme rendent son évaluation à partir d'une référence particulièrement complexe, si ce n'est impossible. La première forme de non-déterminisme concerne l'affectation d'un document à un thème : la décision est basée sur le résultat d'une mixture de multinomiales, elle peut donc varier d'une exécution à la suivante. L'autre forme du non-déterminisme est plus problématique voire bloquante : dans CRP, les systèmes de multinomiales peuvent aboutir à la décision de créer un nouveau

thème ; le nombre de thèmes du modèle final ne peut donc être déterminé a priori. Un contrôle relatif est possible à partir des paramètres α et β qui vont inciter CRP à créer plus ou moins de thèmes. Deux cas extrêmes sont identifiables : la création d'un seul thème auquel appartiennent tous les documents et la création d'autant de thèmes que de documents. L'objectif revient à optimiser les paramètres α et β pour obtenir un nombre de thèmes adéquat, mais sans garantie de résultat. Nos multiples expérimentations avec CRP n'ayant pas abouti à des modèles que nous pouvions évaluer, nous en avons alors choisi un autre de la même famille mais déterministe quant au nombre de thèmes construits.

Nous nous proposons d'effectuer une étude comparative à partir du modèle de thèmes LDA (Latent Dirichlet Allocation, [Blei et al., 2003]) en exploitant les résultats de l'ACI. De multiples exécutions avec et sans modification du texte nous permettront de comparer les versions et d'aboutir à une conclusion.

Une grande variété de méthodes proposent des solutions de classification, reposant sur des hypothèses diverses. Nous avons retenu LDA car c'est le modèle qui a produit les meilleurs résultats sur notre corpus de documents succincts et hautement spécialisés. LDA est un algorithme de *classification non supervisée* statistique, testé et évalué par de nombreux chercheurs. L'algorithme tente de décrire des thèmes dont le nombre est prédéfini à partir de la distribution du vocabulaire au sein des documents. Les thèmes identifiés par LDA sont définis par extension et non par intension, c'est-à-dire qu'ils sont uniquement définis par les éléments qui les constituent (contrairement à une étiquette comme *Mathematics*). Pour cela, LDA décrit les documents comme des combinaisons des thèmes qu'il a identifiés ; l'hypothèse étant que les documents sont générés à partir des distributions de vocabulaires correspondant à leurs thèmes respectifs. LDA pose l'hypothèse du sac-de-mots, à savoir que l'ordre et la nature des mots ne sont pas pris en compte, contrairement à ce que nous proposons à travers l'ACI. Nous précisons ici le vocabulaire de la classification automatique pour y situer LDA. En termes de méthodes, il faut distinguer :

classification supervisée : Un *corpus d'entraînement* est fourni à l'algorithme afin d'*apprendre* à prédire la catégorie à laquelle un nouvel élément appartient. La classification supervisée nécessite un corpus d'entraînement, ce qui peut être un obstacle à son utilisation car le corpus est généralement constitué manuellement et validé par plu-

sieurs personnes afin d'en assurer la qualité.

classification non supervisée : L'algorithme crée des groupes (*clusters*) à partir d'informations extraites du corpus à analyser, ce sans corpus d'apprentissage mais avec un ensemble de règles statistiques ou de mesures de similarité.

En termes de résultats d'algorithmes statistiques, il faut distinguer les sorties :

dures/strictes : l'algorithme fournit une réponse binaire relative à l'appartenance d'un élément à une classe,

floues/non-strictes : l'algorithme fournit une probabilité ou un score relatif à l'appartenance d'un élément à une classe.

Le choix d'un algorithme non supervisé comme algorithme de classification permet de se passer de corpus d'entraînement et donc de pouvoir traiter des documents de natures variées. Avec un algorithme supervisé, il est nécessaire de fournir un corpus d'entraînement par thème : le corpus permettant d'apprendre à distinguer le domaine du *médical* de celui de l'*agriculture* ne permet pas de distinguer immédiatement la *physique* de la *géographie*.

La construction du modèle passe par des représentations de documents sous forme de tableaux de nombres ; le tableau de nombres représentant un document donne la fréquence de chaque terme du vocabulaire dans ce document : c'est l'hypothèse du sac-de-mots. La sortie de LDA est :

1. un ensemble de k thèmes, chaque thème étant une distribution de probabilités sur l'ensemble des éléments constitutifs, qui peut donc être représentée sous forme d'un vecteur/tableau de $|\mathcal{V}|$ nombres, $\mathcal{V}$ désignant l'ensemble du vocabulaire,
2. la représentation des documents comme distributions de probabilités sur l'ensemble des k thèmes : un document est donc représenté sous la forme d'un vecteur de k nombres.

Exemple de sortie :

thème/vocabulaire	mot_1	mot_2	...	$\text{mot}_{\lvert\mathcal{V}\rvert}$
thème_1	0,002	0,002	...	0,002
thème_2	0,003	0,003	...	0,003
...	...	...	...	...
thème_k	0,000	0,000	...	0,000

document/thème	thème_1	thème_2	...	thème_k
document_1	0,1	0,8	...	0,0
document_2	0,1	0,1	...	0,7
...	...	...	...	...
$\text{document}_{\lvert\mathcal{D}\rvert}$	0,6	0,2	...	0,05

LDA pose l'hypothèse que les documents à analyser sont générés par des modèles statistiques correspondant à des thèmes distincts. Ces thèmes correspondent à des variables dites *latentes* et l'objectif de l'algorithme revient à construire les modèles statistiques présupposés. Les modèles relatifs aux différents thèmes consistent en des distributions du vocabulaire. Pour cela, LDA construit des modèles de cooccurrences évolués et estime la similarité entre ces modèles. Il est à noter que LDA ne classifie pas strictement les documents mais fournit un vecteur d'associations thème - poids au sein duquel les similarités entre un document et l'ensemble des thèmes sont représentées. La construction du modèle LDA est influencée par deux hyperparamètres α et β fixés a priori. α permet d'augmenter ou diminuer la taille de l'intersection entre les vocabulaires respectifs des thèmes construits. De manière similaire, β permet d'augmenter ou diminuer le nombre de thèmes par document. Plusieurs recherches ont porté sur une paramétrisation optimale de α et β (références), mais comme nous le verrons en partie 3.1.3, notre expérience n'est pas impactée par ces paramètres.

Des chercheurs ont déjà obtenu une amélioration des modèles de thèmes avec un ajout de termes plus complexes, mais se basant soit sur des thesauri [Nokel and Loukachevitch, 2016], soit sur des modèles statistiques de n-grammes et d'alignement de n-grammes [Wang et al., 2007, Blei and Lafferty, 2009].

Dans la continuité de ces expériences, nous tentons ici d'appliquer d'autres méthodes de traitement automatisé du langage telles que l'ACI ou l'extraction via des patrons morphosyntaxiques, cette dernière étant une méthode large-

ment utilisée pour l'extraction de termes candidats. De plus, nous présentons une étude comparative des mesures de similarité textuelle relativement à leur espace de représentation et à la qualité des modèles construits.

4.1.3 Exploitation de formes surfaciques correspondant à des termes candidats dans un modèle de thèmes

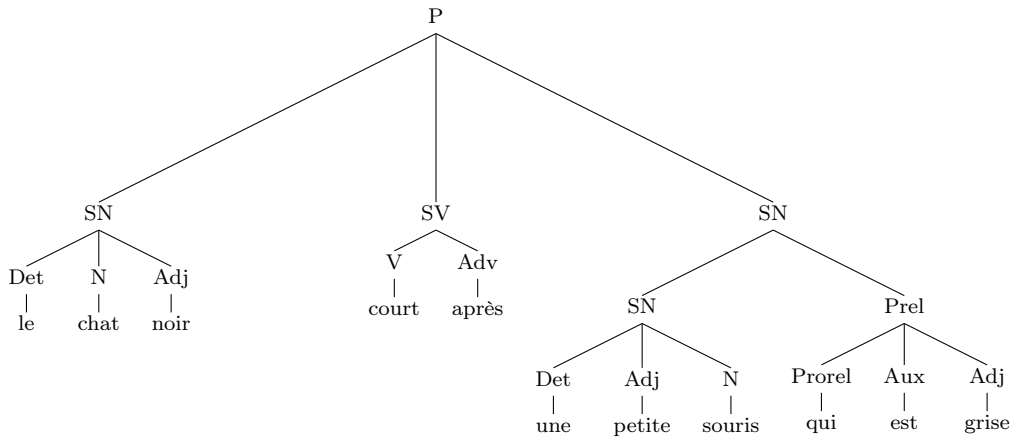
Cette première expérience porte sur l'incorporation de formes plus complexes que les mots simples dans l'espace de représentation lors de la contruction de modèles de thèmes. Là où ne sont généralement pris en compte que les mots simples, nous expérimentons l'introduction d'éléments plus complexes, issus de traitements linguistiques. Nous présentons nos résultats à partir de deux méthodes d'extraction terminographique :

1. ACI : Nous avons extrait les groupes nominaux et verbaux construits à partir d'une Analyse en Constituants Immédiats. Nous avons réalisé plusieurs constructions de modèles de thèmes pour expérimenter les différentes combinaisons : groupes uniquement nominaux, uniquement verbaux, groupes verbaux et nominaux.
2. Patrons : Nous avons extrait les termes candidats correspondant à plusieurs patrons morphosyntaxiques. Contrairement à l'ACI, nous n'avons pas exploité les structures verbales dans cette partie. En revanche, nous présentons des résultats sur quatre patrons morphosyntaxiques – dont trois de la littérature.

Nous développons en premier lieu nos résultats à partir de l'ACI en page 107 avant de présenter nos résultats avec les patrons morphosyntaxiques en page **??**. Nous présentons ensuite les deux écueils principaux que nous avons pu rencontrer lors de nos expérimentations : une dimensionnalité excessive (page 111) et un non-déterminisme significatif (page 113).

4.1.3.1 Exploitation de syntagmes complexes issus de l'ACI

L'ACI peut se définir comme l'identification des groupes fonctionnels constitutifs d'une phrase (verbaux, adjectivaux, prépositionnels, etc.). L'arbre suivant illustre ce processus de délimitation et de hiérarchisation opéré lors d'une ACI.



Où : SN signifie Syntagme Nominal, SV Syntagme Verbal, PREL Proposition subordonnée RELative. A partir de cette analyse nous pouvons extraire les groupes souhaités pour nos expériences, en l'occurrence les groupes verbaux et nominaux. Il est à noter que d'autres analyses valables de la phrase peuvent être retenues : la proposition faite ici n'a qu'une valeur d'exemple à des fins d'illustration.

Nous nous proposons d'augmenter la dimensionnalité de l'espace de représentation de chaque document en y incorporant les syntagmes identifiés. Considérant le fait que l'algorithme LDA fonctionne sur le principe du *sac-de-mots*, il nous suffit d'intégrer des syntagmes issus des résultats de l'ACI dans le texte original. La question de la méthode d'intégration reste cependant ouverte. Comme l'illustre l'exemple d'ACI ci-dessus, l'analyse produit une hiérarchie de séquences de mots, parfois avec des récursions de fonctions. Le syntagme nominal *une petite souris qui est grise* a été identifié, de même que le sous-syntagme nominal *une petite souris*. Nous avons choisi d'ajouter à l'espace de représentation les deux syntagmes nominaux repérés : ne récupérer que le plus *bas* dans l'arbre provoquerait un important silence sur des formes plus complexes potentiellement interessantes, et inversement avec le plus *haut* dans l'arbre qui risque d'être trop spécifique. Pour ces raisons, nous retenons l'ensemble de la hiérarchie des syntagmes nominaux et verbaux lors de l'analyse, au risque d'augmenter à l'excès la dimensionnalité de l'espace de représentation des documents. L'augmentation de la dimensionnalité reste cependant acceptable en appliquant certains filtres génériques, notamment en éliminant les hapax. Sur notre corpus décrit en partie 4.1.1, l'introduction brute de syntagmes augmente l'espace de représentation par un facteur 14 relativement au corpus sans traitement : l'espace de représentation du corpus

standard est de dimension 37 000, celui avec syntagmes de 511 000. Après élimination des hapax, le facteur d'augmentation passe à 3,4 avec 22 600 dimensions pour le corpus standard et 77 500 avec les syntagmes. Le tableau suivant présente l'ensemble des tailles de vocabulaire pour les différentes expériences que nous avons menées, par catégorie.

		LDA		LDA-L		LDA-LS		LDA-LSF	
Catégorie	$\|classes\|$	$\|\mathcal{V}\|$	$\|hapax\|$	$\|\mathcal{V}\|$	$\|hapax\|$	$\|\mathcal{V}\|$	$\|hapax\|$	$\|\mathcal{V}\|$	$\|hapax\|$
Physics	2 classes	14 503	7 181	12 235	6 325	72 982	60 188	71 503	59 785
	3 classes	17 619	8 972	15 212	8 093	103 259	85 730	101 226	85 273
	4 classes	21 807	11 214	19 352	10 439	140 250	116 900	137 536	116 339
	5 classes	24 945	12 996	22 271	12 223	176 745	148 259	173 957	147 982
Condensed Matter	2 classes	12 341	6 009	10 737	5 475	72 152	59 152	70 671	58 857
	3 classes	16 847	8 465	15 161	8 001	108 478	89 952	105 952	89 288
	4 classes	19 563	10 180	17 883	9 724	137 719	114 831	134 701	114 140
	5 classes	21 752	11 636	19 972	11 194	168 283	141 080	165 352	140 722
Mathematics	2 classes	10 725	5 305	9 341	4 815	57 501	46 672	37 760	29 404
	3 classes	13 871	7 125	12 305	6 652	82 955	68 295	45 573	36 359
	4 classes	16 492	8 669	14 851	8 210	108 729	90 202	72 430	59 665
	5 classes	18 464	9 925	16 777	9 515	128 228	106 887	78 507	65 295
H. E. Physics	2 classes	10 843	5 487	9 539	5 005	50 447	40 622	48 221	39 796
	3 classes	13 871	7 220	12 288	6 731	78 222	64 411	75 704	63 626
	4 classes	16 884	8 974	15 133	8 533	106 955	89 103	104 060	88 317

TABLE 4.1 – Evolution de la taille du vocabulaire ($|\mathcal{V}|$) et du nombre d'hapax ($|hapax|$) pour toutes les catégories.

Les étiquettes des colonnes désignent les différents prétraitements effectués avant l'exécution de LDA :

LDA : Aucun prétraitement sur le corpus,

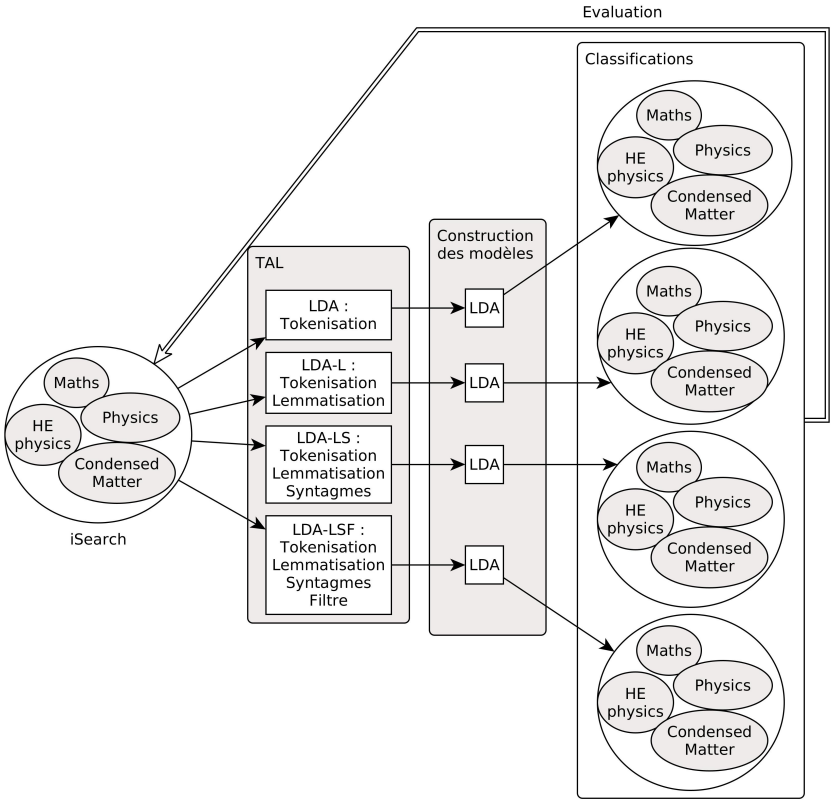
LDA-L : Le corpus est lemmatisé,

LDA-LS : Le corpus est lemmatisé et les syntagmes extraits des documents selon la méthode décrite ci-dessus,

LDA-LSF : Le même traitement que LDA-LS est appliqué, mais une contrainte est appliquée aux mots simples – non-extraits par l'ACI. Seuls sont retenus les noms communs, les verbes et les adjectifs en plus des syntagmes verbaux, nominaux et adjectivaux.

La figure 4.1 présente l'organisation générale des expériences présentées dans cette partie :

FIGURE 4.1 – Organisation générale des expériences



Pour transformer les distributions de probabilités que produit LDA en classification stricte, nous avons implémenté une méthode correspondant à notre idée évoquée en partie **??**.

4.1.3.2 Problématique de la taille croissante du vocabulaire

L'automatisation d'un processus de traitement des données requiert une estimation de la quantité de données à traiter. Nous l'avons vu en section 2, les algorithmes de TAL et de fouille de données s'appuient sur des hypothèses mathématiques diverses – cooccurrences, processus de Dirichlet, lois normales, manipulations de matrices, etc. Ces hypothèses sont affectées de manière variable par la *dimensionnalité* de l'espace de représentation. Nous avons identifié trois problématiques liées à l'espace de représentation relativement à l'exploitation de formes complexes pour la construction de modèles de thèmes.

hapax : Nous avons pu observer une proportion d'hapax de l'ordre de 50% environ sur un extrait du corpus iSearch [Lykke et al., 2010] sans traitement linguistique ; autrement dit la moitié des mots du vocabulaire utilisé ne le sont qu'une seule fois. L'intuition voudrait qu'ajouter des procédés de conflation de la variation – racinisation, lemmatisation, etc. – permettrait de réduire cette proportion. Nos observations sur le même extrait de corpus tendent cependant à montrer le contraire : la proportion d'hapax augmente systématiquement entre 3 et 8% après l'application d'une lemmatisation [Manning et al., 2014a]. Ce phénomène peut s'expliquer par une réduction de la taille du vocabulaire – 10% en moyenne. Cette réduction du vocabulaire augmente la proportion d'hapax, bien que moins nombreux que sans lemmatisation – 5% de moins en moyenne. Les hapax restants posent la question de leur exploitation pour la construction de modèles de thèmes : en tant qu'hapax, ils ne sont pas utiles à la construction des modèles en synchronie. Ils pourront cependant être utiles pour l'extraction terminographique – identification de termes à partir de relations entretenues entre autres avec des hapax – ou pour une évolution en diachronie – enrichissement de l'espace de représentation, i.e. du vocabulaire, avec l'apport progressif de nouveaux documents.

dimensionnalité : Indépendamment de la question des hapax se pose celle de la dimensionnalité. Nous l'avons vu, les modèles de thèmes

s'appuient sur des représentations multidimensionnelles où chaque dimension correspond à un élément du vocabulaire. Certaines techniques de réductions matricielles ont été proposées – par exemple SVD pour LSA[Laham, 1997] – mais la problématique demeure avec l'augmentation constante de la production de données. Pour des modèles comme LDA, la réduction dimensionnelle passe par un seuil de fréquence minimale lors de la construction du vocabulaire ou par l'exploitation des n éléments les plus fréquents uniquement. Le choix de la valeur du seuil ou de n se fait à partir d'un équilibre entre temps de traitement, quantité de données et configuration des machines sur lesquelles le programme va s'exécuter. Dans notre contexte et d'après les observations que nous avons pu faire sur l'extrait de corpus d'iSearch, nous avons fixé n =40 000. Cette limite couvre l'ensemble du vocabulaire *pertinent* i.e. tous les mots qui ne sont pas des hapax. Les machines à notre disposition – processeur 4 cœurs 3.6GHz et 16Go de RAM – nous permettent de réaliser nos expériences en un temps raisonnable – environ 4h pour plusieurs centaines de contructions de modèles.

apprentissage : L'apprentissage à partir d'un corpus de référence induit la problématique du surapprentissage, à savoir de la construction d'un modèle ad-hoc : le modèle correspond aux données d'apprentissage mais n'est pas généralisable à des données inconnues. Nous nous épargnons cependant cet écueil avec LDA : la construction du modèle est non supervisée, mais notre évaluation passe par un corpus de référence. Il apparaît cependant sous une autre forme La spécification de k : la spécification du nombre de thèmes à modéliser est liée au corpus de référence, mais la valeur de k peut ne pas trouver de justification mathématique, notamment quand il augmente. Là où le corpus de référence présente $|cats|$ catégories distinctes, la cohérence maximale pour LDA peut être une partition en $k = |cats| \pm n$ thèmes avec $n > 0$. Ce phénomène peut expliquer certains résultats que nous présentons plus loin dans cette section.

Nous venons d'aborder diverses problématiques liées à l'espace de représentation. Une autre problématique inhérente à l'algorithme de construction de modèles de thèmes est à prendre en considération : le non-déterminisme.

4.1.3.3 Non-déterminisme des algorithmes de construction de modèles de thèmes

L'algorithme de construction de modèle de thèmes que nous avons choisi n'est pas déterministe. Cela signifie que les résultats peuvent varier pour un même jeu de données et un même paramétrage. Ce non-déterminisme s'explique par l'utilisation de multinomiales dans la construction des modèles, formules qui s'appuient en partie sur une fonction aléatoire. Dans nos expériences, nous avons pu observer d'importantes différences en comparant les modèles construits pour un même corpus et les mêmes paramètres – nous présenterons les résultats en détail plus loin dans cette section. Afin de pallier pour partie ces variations de modèles, nous avons construit 100 fois chaque modèle pour chaque combinaison de catégories. Les mesures qualitatives – *F-mesure* – collectées nous permettront ensuite de représenter la qualité de chaque modèle sous la forme d'un diagramme de Tukey, particulièrement éloquent dans notre contexte. Bien que non exhaustive, cette représentation des résultats illustre clairement des tendances dans la construction des modèles.

4.1.3.4 Résultats et conclusions

Nous avons traité les quatre grandes catégories comme quatre corpora indépendants. Les variations du score d'alignement pour ces catégories sont représentées synthétiquement dans les figures 4.2, 4.3, 4.6 et 4.7. Sur les quatre figures et pour les corpus à deux classes, nous pouvons observer des *F-mesures* qui varient de 0,88 à 0,98. Ces valeurs très élevées montrent que la catégorisation du corpus est de bonne qualité et que le choix de LDA comme algorithme de construction de modèles de thèmes est approprié.

Les figures 4.2 et 4.3 — respectivement *Physics* et *Condensed Matter* — illustrent une augmentation de la variation des résultats de l'algorithme en fonction du nombre de classes (écart inter-quartile), mais elles ne présentent que peu de variations entre les jeux d'exécutions — indépendamment du nombre de classes. Cela peut s'expliquer par une analyse du vocabulaire exploité par LDA pour les différents jeux.

La table 4.1 présente les tailles de vocabulaires et les nombres d'hapax (termes n'apparaissant qu'une fois dans le corpus) en fonction du corpus et du jeu d'exécution. Les données des jeux LDA-LS et LDA-LSF nous informent qu'en moyenne, pour ces jeux, 83% du vocabulaire est constitué d'hapax. En comparaison, les vocabulaires des jeux LDA et LDA-L comprennent 53%

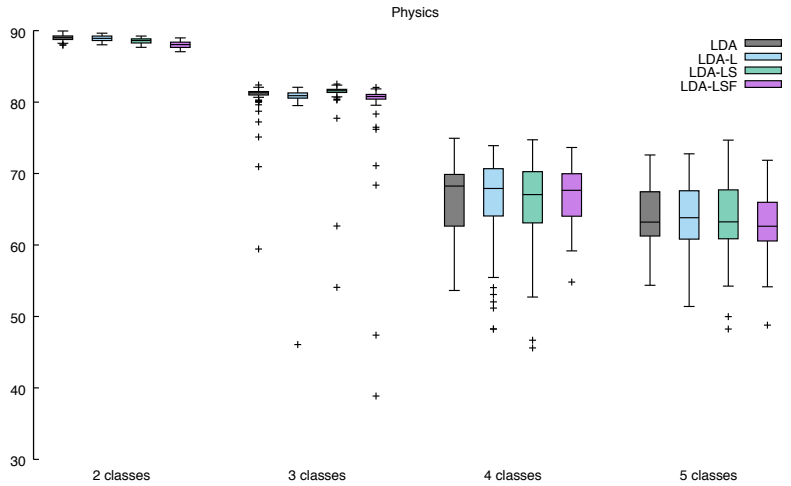


FIGURE 4.2 – Représentations des scores obtenus pour la détection des sous-catégories de «Physics»

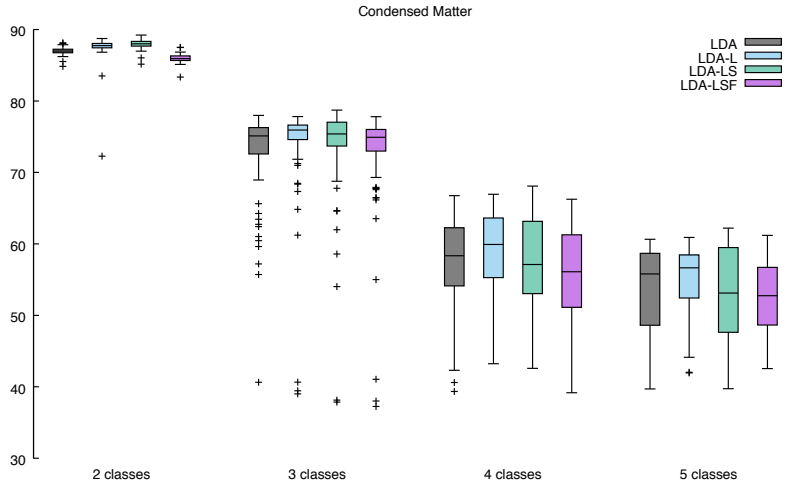


FIGURE 4.3 – Représentations des scores obtenus pour la détection des sous-catégories de «Condensed Matter»

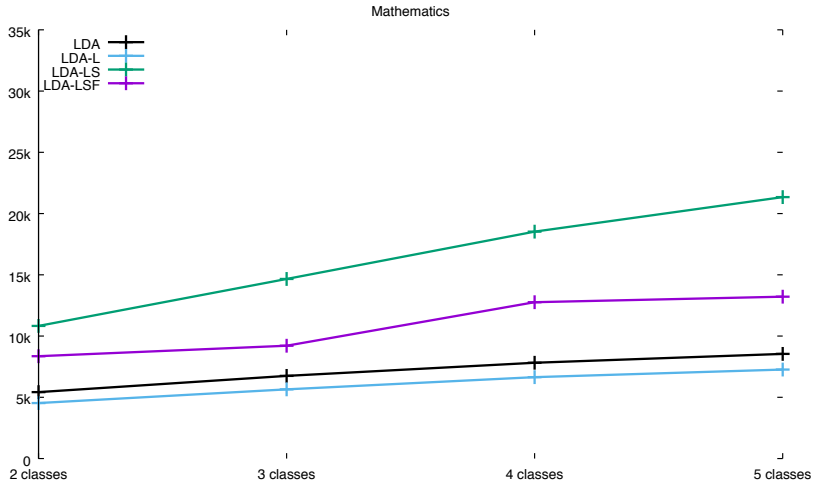


FIGURE 4.4 – Evolution du nombre de termes non hapax en fonction du nombre de classes pour le corpus *Mathematics*

d'hapax en moyenne. Cela peut s'expliquer par notre choix d'extraire le maximum de syntagmes. Les figures 4.5 et 4.4 montrent l'évolution du nombre de termes non hapax pour les catégories *Physics* et *Mathematics* : nous pouvons y remarquer des différences de tailles et d'évolutions. Ces observations nous poussent à prévoir une future étude sur la construction et la distribution tant des unigrammes que des syntagmes.

Il est intéressant de noter que pour les corpora *Physics* et *Condensed Matter*, l'ajout de syntagmes n'apporte que peu voire pas de gain de performance. Nous ne pouvons cependant pas en tirer de conclusion générale dans la mesure où les résultats sur les corpora *Mathematics* et *High Energy Physics* sont de tout autre nature.

Contrairement aux figures 4.2 et 4.3, les figures 4.6 et 4.7 présentent d'importantes variations entre les jeux d'exécutions.

Nous pouvons constater une perte d'une trentaine de points de *F-mesure* pour tous les jeux sur le corpus *High Energy Physics* (figure 4.6) lors de l'ajout de la troisième classe. Alors qu'à deux classes tous les scores de tous les jeux sont supérieurs à 0,95, la troisième classe semble induire l'algorithme en erreur. Une interprétation possible est que la distribution du vocabulaire de la troisième classe est similaire à la distribution du vocabulaire d'une classe

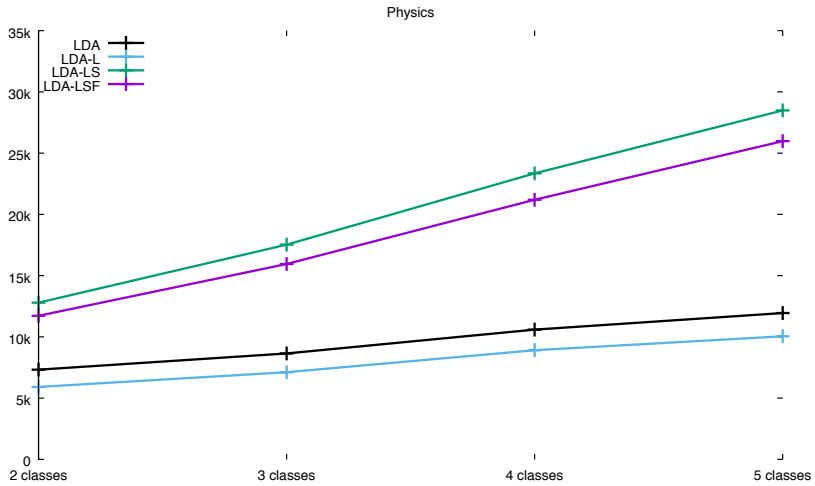


FIGURE 4.5 – Illustration de l'évolution du nombre de termes non hapax en fonction du nombre de classes pour la catégorie *Physics*

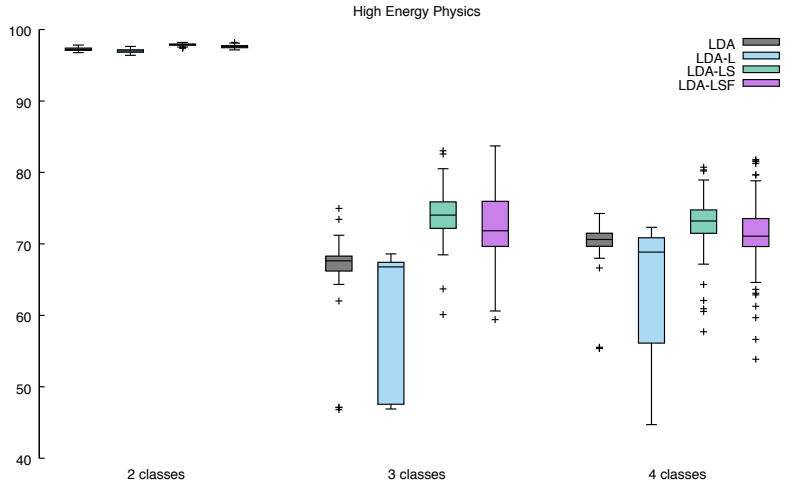


FIGURE 4.6 – Représentations des scores obtenus pour la détection des sous-catégories de «High Energy Physics»

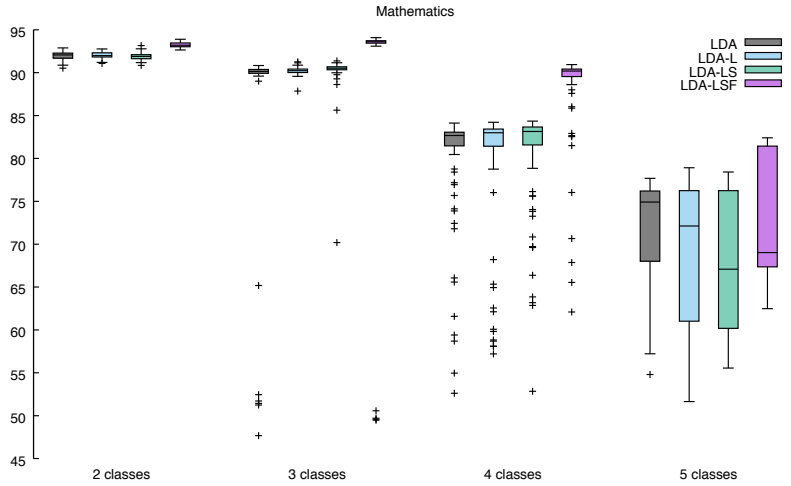


FIGURE 4.7 – Représentations des scores obtenus pour la détection des sous-catégories de «Mathematics»

déjà existante. Cette interprétation est étayée par l'ajout de la quatrième classe, qui ne provoque que de faibles baisses dans les jeux avec syntagmes et des améliorations significatives pour les deux autres jeux. Nous pouvons émettre l'hypothèse que la distribution du vocabulaire de la quatrième classe est particulièrement distincte des distributions des classes existantes, ce qui expliquerait une quasi-stabilité voire une amélioration des résultats alors que l'inverse est attendu quand une classe supplémentaire est ajoutée. Les résultats sur le corpus *High Energy Physics* semblent indiquer que l'exploitation des syntagmes est utile dans les cas où le vocabulaire des thèmes à identifier est particulièrement partagé.

C'est sur le corpus *Mathematics* (figure 4.7) que nous avons obtenu les meilleurs résultats en termes de performance de l'exploitation des syntagmes. Aucune chute des performances aussi importante que précédemment ne peut être observée et il y a d'importantes variations entre les jeux d'exécutions. Nous pouvons observer que dans trois jeux sur quatre, le jeu LDA+LSF est significativement plus performant que les autres.

Les différences de résultats doivent être mises en relation avec les vocabulaires respectifs de chaque catégorie : nous avons obtenu les meilleurs résultats avec les syntagmes sur les catégories avec les vocabulaires les plus

restreints ($|\mathcal{V}_{\text{Mathematics}}| = 78\ 507$ et $|\mathcal{V}_{\text{HighEnergyPhysics}}| = 104\ 060$) et de moins bons sur les catégories avec de plus larges vocabulaires ($|\mathcal{V}_{\text{Physics}}| = 173\ 957$ et $|\mathcal{V}_{\text{CondensedMatter}}| = 165\ 352$).

4.1.4 Exploitation de termes candidats et de leur score dans un modèle de thèmes

Nous présentons ici brièvement une expérience que nous avons menée en lien avec les modèles de thèmes et la terminographie, mais qui s'est avérée non-concluante. Nous avons vu que l'extraction terminographique automatique s'appuie fréquemment sur des scores de potentiels terminologiques. Nous avons également vu que les algorithmes de construction de modèles de thèmes comme LDA s'appuient sur les fréquences d'apparition des lexies du vocabulaire. Nous avons voulu observer l'effet de l'incorporation du potentiel terminologique dans les fréquences utilisées pour la construction des modèles. De la même manière que la *NC-valeur* est une combinaison linéaire de deux scores, nous avons expérimenté avec une combinaison linéaire de la fréquence d'une lexie et sa *NC-valeur*.

$$f^*(\sigma(lexie)) = m \times F(\sigma(lexie)) + n \times NC(\sigma(lexie))$$

Avec $0 \leq m, n \leq 1$ et $m + n = 1$ et où $\sigma(lexie)$ désigne la forme de surface de la lexie – voir section 2.1.2.

A partir de cette formule, nous avons expérimenté avec différentes combinaisons de valeurs pour n et m en remplaçant la fréquence d'une lexie $F(\sigma(lexie))$ normalement utilisée par $f^*(\sigma(lexie))$. Pour rappel, la *NC-valeur* ne s'applique qu'aux lexies composées de plusieurs mots, aussi :

$$|\sigma(lexie)| = 1 \implies f^*(\sigma(lexie)) = F(\sigma(lexie))$$

Le résultat de cette transformation est ensuite passé à l'algorithme de construction de LDA. La méthode d'évaluation suit celle évoquée en section 3, qui est celle employée dans l'expérience précédente.

Les résultats de cette expérience sont sans équivoque : l'intégration de la *NC-valeur* ne fait que nuire à la qualité des modèles construits. Nous avons expérimenté avec les combinaisons suivantes :

m	n
0	1
0,2	0,8
0,4	0,6
0,5	0,5
0,6	0,4
0,8	0,2
1	0

Une incrémentation de n – une augmentation de l'impact de la *NC-valeur* dans l'espace de représentation – mène systématiquement à une baisse de la *F-mesure* finale. Le postulat étant de faire bénéficicer les modèles de thèmes de connaissances terminograhiques, cette expérience a montré que la combinaison d'espaces de représentation n'est pas viable au travers d'une simple combinaison linéaire.

4.1.5 Calculs de corrélations entre distributions de vocabulaires et modèles de thèmes

Dans nos expériences précédentes menées avec LDA, nous avons observé des variations dans la qualité des résultats selon les données placées en entrée, même avec seulement deux catégories. De plus, du fait du non-déterminisme de notre algorithme, nous avons pu observer dans certains cas d'importantes variations de qualité entre plusieurs exécutions sur un même jeu de données. Enfin, l'amplitude de ces variations ne fait que s'accroître avec l'augmentation du nombre k de thèmes.

Nous postulons que la construction de LDA varie selon les distributions de vocabulaires des thèmes à construire, i.e. du vocabulaire des catégories placées en entrée.

Nous présentons ici une expérience de comparaison afin de déterminer une potentielle corrélation entre les résultats de LDA et une similarité entre des paires de catégories du corpus iSearch. Dans la continuité des expériences déjà présentées, nous profitons de cette expérience pour également comparer les corrélations entre espace de représentation simple – mots simples – et espace complexe – avec mots composés [**Delamaire, Amaury** et al., 2019b].

L'expérience peut se résumer comme suit. Nous avons d'abord réuni et

implémenté un grand nombre de mesures de similarité textuelle. Ensuite, pour chaque paire de catégories de notre corpus, nous avons exécuté plusieurs constructions de LDA pour prendre en compte le nom déterminisme et obtenir une *F-mesure* moyenne selon la méthode présentée plus haut. Ces *F-mesures* ont ensuite été corrélées aux similarités textuelles.

4.1.5.1 Espace de représentation

Nous utilisons la représentation habituelle sous forme de sacs-de-mots pour les documents analysés, bien que la définition de *mot* varie entre LDA standard et LDA avec extraction de syntagmes complexes. Dans l'espace de représentation standard, la segmentation en mots s'appuie sur quelques expressions rationnelles simples et n'implique aucune conflation de la variation – lemmatisation, racinisation. A contrario, LDA avec les syntagmes complexes (LDA+NLP) s'appuie sur une chaîne complète de TAL pour extraire les syntagmes pertinents et prendre en compte les différentes graphies d'une lexie donnée. La chaîne de traitements inclut un étiqueteur morphosyntaxique ainsi qu'une analyse en dépendances, suivis par une extraction de syntagmes restreinte aux groupes nominaux, adjectivaux et verbaux.
Exemple : la phrase *Neural networks are useful tools* permet l'extraction des syntagmes *neural_ network*, *useful_ tool* et *be_ useful_ tool*, les deux premiers étant nominaux et le dernier verbal. Afin de calculer des similarités textuelles entre des catégories i.e. des corpora, chaque catégorie est considérée comme la concaténation des documents qu'elle contient.

4.1.5.2 Corpus, algorithmes et modèles

Le corpus exploité dans cette expérience est le même que celui utilisé pour les expériences précédentes, à savoir un extrait du corpus iSearch [Lykke et al., 2010] composé de 4 catégories et 19 sous-catégories – voir section 4.1.1. Les contraintes appliquées aux documents sont les mêmes. Ces 19 sous-catégories nous permettrons de calculer des corrélations entre les similarités textuelles de $n \times \frac{n-1}{2} = 19 \times \frac{19-1}{2} = 171$ paires distinctes avec les *F-mesures* respectives. Comme pour les expériences précédentes, nous avons retenu LDA comme modèle de thèmes. L'algorithme de construction du modèle reste inchangé et les outils de TAL employés sont tous issus de [Manning et al., 2014a]. Nous procédons également à la même transformation d'une classification floue vers une classification stricte – section 3.1.3 – afin d'obtenir les

F-mesure à corréler avec les similarités textuelles. L'alignement entre classes découvertes par LDA et catégories du corpus de référence suit la méthode présentée en section 3.1.3.

4.1.5.3 Cadre de travail

L'objectif de cette expérience est d'estimer la capacité de LDA à distinguer les documents de deux catégories données a priori, i.e. avant de construire le modèle. Cette estimation trouve deux cas d'utilisation : supervisé ou non. Dans un contexte supervisé, cette mesure permettrait d'obtenir une idée de la qualité du modèle contruit : s'il y a corrélation entre *F-mesure* et similarité textuelle, le calcul de la seule similarité permet d'estimer la *F-mesure*. Dans un contexte non supervisé, la similarité pourrait simplement servir de mesure de qualité du modèle en lieu et place de la *F-mesure* ou de limite de convergence lors de la construction du modèle. Afin de déterminer la viabilité de notre hypothèse, ainsi qu'une métrique spécifique à notre objectif, nous mesurons diverses corrélations entre un maximum de mesures de similarité et les *F-mesures* obtenues.

Pour mener notre expérience sur notre corpus catégorisé, nous construisons d'abord un modèle LDA à deux thèmes sur chacune des paires de catégories. Pour pallier le non-déterminisme de l'algorithme de construction – voir section 4.1.3.3 – de multiples constructions de modèles pour une même paire de catégories nous permettent d'obtenir une *F-mesure* moyenne plus représentative bien que non idéale. Pour chaque mesure de similarité, ainsi que pour la *F-mesure*, nous disposons alors d'un vecteur de 171 valeurs, une valeur par paire de catégories. Nous calculons ensuite les coefficients de corrélations de Pearson (R) ainsi que de Spearman (ρ) entre ces vecteurs.

Contrairement à une expérience précédente (section 4.1.3.1) qui a nécessité 100 constructions de modèles pour obtenir les diagrammes de Tukey, notre présente expérience peut se limiter à 10 constructions. Les diagrammes 4.2, 4.3, 4.6 et 4.7 démontrent une stabilité certaine pour $k = 2$, aussi nous sommes nous limités à 10 constructions pour notre expérience sur des paires de catégories.

4.1.6 Mesures de similarité textuelle

Considérant la taille raisonnable de notre corpus, nous pouvons expérimenter avec un nombre important de métriques. Comme évoqué, une catégo-

rie $\mathcal{P}$ composée de documents du corpus est considérée comme un ensemble de lexies associées à leur fréquence – un sac-de-mots. Nous notons P_i la fréquence absolue de la $i^{\text{ème}}$ lexie. Chaque catégorie est donc représentée par le vecteur de fréquences de ses lexies : $P = (P_i), i = 1, n$ où n est la taille du vocabulaire. Certaines métriques exploitent les fréquences relatives p_i où $p_i = \frac{P_i}{\sum P_j}$.

Quand nous comparons deux catégories $\mathcal{P}$ et $\mathcal{Q}$, nous utilisons leur vecteur de fréquences relatives ou absolues, respectivement p et q et P et Q. Dans certains cas nous employons des méthodes ensemblistes – comme $\cap$, $\cup$ ou $\setminus$ – entre le vocabulaire de $\mathcal{P}$ ($Term(\mathcal{P})$) et celui de $\mathcal{Q}$ ($Term(\mathcal{Q})$).

Pour les mesures de correlations non-symétriques et celles de divergences, nous appliquons respectivement les tranformations suivantes :

$$distance_x(\mathcal{P}, \mathcal{Q}) = |1 - corr_x(\mathcal{P}, \mathcal{Q})|$$

$$distance_x(\mathcal{P}, \mathcal{Q}) = \frac{div_x(\mathcal{P}, \mathcal{Q}) + div_x(\mathcal{Q}, \mathcal{P})}{2}$$

La table 4.2 présente les différentes métriques que nous avons retenues pour notre expérience, organisées selon leur typologie.

Divergence based metrics	
Jensen-Shanon	$div_{JS}(\mathcal{P}, \mathcal{Q}) = \frac{div_{KL}(p, m) + div_{KL}(q, m)}{2}$ with $m = \frac{p+q}{2}$
Weighted Euclidian (WED)	$div_{WED}(\mathcal{P}, \mathcal{Q}) = \sqrt{\sum_{p_i \neq 0} p_i (p_i - q_i)^2 + \sum_{p_i = 0} q_i^2}$
Geometry based metrics	
Chord	$dis_{chord}(\mathcal{P}, \mathcal{Q}) = \sqrt{2 - 2\cos(p, q)}$
Cosine	$dis_{cos}(\mathcal{P}, \mathcal{Q}) = 1 - \cos(p, q)$
Minkowski based metrics	
Minkowski	$dis_{Minkowski}(\mathcal{P}, \mathcal{Q}) = \left(\sum \lvert p_i - q_i \rvert^{\alpha} \right)^{\frac{1}{\alpha}}$
Manhattan	$\alpha = 1$
Weighting based metrics	
Canberra	$dis_{Canberra}(\mathcal{P}, \mathcal{Q}) = \sum \frac{\lvert p_i - q_i \rvert}{p_i + q_i}$
χ^2	$dis_{\chi^2}(\mathcal{P}, \mathcal{Q}) = \sum \frac{(p_i - q_i)^2}{p_i + q_i}$
Set operations based metrics	

Alt-intersection	$dis_{altIntersection}(\mathcal{P},\mathcal{Q}) = \frac{1}{card(Term(\mathcal{P}) \cap Term(\mathcal{Q})) + 1}$
Jaccard	$dis_{Jaccard}(\mathcal{P},\mathcal{Q}) = 1 - \frac{card(Term(\mathcal{P}) \cap Term(\mathcal{Q}))}{card(Term(\mathcal{P}) \cup Term(\mathcal{Q}))}$
Correlation based metrics	
Kendall τ	$corr_{\tau}(\mathcal{P},\mathcal{Q}) = \frac{card(concordant\ pairs) - card(discordant\ pairs)}{n(n-1)/2}$
Kendall τ B	$corr_{\tau B}(\mathcal{P},\mathcal{Q}) = \frac{card(concordant\ pairs) - card(discordant\ pairs)}{\sqrt{((n(n-1)/2) - n_1)((n(n-1)/2) - n_2)}}$ where n_1 is the cumulated number of possible tied pairs of values from $\mathcal{P}$ to $\mathcal{Q}$, and reciprocally for n_2
Pearson	$corr_{R}(\mathcal{P},\mathcal{Q}) = \frac{covariance(p,q)}{standard.dev(p) \times standard.dev(q)}$
Other metrics	
Soergel	$dis_{Soergel}(\mathcal{P},\mathcal{Q}) = \frac{\sum \lvert p_i - q_i \rvert}{\sum \max(p_i, q_i)}$
Wave hedges	$dis_{waveHedges}(\mathcal{P},\mathcal{Q}) = 1 - \sum \frac{\min(p_i, q_i)}{\max(p_i, q_i)}$
Bhattacharyya	$dis_{Bhattacharyya}(\mathcal{P},\mathcal{Q}) = \lvert \ln\big(\sum \sqrt{\lvert p_i - q_i \rvert}\big) \rvert$
Kešelj weighted	$dis_{Keselj}(\mathcal{P},\mathcal{Q}) = \sum_{p_i \times q_i \neq 0} \frac{(p_i - q_i)^2}{(p_i + q_i)^2} + card(Term(\mathcal{P}) \setminus Term(\mathcal{Q})) + card(Term(\mathcal{Q}) \setminus Term(\mathcal{P}))$
Hellinger	$dis_{Hellinger}(\mathcal{P},\mathcal{Q}) = \frac{1}{\sqrt{2}} \sqrt{\sum (\sqrt{p_i} - \sqrt{q_i})^2}$

TABLE 4.2: Métriques de similarité textuelle regroupées selon leur typologie

Dans cette table, nous avons regroupé des métriques en considérant soit leur hypothèse sous-jacente, soit leur formalisation mathématique. A des fins de rigueur, nous avons diversifé la nature des métriques comparées. Nous pouvons ainsi regrouper les métriques de Minkowski comme Manhattan (*Taxi cab*). Nous pouvons également regrouper les métriques géométriques comme *chord* et cosinus. Certaines mesures ne s'appuient que sur des opérations ensemblistes, comme les mesures d'intersection et d'alt-intersection. Enfin, des mesures sont dérivées de mesures de corrélation – Kendall et Pearson. Les métriques restantes que nous avons retenues – Soergel, Hellinger, etc. –

ne peuvent être regroupées dans un ensemble cohérent.

4.1.6.1 Résultats

Nous présentons ici les résultats que nous avons obtenus pour les mesures de similarité que nous avons présentées. La table 4.3 ci-dessous synthétise l'ensemble des mesures que nous avons calculées, *F-mesures* et similarités, ainsi que des mesures statistiques comme les écarts-types, minima, maxima et moyenne. La table 4.3 contient également les R et ρ de Pearson et Spearman calculés entre les *F-mesures* et les valeurs de similarité. Enfin, la table contient également les mêmes calculs mais pour LDA+NLP.

Les scores de Pearson et de Spearman **encadrés et en gras** indiquent une corrélation élevée. Les valeurs élevées avec un **asterisque*** indiquent une plus forte corrélation pour LDA+NLP que pour LDA standard. Les valeurs soulignées dénotent des corrélations très basses. Une valeur de corrélation est considérée haute si elle est dans l'intervalle $[-1; -0,7]$ ou $[0,7; 1]$; elle est considérée basse si dans l'intervalle $[-0,5; 0,5]$. Les valeurs non comprises dans ces intervalles sont considérées comme non pertinentes.

STANDARD LDA						
	Minimum	Maximum	Average	Standard Deviation	Pearson Correlation	Spearman correlation
F-mesure	71,980	99,800	94,957	4,025	1,000	1,000
Jensen-Shanon	0,066	0,275	0,151	0,036	**0,752**	**0,917**
WED	0,002	0,016	0,007	0,003	0,692	**0,901**
Chord	0,113	0,426	0,234	0,047	0,629	**0,794**
Cosine	0,006	0,091	0,029	0,012	0,544	**0,793**
Manhattan	0,390	1,053	0,684	0,110	**0,752**	**0,886**
Canberra	2 583	5 110	3 984	406	0,438	0,474
Chi-Square	0,215	0,841	0,474	0,108	**0,754**	**0,909**
Alt-Intersection	3,131	7,496	4,738	0,674	0,629	**0,781**
Jaccard	0,456	0,729	0,597	0,052	**0,753**	**0,878**
Kendall	0,376	1,297	0,742	0,155	**0,736**	**0,912**
Kendall Tau B	0,525	0,872	0,685	0,056	**0,762**	**0,870**
Pearson	0,006	0,092	0,030	0,012	0,544	**0,793**
Soergel	0,327	0,690	0,506	0,061	**0,782**	**0,886**

Wave hedge	2 822	5 354	4 246	403	0,385	0,396
Bhattacharyya	3,531	4,045	3,835	0,083	0,668	**0,735**
Kešelj	2 266	4 809	3 623	412	0,497	0,555
Hellinger	0,285	0,599	0,432	0,055	**0,794**	**0,922**
LDA+NLP						
	Minimum	Maximum	Average	Standard Deviation	Pearson Correlation	Spearman correlation
F-mesure	73,955	99,891	95,866	3,453	1,000	1,000
Jensen-Shanon	0,115	0,474	0,267	0,069	**0,766***	**0,953***
WED	0,004	0,048	0,013	0,006	0,589	**0,930***
Chord	0,527	1,339	0,930	0,143	**0,801***	**0,928***
Cosine	0,139	0,897	0,449	0,135	**0,737***	**0,928***
Manhattan	0,623	1,596	1,071	0,184	**0,790***	**0,944***
Canberra	1 546	6 989	5 573	579	0,168	0,009
Chi-Square	0,376	1,418	0,822	0,198	**0,772***	**0,952***
Alt-Intersection	0,001	8,913	5,568	1,447	0,451	0,699
Jaccard	0,467	0,820	0,725	0,044	**0,814***	**0,903***
Kendall	0,392	1,575	0,928	0,173	0,659	**0,756**
Kendall Tau B	0,428	0,876	0,690	0,073	0,214	0,413
Pearson	0,153	0,928	0,480	0,140	**0,753***	**0,930***
Soergel	0,475	0,888	0,691	0,078	**0,827***	**0,944***
Wave hedge	1 700	7 227	5 801	603	0,116	−0,062
Bhattacharyya	3,421	4,319	4,152	0,070	0,522	0,428
Kešelj	1 339	6 677	5 258	552	0,236	0,104
Hellinger	0,372	0,780	0,576	0,075	**0,809***	**0,958***

TABLE 4.3: Table de résultats présentant les scores de corrélation de Pearson et de Spearman entre mesures de similarité textuelle et qualité de LDA, cette dernière étant exprimée par une *F-mesure*.

Nous pouvons observer dans la table 4.3 que LDA+NLP surpasse LDA en termes de *F-mesure*, ce qui confirme l'intérêt du TAL pour la construction des modèles de thèmes. Nous apprenons également que les distances de Canberra, Kešelj et *wave hedges* ne semblent pas du tout corrélées à la

F-mesure dérivée de LDA, avec ou sans TAL. Avec de très faibles coefficients de Pearson et de Spearman, nous pouvons les éliminer des solutions possibles à notre problématique. Ces faibles corrélations peuvent être expliquées par l'important nombre d'hapax dans le vecteur de fréquences, qui a un effet négatif sur ces trois métriques. Cette hypothèse est étayée par l'ajout du TAL dans LDA+NLP : l'incorporation de lexies complexes augmente fortement le nombre d'hapax, ce qui induit des corrélations moindres. Cette baisse de corrélation trouve sa source dans le calcul des fréquences relatives des lexies, qui va attribuer des fréquences différentes à des hapax selon la longueur du vecteur.

La majorité des métriques considérées montrent une forte corrélation avec les *F-mesures*. Afin de comparer leur performance, nous devons nous appuyer sur des comparaisons entre leurs coefficients de Spearman et de Pearson respectifs. Calculer les corrélations de Spearman nous permet de comparer deux métriques qui ont des valeurs de Pearson élevées et proches. Par exemple, les corrélations de Pearson pour la *cross-entropy* et pour Kullback-Leibler ont des valeurs absolues très proches ($-0{,}706$ et $0{,}714$ respectivement), mais leur coefficient de Spearman est discriminant : $-0{,}922$ pour *cross-entropy* et $0{,}785$ pour Kullback-Leibler.

Avec les corrélations les plus élevées, la distance de Hellinger apparaît être la métrique la plus corrélée à la classification fournie par LDA sans TAL. Bien qu'Hellinger obtienne également de bons scores de corrélations pour LDA+NLP, plusieurs métriques apparaissent pertinentes. Les distances de Soergel et d'Hellinger apparaissent toute deux pertinentes avec des corrélations très élevées et très proches : $0{,}827$ et $0{,}809$ pour leur coefficient de Pearson respectif ; $0{,}944$ et $0{,}958$ pour leur coefficient de Spearman. Nous pouvons également observer une amélioration générale parmi les métriques pertinentes avec l'ajout de procédés du TAL. Aux côtés de Hellinger et de Soergel, qui sont les plus corrélées avec LDA, les distances *chord* et Jaccard montrent également des corrélations élevées. En plus des trois métriques non pertinentes évoquées plus haut, plusieurs métriques ont également présenté des baisses significatives de corrélation avec l'ajout du TAL : alt-intersection, Kendall τ, Kendall τ B et Bhattacharyya. Ces baisses se justifient également par le nombre d'hapax, qui provoque une baisse des corrélations déjà moyennes.

4.1.6.2 Conclusions et perspectives

Nous avons mesuré les coefficients de corrélation de Pearson et de Spearman entre la qualité du modèle de thèmes LDA et plusieurs mesures de similarité textuelle afin d'établir une corrélation entre qualité de classification et similarité textuelle entre corpora. Nous avons également comparé une version standard de LDA avec une version qui inclut des lexies complexes. Nous avons montré que trois métriques – Canberra, Kešelj et *wave hedges* – peut-être rejetées indépendamment de la longueur ou de la forme des sacs-de-mots – avec ou sans lexies complexes. Deux métriques ont montré de fortes corrélations avec les résultats de LDA avec ou sans lexies complexes : les distances de Hellinger et de Soergel. Alors que Hellinger est davantage corrélée au résultat de LDA sans lexie complexe, la distance de Soergel semble prometteuse dans les deux cas. Les distances de Jaccard et de cosinus ont également montré de fortes corrélations, qui ont présenté des améliorations significatives avec l'extraction de lexies complexes – améliorations des coefficients de Pearson et de Spearman.

Nous avons montré avec succès que la qualité du modèle LDA est fortement corrélée à la similarité textuelle des documents à analyser, plus spécifiquement avec les mesures d'Hellinger et de Soergel. Nous avons également pu observer des divergences de comportement avec et sans lexies complexes, sans pour autant pouvoir dégager de tendance générale.

4.2 Extraction automatique de syntagmes terminologiques à partir d'un motif élargi

Nous présentons ici notre expérience portant sur une extraction terminographique à partir de patrons morphosyntaxiques [**Delamaire, Amaury** et al., 2020] (à paraître). Nous l'avons vu en partie 2, différents patrons ont été proposés dans la littérature. [Frantzi et al., 1998] ont proposé une évaluation de ces patrons à partir d'annotations d'experts, permettant de les comparer. Nous proposons de compléter cette méthode d'évaluation à partir d'un corpus de référence et l'application de mesures standards utilisées dans le domaine de la recherche d'information. Nous présentons dans la partie 4.2.2 le corpus exploité ainsi que les prétraitements. Dans la partie 4.2.4 nous développons la méthode d'évaluation utilisée. Dans la partie 4.2.3 nous complétons le "meilleur" patron identifié par [Frantzi et al., 1998]. Nous pré-

sentons les résultats de notre étude comparative en partie 4.2.5 avant de présenter nos conclusions et perspectives en partie 4.2.6.

4.2.1 Présentation de l'expérience

Nous avons comparé la qualité d'extraction de quatre patrons morpho-syntaxiques associés à des calculs de potentiels terminologiques [**Delamaire, Amaury** et al., 2019c].

4.2.1.1 Méthode de calcul de potentiel terminologique : la *NC-valeur*

La fréquence absolue peut donner une idée de la puissance d'un terme candidat : plus elle serait importante, plus le terme candidat pourrait faire partie d'une terminologie. Cependant cette intuition initiale est mise à mal pour les termes candidats composés de plus de deux mots, car toute sous-séquence est au moins aussi fréquente. La notion de *C-valeur* a pour objectif de pallier à ce défaut.

4.2.1.1.1 Calcul de la *C-valeur* Soient $\mathcal{CT}$ un ensemble de termes candidats et $\mathcal{D}$ un ensemble de documents. Nous avons introduit auparavant la notion de fréquence pour un terme candidat x, $F(x, \mathcal{D})$ et la notion de supra-termes, $Supra(ct, \mathcal{CT})$. Afin d'alléger les notations, on sous-entendra $F(x) = F(x, \mathcal{D})$ et $Supra(x) = Supra(x, \mathcal{CT})$.

La notion de *C-valeur*, qui est une mesure du potentiel terminologique d'un terme candidat, se définit comme :

$$C\text{-}valeur(x) = \begin{cases} F(x) \log_2 |x| & \text{si } Supra(x) = \emptyset \\ \log_2 |x| (F(x) - \frac{1}{|Supra(x)|} \sum_{y \in Supra(x)} F(y)) & \text{sinon.} \end{cases}$$

La formule ci-dessus illustre les deux cas à distinguer : si le terme candidat est inclus dans un autre terme candidat ou non. Les termes candidats qui apparaissent fréquemment en tant que sous-chaînes d'un autre terme candidat sont favorisés. Le processus de calcul de la *NC-valeur* ne consiste cependant pas uniquement en une paire de formules à appliquer à une liste de termes candidats ; il inclut également la reconnaissance de termes candidats absents de la liste. Cette méthode d'enrichissement de la liste de candidats

s'appuie sur la liste de termes candidats extraite pour en détecter de nouveaux. L'objectif de cet enrichissement est de pouvoir détecter des termes qui n'apparaissent que comme sous-chaînes d'autres termes mais jamais indépendamment.

Exemple : Soient les termes extraits, avec une fréquence respective de 1 :
— *recurrent neural network*
— *artificial neural network*
— *convolutional neural network*

Le syntagme *neural network* sera reconnu comme potentiellement terminologique car il s'agit d'une sous-chaîne des autres termes candidats. Les autres sous-chaînes des termes candidats (*recurrent neural*, *artificial neural* et *convolutional neural*) seront également extraites mais avec une importance moindre, de fait de leur fréquence plus faible.

L'extraction de sous-chaînes avec un potentiel terminologique nul est problématique car nuisible à la qualité de la terminologie construite et elle augmente le temps de traitement. Alors que [Frantzi et al., 1998] n'évoquent aucun filtre particulier à ce niveau de l'extraction – ils se fient probablement aux fréquences en tant que sous-chaînes et à leur liste de mots vides – il semblerait opportun d'appliquer les mêmes contraintes aux termes candidats initiaux qu'à ceux extraits en tant que sous-chaînes (existence dans une base de connaissances externes, correspondance avec un motif lexico syntaxique, etc.). L'absence de contraintes chez [Frantzi et al., 1998] peut également s'expliquer par la langue de leur corpus (anglais), moins sujette à l'usage des prépositions que le français dans la construction des groupes nominaux – *réseau de neurones* se traduit par *neural network*. De plus, appliquer une contrainte de formation à la liste placée en entrée et à l'extraction de nouveaux termes candidats remet en cause l'indépendance des deux étapes.

4.2.1.1.2 Mots vides : quel traitement dans l'application originale de la formule ? [Frantzi et al., 1998] ne détaillent pas particulièrement l'usage qui est fait de leur liste de mots vides, ce qui pose certaines questions quant à leur gestion :

Faut-il revoir la limite des mots vides dans le contexte d'extraction automatique de termes ? En TAL, les listes de mots vides sont généralement constituées de mots dits fonctionnels ou grammaticaux, qui ne véhiculent pas de sens particulier. La prise en compte

de mots comme *year* ou *great* chez [Frantzi et al., 1998] nécessite une redélimitation du concept de mot vide. Comme l'évoquent les auteurs, le choix de ces nouveaux mots vides s'est fait empiriquement après observations sur le corpus. Dans la perspective d'une extraction terminologique automatique non-supervisée il n'est pas possible de procéder de la même manière. De plus, comme explicité par les auteurs, ces nouveaux mots vides peuvent potentiellement nuire au rappel.

Comment cette liste est-elle construite ? Autant la construction d'une liste de mots fonctionnels ou grammaticaux est directe, autant celle d'une liste de mots vides non grammaticaux ne l'est pas. Les mots *year* et *great* sont estimés « vides » relativement à leur contexte d'apparition ; ce statut n'est cependant pas absolu comme pour les mots grammaticaux.

A quel moment est-elle utilisée ? La liste de mots vides ou grammaticaux peut être exploitée à différents moments de l'extraction automatique : dans le texte original avant son traitement, dans les termes candidats extraits et dans les candidats issus de sous-chaînes. Dans tous les cas, son exploitation est problématique : remplacer des mots dans une phrase avant son analyse biaise son traitement par les outils de TAL, les termes candidats et sous-chaînes ne contiennent a priori pas de mots grammaticaux. S'ils contiennent des mots vides non grammaticaux, se pose alors la question de la délimitation de la *vacuité* d'un mot.

Vu ces problématiques et la subjectivité des mots retenus dans leur liste de mots vides (*great*, *year*, *just*, etc.), nous avons fait le choix de ne pas en avoir dans nos expériences. En lieu et place de liste de mots vides nous reporterons la contrainte de construction initiale de la liste de termes candidats à son enrichissement. Cela nous permettra de nous passer de listes filtrantes ad-hoc et de gagner en cohérence entre les différentes étapes au prix d'une perte d'indépendance entre ces dernières.

4.2.1.1.3 Calcul du facteur contexte Bien que les recherches basées sur [Frantzi et al., 1998] n'en fassent que peu souvent état, la *NC-valeur* consiste en une combinaison linéaire de deux facteurs : la *C-valeur* que nous avons décrit, et le facteur contexte (FC) que nous présentons maintenant.

D'un point de vue intuitif, un terme candidat qui a le même score terminologique qu'un autre terme candidat s'il apparaît plus souvent avec termes "contexte" plus variés, à savoir, outre un usage fréquent il apparaît avec des mots qui apparaissent souvent et qui ont un sens. De manière logique ces termes contexte serait plutôt des unitermes et avec une certaine fonction grammaticale, ce que nous fait éliminer les articles et les conjonctions.

Un idem w est voisin d'un terme x s'il existe au moins un document d tel que w précède a ou x précède w dans d. On peut formaliser :

$\exists d \in \mathcal{D}, \exists i : (d(i) = w \text{ et } d(i{+}1) = x(1)) \text{ ou } (d(i) = x(length(x)) \text{ et } d(i{+}1) = w)$

On appellera x voisin de w dans le document d à la position i.

Le contexte d'un terme candidat x dans un corpus $\mathcal{D}$ est défini comme l'ensemble des items voisins (les positions varient très peu) à x dans un quelconque document du corpus :

$Contexte(x) = \{w | length(w) = 1, \pi(x) \in \{NN, JJ, VB\}, x \text{ et } w \text{ sont voisins}\}$

Il est normal qu'un même item x puisse apparaître dans le voisinage des différents termes candidats, tout comme un terme candidat x et un des ses voisins occurrent ensemble comme voisins dans un même document ou dans des documents différents. Le poids d'un voisin sera plus fort s'il occurre avec des termes candidats différents.

$$weigth(w) = \frac{F(\{x | w \in Contexte(x)\})}{F(\mathcal{CT})}$$

La fréquence d'occurrence d'un couple (w, x) terme candidat, voisin est aussi mesurable :

$f(w, x) = F(\{(i, d) | i \text{ indice}, d \in \mathcal{D} \text{ tel que } w \text{ voisin de } x \text{ dans } d \text{ à la position } i\}$

Le facteur contexte se défini comme le produit scalaire :

$$F\text{-}contexte(x) = \sum_{w \in Contexte(x)} f(w, x) weigth(w)$$

4.2.1.1.4 Calcul de la *NC-valeur*

$$NC\text{-}valeur(x) = \alpha C\text{-}valeur(x) + \beta F\text{-}contexte(x)$$

avec $\alpha + \beta = 1$, on prend le plus souvent $\alpha = 0.8$ et $\beta = 0.2$

4.2.1.2 Déroulé de l'expérience

Nous venons de le voir, la *NC-valeur* est une combinaison linéaire de deux poids, la *C-valeur* et le facteur contexte (FC). Pour chaque patron, nous avons effectué une extraction sur le corpus ACL-RD-TEC à partir d'annotations morphosyntaxiques issues de Stanford CoreNLP [Manning et al., 2014a]. Plusieurs listes triées ont ensuite été construites afin de déterminer l'incidence de l'introduction du facteur contexte. Les patrons de la littérature que nous comparons sont les suivants :

Patron 1 : $((\text{JJ}|\text{NN})^{+}((\text{JJ}|\text{NN})^{*}(\text{NN IN})^{?})(\text{JJ}|\text{NN})^{*})\text{NN}$

Patron 2 : $\text{JJ}|\text{NN}^{+}\text{NN}$

Patron 3 : NN^{+}NN

Les cinq listes par patron ont été construites à partir des combinaisons linéaires suivantes :

1. $1 \times$ *C-valeur* $+ \; 0 \times$ FC
2. $0,8 \times$ *C-valeur* $+ \; 0,2 \times$ FC
3. $0,5 \times$ *C-valeur* $+ \; 0,5 \times$ FC
4. $0,2 \times$ *C-valeur* $+ \; 0,8 \times$ FC
5. $0 \times$ *C-valeur* $+ \; 1 \times$ FC

Le processus d'extraction produit donc cinq listes triées par patron morphosyntaxique, chaque tri effectué selon une combinaison linéaire de *C-valeur* et FC. Pour les combinaisons 1 et 5, seul l'un des deux poids est pris en compte – respectivement la *C-valeur* et le FC. Une comparaison entre les patrons présentés ici met en évidence une limite de taille intrinsèque pour tous excepté le patron 1. La complexité de sa structure nous a poussés à appliquer la limite après l'extraction, nous permettant d'obtenir des graphies de taille similaire pour les quatre patrons. L'application de la méthode d'évaluation que nous introduisons dans la partie 4.2.4 nous permettra de comparer non seulement les patrons entre eux, mais également les patrons avec eux-mêmes pour des combinaisons de *C-valeur* et FC distinctes.

4.2.2 Données et outils

Nos expériences ont été menées avec Stanford CoreNLP [1] [Manning et al., 2014b], qui permet d'effectuer un étiquetage morphosyntaxique et une lem-

1. `https://stanfordnlp.github.io/CoreNLP/`

matisation rapides et de qualité. Les méthodes d'extraction et de tri ont été implémentées en Java 8. Dans un premier temps nous avons mené des expériences [**Delamaire, Amaury** et al., 2019a] sur le corpus en anglais iSearch [Lykke et al., 2010], catégorisé par thème et constitué de documents techniques. Nous avons en premier lieu réalisé les expériences développées en section 4.1 sur iSearch, ce qui nous a naturellement menés à des expériences purement terminographiques. Néanmoins, en l'absence d'experts capables d'extraire manuellement ou de valider des termes candidats, iSearch ne nous a pas permis de comparer différents systèmes. Nos extractions préliminaires par patrons syntaxiques nous ont cependant permis d'observer certains silences sur des termes qui sont clairement terminologiques, silences auxquels nous tentons de répondre dans ces expériences et que nous avons présentés en section 3.2.

Nous avons retenu le corpus ACL-RD-TEC 1.0 [2] [Handschuh and QasemiZadeh, 2014] comme corpus de référence pour notre expérience. Il est composé d'environ 11 000 articles scientifiques en anglais sur le traitement automatique du langage et ses domaines connexes, où chaque article est associé à sa liste de termes. Les termes ont d'abord été extraits par un patron morpho-syntaxique, avant d'être triés selon leur *C-valeur*. La validation finale a été effectuée manuellement par plusieurs annotateurs. Les propriétés du corpus sont les suivantes :

- 200 méga octets de corps de textes sans balises XML,
- 34 millions de mots,
- 25 000 graphies terminologiques distinctes annotées manuellement,
- 21 500 graphies terminologiques distinctes composées de plusieurs mots,
- 85,8 graphies terminologiques en moyenne par document,
- 40,5 graphies terminologiques composées de plusieurs mots en moyenne par document.

La distinction faite entre *termes techniques* et *autres termes* dans ACL-RD-TEC ne nous a pas paru pertinente, aussi n'avons-nous considéré qu'un ensemble de graphies constitué des *termes techniques* et des *autres termes*. Comme les chercheurs le détaillent [Handschuh and QasemiZadeh, 2014], les deux ensembles ne sont pas exclusifs.

Exemple : Liste exhaustive des graphies annotées pour le premier document du corpus, intitulé *A dialogue-based system for identifying parts for medical systems*

2. `http://pars.ie/lr/acl-rd-tec-terminology`

Termes techniques : *algorithm*, *context-based parser*, *database*, *dialogue manager*, *dialogue system*, *dialogue-based system*, *human-like dialogue*, *identification*, *information technology*, *interactive voice response*, *matching*, *matching algorithm*, *natural language dialogue*, *parser*, *recognition*, *recognition system*, *recognition technology*, *speech recognition*, *speech recognition system*, *speech recognition technology*, *sub-string matching*, *text-to-speech*, *text-to-speech system*, *tutoring system*, *world wide web*

Autres termes : *abbreviations*, *bigram*, *call center*, *case*, *characters*, *data structures*, *dialogues*, *dictionary*, *document*, *domain-specific information*, *domain-specific knowledge*, *feature*, *heuristic*, *heuristics*, *human operator*, *index*, *input string*, *key words*, *keyword*, *knowledge*, *lexicon*, *names*, *natural language*, *phrase*, *process*, *queries*, *recognition quality*, *semantic*, *semantic knowledge*, *sentences*, *spoken natural language*, *sub-string*, *syntax*, *system architecture*, *technology*, *text*, *training*, *training corpus*, *tree*, *tutoring*, *unigram*, *user*, *user utterance*, *utterance*, *word*, *words*

La qualification de *techniques* pour des termes comme *algorithm* ou *world-wide-web* mais pas pour *bigram* ou *data structures* ne nous a pas semblée pertinente dans la mesure où nous n'envisageons pas de partition de l'ensemble des termes. Il est à noter dans l'exemple donné ci-dessus que des formes fléchies apparaissent dans les graphies annotées – *word* vs *words*, *sentences*, *dialogues*, etc. – paramètre qu'il nous faudra prendre en compte lors de l'analyse de nos résultats en section 4.2.5. Le nombre de mots dans les graphies est également significatif : la majorité des graphies – 55% – n'est constituée que d'un seul mot, contrairement à ce que semblent indiquer [Justeson and Katz, 1995], exception faite du domaine de la médecine. La méthode initiale du calcul de la *C-valeur* [Frantzi et al., 1998] ne prévoit de tri que pour les graphies composées de plusieurs mots ; les chercheurs [Handschuh and QasemiZadeh, 2014] ont dû adapter la formule pour leurs fins. La *C-valeur* est (proche d')une mesure de lexicalisation, elle tend donc naturellement à attribuer des scores plus élevés aux unités les plus courtes. Il s'agit très certainement la raison de la surreprésentation des unigrames dans les termes annotés du corpus ACL-RD-TEC relativement aux observations préalables dans le domaine de la terminologie computationnelle. Malgré les biais que le corpus ACL-RD-TEC peut présenter – comme n'importe quel corpus –, il va nous permettre de comparer la qualité de l'extraction de différents patrons morphosyntaxiques,

ainsi que l'impact de l'incorporation du facteur contexte, initialement décrit par [Frantzi et al., 1998]. Parmi les patrons comparés, trois sont de la littérature, nous en proposons un quatrième que nous présentons dans la section suivante et nous détaillons en section 4.2.4 l'application de la méthode d'évaluation de recherche d'information introduite en section 3.3.4.

4.2.3 Exploitation d'un nouveau patron

Plusieurs patrons morphosyntaxiques ont été proposés dans la littérature. Certains nous semblent trop permissifs, notamment du fait des groupes prépositionnels [Justeson and Katz, 1995] ou des conjonctions de coordinationet déterminants [Park et al., 2002], alors que d'autres nous semblent trop contraints. Nous proposons de compléter le patron identifié comme étant le plus pertinent par [Frantzi et al., 1998]. En effet, ils obtiennent les meilleurs résultats avec des séquences de noms communs et adjectifs terminant par un nom :

Patron 3 : $(\text{JJ}|\text{NN})^{+}\text{NN}$

Nous proposons de compléter cette séquence avec d'autres éléments, repris par d'autres chercheurs, notamment [Park et al., 2002]. Nous proposons l'introduction de deux formes verbales dans le patron 3 : les gérondifs – verbes *-ing* – et les participe-passé. Cette modification se justifie par une similarité de comportement et de forme entre gérondifs, participes et adjectifs : ils agissent comme modifieurs d'un nom commun, le tout formant potentiellement une graphie terminologique. De plus, les étiqueteurs morphosyntaxiques confondent parfois les trois ; [Handschuh and QasemiZadeh, 2014] documentent cette problématique lors de la construction du corpus. Ajouter ces deux formes devrait théoriquement permettre de repérer des graphies pertinentes ignorées au préalable. De plus, conformément au jeu d'étiquettes de Stanford CoreNLP – à savoir celui du PENN TREEBANK [3], nous avons respectivement étendu les formes nominales et adjectivales aux formes plurielles (NNS) et comparatives/superlatives (JJ[RS]). Une fois complété et étendu, le patron 3 devient :

Patron 9 : $(\text{JJ}(\text{R}|\text{S})^{?}|\text{NN}(\text{S})^{?}|\text{VB}[\text{GD}]|\text{NNP}(\text{S})^{?})^{+}\text{NN}(\text{S})^{?}$

Où VB est un verbe soit au gérondif (VBG) soit au participe passé (VBD) et NNP un nom propre potentiellement au pluriel (NNPS). Nos expériences

3. https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

préliminaires sur le corpus iSearch [Lykke et al., 2010] nous ont rapidement menés à un constat : la présence de noms propres dans les termes, ignorés dans la plupart des patrons de la littérature. Nous introduiront donc la possibilité d'un nom propre comme modifieur du nom commun final. L'introduction des noms propres permettrait de repérer des graphies telles que *Markov chain* ou *Lie algebra*, qui correspondent effectivement à des termes spécifiques aux mathématiques. Dans ces expériences, nous comparerons les patrons 1, 2, 3 et 9 :

Patron 1 : $((\text{JJ}|\text{NN})^+((\text{JJ}|\text{NN})^*(\text{NN IN})^?)(\text{JJ}|\text{NN})^*)\text{NN}$

Patron 2 : NN^+NN

Nous avons sélectionné un corpus de référence ainsi que les patrons morpho-syntaxiques à comparer. Nous détaillons maintenant notre méthode d'évaluation.

4.2.4 Méthode d'évaluation

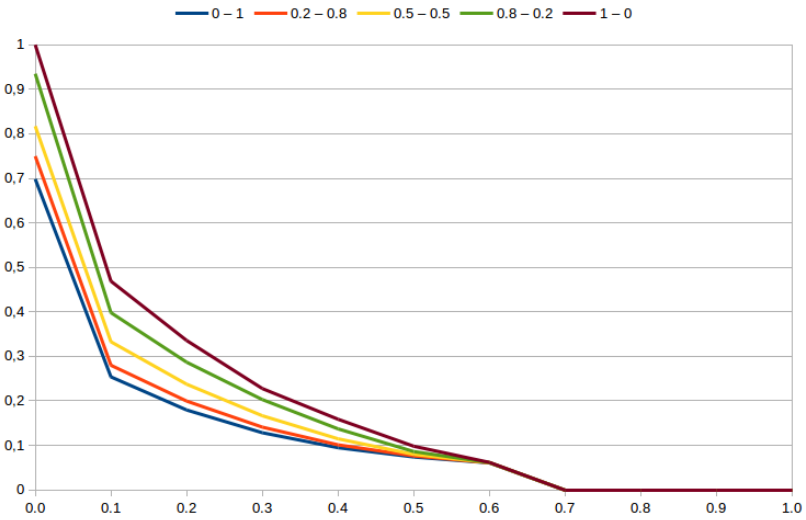
Nous avons présenté notre méthode d'évaluation basée sur la recherche d'information en section 3.3.4. Nous présentons ici son implémentation dans le cadre de notre expérience. Nous avons créé un index référençant les termes annotés et leur document, qui nous servira à construire les ensembles de documents et de termes nécessaires aux calculs de TREC_EVAL. Nous avons divisé le corpus ACL-RD-TEC en 55 blocs de 170 documents environ, avant de procéder aux différentes extractions à partir des quatre patrons et des cinq combinaisons de *C-valeur* et FC, soit 20 évaluations à effectuer. Nous avons pour cela complété l'index avec les *C-valeurs* et FC des termes relativement au bloc dont ils font partie – la *NC-valeur* dépend du nombre de mots du corpus, la position d'un élément dans la liste finale peut donc varier.

4.2.5 Résultats de l'étude comparative

Nous avons comparé la qualité d'extraction de quatre patrons morpho-syntaxiques associés à des calculs de *C-valeur* et *NC-valeur*. Nous l'avons vu précédemment, la *NC-valeur* est une combinaison linéaire de deux poids, la *C-valeur* et le facteur contexte (FC). Pour chaque patron, nous avons effectué une extraction sur le corpus ACL-RD-TEC à partir d'annotations morphosyntaxiques issues de Stanford CoreNLP [Manning et al., 2014b]. Plusieurs listes

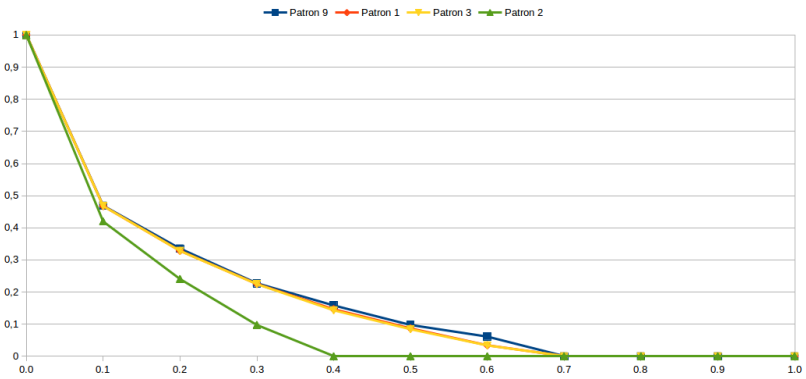
triées ont ensuite été construites afin de déterminer l'incidence de l'introduction du facteur contexte. Chaque tri est effectué selon une combinaison linéaire de *C-valeurs* et FC – voir section 4.2.1. Pour les combinaisons 1 et 5, seul l'un des deux poids est pris en compte – respectivement la *C-valeur* et le FC. Une comparaison entre les patrons présentés ici met en évidence une limite de taille intrinsèque pour tous excepté le patron 1. La complexité de sa structure nous a poussés à appliquer la limite après l'extraction, nous permettant d'obtenir des graphies de taille similaire pour les quatre patrons. Les tendances observées pour chaque patron lors de l'incrémentation du coefficient du facteur contexte montrent que 5 paires de coefficients distincts sont suffisantes.

FIGURE 4.8 – Précisions du patron 9 pour diverses combinaisons de *C-valeur* et FC. 0-1, 0.2-0.8, 0.5-0.5, 0.8-0.2 et 1-0 sont à comprendre comme une paire de coefficients : <Coefficient *C-valeur* > - < Coefficient FC>. Les courbes des autres patrons sont similaires dans leur comportement : la précision décroit systématiquement avec l'incrémentation du coefficient du facteur contexte.



Les figures 1 et 2 permettent d'analyser la distribution des termes et non-termes dans la liste triée construite, distinction faite sur la base des annotations du corpus ACL-RD-TEC. Avec la figure 4.8, nous illustrons l'impact négatif du FC avec le patron 4. Les tendances sont les mêmes pour tous les autres. La figure 2 permet de comparer les quatre patrons pour une combi-

FIGURE 4.9 – Comparaison des précisions des 4 patrons. Ca graphique est moins éloquent : pour rappel, [Handschuh and QasemiZadeh, 2014] ont procédé à un extraction par patron morphosyntaxique avant de trier la liste par *C-valeur*, d'où la grande précision sur les premiers éléments des listes construites.



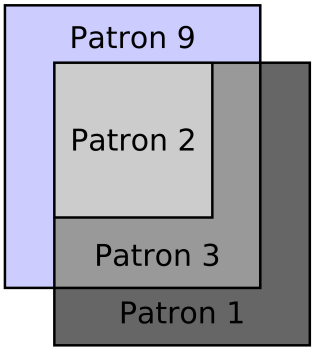
naison unique (optimale) de *C-valeur* et FC. Ces courbes nous permettent d'observer plusieurs phénomènes :

- L'introduction du facteur contexte nuit systématiquement à la qualité des résultats,
- La *C-valeur* produit de très bons résultats relatifs au corpus ACL-RD-TEC : les premiers éléments des listes triées sont essentiellement des vrais termes, les derniers des non-termes,
- Trois des quatre patrons comparés ont des résultats très proches (patrons 1, 3, et 9), le quatrième est très en-deçà (patron 3),
- Une légère prééminence du patron proposé (patron 4) est observable.

La faible qualité de l'extraction à partir du patron 1 peut s'expliquer par le biais morphosyntaxique qui caractérise le corpus d'évaluation : aucun terme annoté dans ACL-RD-TEC ne contient de préposition. Comme vu en partie 4.2.1, cela s'explique par le processus de construction du corpus, qui s'appuie sur un patron morphosyntaxique et sur sur la *C-valeur* sans facteur contexte. L'absence de groupe prépositionnel dans le patron utilisé pour le corpus explique la rapide baisse de qualité du patron 4, cependant les courbes du graphique 4.9 illustrent un phénomène ensembliste relatif aux patrons que nous avons retenus.

Nous pouvons effectivement constater que l'ensemble des graphies recon-

FIGURE 4.10 – Représentation des ensembles de graphies extraits par les patrons comparés.



nues par le patron 2 le sont également par le patron 3 :

$$(\text{NN}|\text{JJ})^{+}\text{NN} \equiv (\text{NN}^{+}\text{NN})|(\text{NN}|\text{JJ})^{+}\text{NN} \equiv \text{PATRON 2}|\text{PATRON 3}$$

Il en est de même entre les patron 2, 3 et 9 :

$((\text{JJ}|\text{NN})^{+}((\text{JJ}|\text{NN})^{*}(\text{NN IN})^{?})(\text{JJ}|\text{NN})^{*})\text{NN}$
$\equiv ((\text{NN}|\text{JJ})^{+}\text{NN})|((\text{JJ}|\text{NN})^{+}((\text{JJ}|\text{NN})^{*}(\text{NN IN})^{?})(\text{JJ}|\text{NN})^{*})\text{NN}$
$\equiv$ PATRON 3|PATRON 1

De même pour les patrons 3 et 9 :

$(\text{NN}|\text{JJ}|\text{VB}[\text{GD}]|\text{NNP})^{+}\text{NN}$
$\equiv(\text{NN}|\text{JJ})^{+}\text{NN}|(\text{NN}|\text{JJ}|\text{VB}[\text{GD}]|\text{NNP})^{+}\text{NN}$
$\equiv$PATRON 3|PATRON 9

Alors que le graphique 4.9 montre une équivalence relative entre les patrons, le patron 4 introduit ici permet l'extraction de nouvelles graphies :

- *automated information*
- *named entity matcher*
- *binarized PCFGs*
- *developing chart*
- *scoping pattern*

La comparaison des résultats des différentes combinaisons linéaires semble contre-intuitive. La théorie de Harris veut que le sens d'un mot est déterminable par les mots qui l'entourent, théorie corroborée par les résultats de [Mikolov et al., 2013a] – entre autres. A contrario, nous pouvons observer

ici une dégradation constante de la qualité du tri relative à l'augmentation du facteur contexte. La raison peut être la taille de la fenêtre retenue lors du calcul des poids des mots contexte dans le facteur contexte : nous avons retenu la phrase comme fenêtre de contexte. La réduire aux mots strictement mitoyens (précédent et suivant) pourrait améliorer les résultats observés ici.

4.2.6 Conclusions et perspectives

Les patrons morphosyntaxiques sont largement utilisés pour la construction automatique de terminologie. Nous avons réalisé une étude comparative de plusieurs de ces patrons, parmi lesquels un patron modifié en fonction de nos observations. Nous introduisons également une méthode d'évaluation de terminologies basée sur celles de la recherche d'information. Nos résultats mettent en évidence les biais introduits par l'annotation manuelle ainsi qu'une légère prééminence du nouveau patron proposé. Nous avons également pu observer une dégradation de la qualité des résultats lors de l'intégration du facteur contexte dans le calcul de la *NC-valeur*. Nous avons montré qu'une utilisation des mesures de recherche d'information classiques pouvait permettre de comparer différents systèmes d'extraction de termes candidats, ce à partir d'un corpus annoté. Cette expérience est à reproduire sur d'autres terminologies préconstruites et avec davantage de patrons. Les fenêtres de calculs des facteurs contextes dans la *NC-valeur* sont également à varier afin d'en analyser l'impact.

Chapitre 5

Conclusions et perspectives

La construction automatique de terminologies est une tâche complexe du TAL, dont les avancées peuvent bénéficier à d'autres procédés. Il n'y a que peu de consensus parmi les chercheurs sur la formalisation, sur les définitions ainsi que sur les méthodes à employer. Qu'ils s'agissent des limites linguistiques théoriques des lexies terminologiques, des méthodes d'automatisation ou de l'évaluation, la construction de terminologies n'en est pour l'instant qu'à un niveau de recherches préliminaires. Le peu de consensus sur la notion de *terminologie* en général relève probablement des multiples applications des méthodes d'extraction automatique. La définition proposée dans le Larousse reflète cette multiplicité d'aspects :

> Ensemble des termes, rigoureusement définis, qui sont spécifiques d'une science, d'une technique, d'un domaine particulier de l'activité humaine.
> Discipline qui a pour objet l'étude théorique des dénominations des objets ou des concepts utilisés par tel ou tel domaine du savoir, le fonctionnement dans la langue des unités terminologiques, ainsi que les problèmes de traduction, de classement et de documentation qui se posent à leur sujet. [1]

La définition du Larousse distingue clairement deux aspects : la ressource créée – un ensemble de termes – et le sujet de recherche. De fait, l'automatisation du processus recouvre les deux à la fois. Le chercheur doit déterminer des heuristiques de reconnaissance de lexies terminologiques à partir de connaissances linguistiques afin de construire une ressource correspondant

1. https://www.larousse.fr/dictionnaires/francais/terminologie/77407

aux besoins exprimés. C'est notamment des variations dans la nature de cette ressource qu'émanent certaines dissensions entre chercheurs. Les définitions et limites de ce qui constitue une lexie terminologique varient selon les objectifs voulus par ces derniers. Divers rapprochement ont ainsi été faits avec d'autres domaines du TAL et de la science des données en général.

Un rapprochement a notamment été fait entre extraction terminologique et indexation de documents, donc avec le domaine de la recherche d'information. L'indexation permet à un moteur de recherche de répondre à une requête émise par un utilisateur ; l'extraction terminologique y est exploitée pour y ajouter des connaissances. L'indexation des lexies terminologiques concerne les moteurs de recherche spécialisés ; elle leur permet de mieux répondre aux attentes de l'émetteur de la requête par la reconnaissance de concepts spécifiques.

D'un point de vue pratique, l'extraction n'aboutit pas à la construction d'une terminologie mais à l'enrichissement d'une indexation préexistante. Les lexies reconnues à ces fins ne sont donc pas évaluées par un expert, elles sont évaluées en regard de la qualité de l'indexation. L'évaluation des systèmes de recherche d'information est quant à elle particulièrement consensuelle, avec des mesures et méthodes largement acceptées.

Un rapprochement a également été fait entre ontologies et terminologie. Une ontologie est une structure de données permettant la description de concepts et des liens qu'ils peuvent entretenir. Les liens peuvent décrire des relations hiérarchiques – subsomption, IS A (est un), marqueur de relation hiérarchique – ou sémantiques – toutes les autres relations inter-conceptuelles possibles.

Comme l'indique la définition du Larousse, une terminologie consiste en un ensemble de termes, mais son étude consiste – entre autres – à les organiser. Un rapprochement vers des structures de connaissances plus sophistiquées que des lexiques est donc pertinent et reste dans les limites de la définition. L'organisation des termes du lexique pose cependant des questions spécifiques aux méthodes d'automatisation et d'évaluation. La détection de liens sémantiques entre les lexies extraites est une tâche à part entière, de même que l'évaluation de l'ontologie finale.

Les choix liés à la structure de l'ontologie et/ou à la construction des concepts et relations induisent également une variation dans la nature de la ressource construite : au même titre que la définition de *terminologie* varie

d'un chercheur à l'autre, les contraintes appliquées sont rarement les mêmes. Les variations peuvent se trouver à divers niveaux de traitement, qui rendent difficilement comparables les différentes solutions proposées dans la littérature. Nous pouvons notamment identifier deux pôles importants générateurs de variations : la construction/identification des concepts et les contraintes sur les relations.

La construction des concepts terminologiques dans une ontologie s'appuie notamment sur des étapes de conflation de la variation pour trouver les formes canoniques des lexies, hors cette étape est très inégalement appliquées chez les chercheurs. Nous avons pu observer des traitements simples (lemmatisation, racinisation) mais également des traitements bien plus complexes (acronymes, synonymes, variation grammaticale, dérivation, etc.), traitements qui nécessitent du temps à mettre en place et qui produisent des inégalités dans la construction des concepts.

Les variations de contraintes sur les relations portent quant à elles sur le choix des chercheurs d'étendre leur structure de données au delà de la subsomption : une structure avec des relations uniquement de la forme IS A n'est pas comparable à une ontologie – une ontologie avec uniquement des relations IS A est une hiérarchie.

Nous avons proposé dans nos recherches un rapprochement avec les modèles de thèmes. Les modèles de thèmes sont des modèles statistiques supposés représenter les thèmes inhérents à un corpus au travers de distributions de fréquences sur un vocabulaire donné. L'hybridation entre terminologie et modèles de thèmes trouve sa justification sur le plan théorique : la terminologie consiste en un ensemble de lexies spécifiques à un domaine, un thème est décrit par une distribution de fréquences sur un vocabulaire. Pour la terminologie comme pour les modèles de thèmes, le postulat linguistique porte sur l'aspect discriminant du vocabulaire employé : le vocabulaire terminologique comme discriminant du domaine auquel appartient le corpus, les distributions de fréquences comme discriminants entre les thèmes.

Nous avons expérimenté avec diverses formes d'hybridation qui ont abouti à la confirmation de ce postulat. Nous avons notamment pu observer un gain de qualité lors de l'intégration de graphies terminologiques dans le vocabulaire d'un modèle de thèmes. Le bénéfice observé pour la qualité des modèles a ensuite été confirmé par des calculs de corrélations qui ont mis en évidence la validité du postulat. Une reproduction de l'expérience avec un autre algorithme – de *clustering* cette fois – nous a permis d'observer un gain de

qualité équivalent à partir des mêmes données et traitements linguistiques.

Bien que nous sommes parvenus à prouver la pertinence de l'hybridation entre terminologie et modèles de thèmes, celle-ci ne vaut que pour certains aspects : nous ne sommes pas parvenus à intégrer le poids terminologique dans l'espace de représentation d'un modèle de thèmes de manière pertinente, toutes nos tentatives se sont soldées par une baisse de qualité du modèle. Cette hybridation reste à investiguer davantage avec d'autres modèles de thèmes et/ou des combinaisons d'espaces de représentation plus élaborées.

Les natures variées de la ressource créée aboutissent à de nombreuses propositions de chercheurs spécifiques à leur contexte. De fait, peu de solutions proposées sont suffisament généralisables. Certaines étapes de la construction de la ressource terminologique s'établissent néanmoins peu à peu comme des méthodes de références.

L'étape indispensable et préliminaire à toute construction automatique de ressource terminologique consiste à identifier les graphies concernées dans un texte. Parmi les nombreuses étapes de la construction qui sont sujettes à dissensions, la méthode d'extraction des graphies est relativement consensuelle. Diverses méthodes ont été proposées à ces fins dans la littérature, mais depuis les années 90, la méthode des patrons morphosyntaxiques s'est établie comme référence. Un patron morphosyntaxique décrit une séquence d'étiquettes éponymes *acceptables* pour qu'une graphie soit considérée comme terminologique. Les chercheurs peuvent définir un patron unique – généralement sous forme d'une expression rationnelle – ou une liste de patrons, sans distinction sur le traitement. Les patrons sont particulièrement intéressants car aisés à modifier et à comprendre, ils sont cependant dépendants de la langue analysée. De plus, l'application de patrons ne nécessite qu'un étiquetage morphosyntaxique préalable, contrairement à des méthodes qui peuvent nécessiter des analyses plus coûteuses.

C'est également depuis les années 90 que les travaux en terminologie se sont concentrés spécifiquement sur les groupes nominaux, supposés véhiculer davantage de spécialisation que les autres groupes syntaxiques. Les patrons morphosyntaxiques sont donc généralement utilisés pour identifier ces groupes dans des documents spécialisés, soutenus par des contraintes structurelles – les déterminants sont par exemple ignorés car ils sont uniquement générateurs de variations et ne véhiculent aucun sens. Du point de vue linguistique, davantage que les groupes nominaux, ce sont les noms communs

et certains de leurs satellites qui sont considérés comme pertinents.

La diversité des propositions de patrons porte sur les satellites possibles ainsi que leur nombre pour une graphie terminologique. Si – pour l'anglais – tous les chercheurs s'accordent à extraire les adjectifs directement voisins du nom central, il n'en est pas de même pour d'autres catégories comme les conjonctions de coordination, les verbes conjugués ou les propositions subordonnées relatives. La littérature diverge également quant à la taille maximale des graphies terminologiques. La problématique de la gestion des entités nommées donne lieu à des postulats forts sur leur absence/présence dans les graphies terminologiques. Alors que de nombreux auteurs les excluent des graphies extraites, nous avons pu observer de nombreuses lexies terminologiques contenant – et non consistant en – une entité nommée, à tout le moins sur notre corpus constitué de documents hautement spécialisés.

Une partie des graphies reconnues par un patron peut l'être à tort, de même que certaines graphies pertinentes peuvent être ignorées. Une réponse possible à cette problématique vient des étapes préalables à l'application des patrons, à savoir l'étiquetage morphosyntaxique. La propagation d'erreurs produites par l'étiqueteur influe sur l'extraction : un mot mal étiqueté peut aboutir à un bruit ou à un silence. Les étiqueteurs – anglais – actuels sont particulièrement performants, ne produisent que peu d'erreurs mais sont sensibles à la nature et à la qualité du texte analysé. Une erreur typographique, un mauvais résultat d'OCR (*Optical Character Recognition*) ou une ambiguité linguistique sont générateurs d'erreurs pour un étiqueteur morphosyntaxique.

Dans nos expériences sur plusieurs corpora, nous avons pu observer deux erreurs récurrentes de notre étiqueteur qui ont trait à la confusion entre certaines catégories grammaticales dans des contextes spécifiques. Plus précisément, nous avons observé (pour l'anglais) :

— La reconnaissance d'un gérondif à la place d'un adjectif. Les gérondifs désignent les formes verbales en *-ing* en anglais. Si les gérondifs jouent le même rôle que les adjectifs quand directement accolés à un nom, peu de propositions de patrons les incluent.

— La reconnaissance d'un participe passé à la place d'un adjectif. De la même manière et pour les mêmes raisons que les gérondifs, les participes passés sont générateurs d'erreurs.

Malgré le fait que la proposition d'intégrer gérondifs et participe-passé soit théoriquement et pragmatiquement justifiée, son impact reste difficile à mesurer du fait du peu de ressources de référence à disposition.

Nous avons évoqué la variété des contextes d'application d'une analyse terminologique et la même variété relative aux résultats produits. Cette diversification pose les questions de l'évaluation et de la généralisation des solutions proposées, questions pour l'instant sans réponse. L'annotation terminologique ne montre qu'un faible accord annotateur, ce qui dénote le peu de consensus sur le sujet et qui justifie le peu de ressources à disposition. L'annotation des graphies pertinentes est une tâche longue et fastidieuse qui nécessite de multiples occurrences pour une même graphie, et qui peut parfois être *corrompue* par des a priori structurels ou des contraintes ad-hoc.

Une méthode récurrente de construction de terminologie de référence consiste à extraire un nombre limité de graphies qui sont ensuite validées manuellement. Toute la problématique est véhiculée par le mot *limité* dans la phrase précédente : les chercheurs veulent épargner du temps de travail aux experts qui évaluent les termes, ils restreignent donc les graphies à celles qui sont le plus probablement terminologiques.

Pragmatiquement fondée, cette restriction peut créer un corpus de référence ad-hoc, qui ne sera généralisable qu'à peu d'autres recherches. Nous avons par exemple pu observer une campagne d'annotation d'experts sur des graphies extraites à partir d'un patron. De fait, si une catégorie grammaticale est ajoutée au patron, son impact ne pourra pas être estimé. Les limites peuvent porter de manière indifférente sur le nombre de mots de la lexie, sa fréquence, ses propriétés morphosyntaxiques ou sur un poids déterminé par une formule donnée. Dans tous les cas, l'application de cette limite induit nécessairement une proportion de silences malgré l'annotation manuelle, silence qui deviendra du bruit lors de l'évaluation d'autres patrons à partir de cette référence.

Au delà des limitations pragmatiques des annotateurs, une de leurs sources de désaccord émane de la *granulosité* de la terminologie. Le jugement de l'aspect terminologique d'une lexie se base sur le contexte d'apparition de cette dernière. Dans nos expériences sur des hiérarchies de catégories, nous avons pu observer des variations de jugement d'aspect terminologique pour une même lexie mais dans des contextes différents. L'aspect terminologique d'une graphie ne sera pas estimé de la même manière selon qu'elle appartienne à un corpus de MATHÉMATIQUES ou à un corpus d'ALGÈBRE QUANTIQUE (sous-corpus du premier), ce alors que les deux catégories partagent une partie importante de leur vocabulaire.

A partir de ces différents constats, nous avons proposé une formalisation et une terminologie répondant à différentes exigences. Nous avons pris

le temps de définir et limiter clairement des concepts simples mais distincts – *mot, terme, terme candidat, lexie, graphie, etc.* – ainsi que de formaliser le processus de construction de terminologies à partir de ce vocabulaire. Nos expériences nous ont permis de constater un manque de communication entre les différentes communautés de chercheurs – mathématiciens, linguistes, terminologues, analystes, etc. – qui aboutit à des propositions fortement déséquilibrées dans le sens de la spécialité des chercheurs. Alors qu'un analyste estime la qualité d'une terminologie par des mesures de comparaison automatique, le terminologue le fait généralement à la main. Les objectifs possibles des ressources construites ne sont pas les mêmes, leurs méthodes de construction et d'évaluation non plus. Dans nos recherches, nous avons proposé un cadre qui se veut équilibré entre les différents aspects impliqués dans la construction automatique de terminologie. Nous avons pris le temps de définir un cadre formel strict afin de nous permettre d'échanger sur nos expériences sans ambiguité : l'utilisation indifférenciée de mots comme *mots* ou *termes* prête à confusion dans notre contexte de recherche.

Nous avons tenté d'apporter des éclaircissements sur la diversité des recherches en construction de terminologie automatique. Nous avons également validé plusieurs de nos hypothèses dans ce contexte. Comme nous l'avons vu, le consensus ne règne pas parmi les chercheurs impliqués. L'état actuel des recherches ne permet pas encore de trouver de consensus sur des méthodes d'automatisation, consensus qui reste à déterminer pour les méthodes manuelles. De fait, les perspectives de recherche restent très larges pour un domaine relativement récent. Nous évoquons brièvement les perspectives générales de ce domaine de recherche avant de nous limiter à présenter des perspectives relatives à nos expériences spécifiquement, à savoir sur des hybridations possibles entre une terminologie et des algorithmes de classification.

De manière générale, les chercheurs en terminologie en sont encore à déterminer i) la nature des graphies à extraire, ii) la mesure de potentiel terminologique optimale ainsi que iii) la méthode d'évaluation la plus adaptée et la moins coûteuse.

i) Nous avons évoqué le postulat qui consiste à ne considérer que les groupes nominaux comme potentiellement terminologiques. Cette hypothèse est régulièrement remise en question, notamment par des chercheurs qui s'intéressent à la terminologie verbale. A cela s'ajoute les limites des groupes nominaux qui

varient selon les propositions : il n'y a pas de consensus sur des limites structurelles précises. La première perspective pour une automatisation consensuelle, strictement linguistique, consisterait à (re)définir la notion d'aspect terminologique et de ce qu'elle implique relativement à la nature du terme considéré – un adjectif est à considérer avec un nom, un verbe avec ses acteurs, etc.

ii) De nombreux chercheurs ont proposé des métriques de potentiel terminologique et d'autres chercheurs en ont fait des analyses comparatives. Si certaines métriques se démarquent pour leur efficacité, aucune ne remplit encore tous les critères qualitatifs : les meilleures sont basées sur les fréquences et ignorent de fait les termes peu fréquents. Des chercheurs ont proposé des combinaisons de métriques pour pallier cet écueil ; la nature et la méthode de combinaison des métriques retenues donnent lieu à de nombreuses recherches encore aujourd'hui.

iii) Nous avons développé les difficultés liées à l'évaluation d'une terminologie. Aucune solution pratique n'a encore été proposée pour évaluer une terminologie sans biais, de manière objective. Les évaluations de terminologies sont souvent limitées à des sous-ensembles de termes, ce qui n'est pas satisfaisant. La méthode d'évaluation que nous avons employée présente les mêmes biais que celles proposées dans la littérature, nous avons été confrontés aux mêmes limites pragmatiques que les autres chercheurs. Un protocole d'évaluation strict reste à déterminer, protocole qui doit permettre d'évaluer les évolutions des différentes étapes de la construction de terminologie – changements de patrons, de métriques, de conflations, etc.

Concernant nos expériences avec les modèles de thèmes et algorithmes de classification, diverses perspectives de recherche sont envisageables. S'il y a déjà eu de nombreuses recherches sur l'exploitation de lexies complexes – de plusieurs mots – pour la classification, il n'en existe à notre connaissance aucune concernant l'exploitation de connaissances strictement terminologiques. De manière générale, nos résultats sont à confirmer à partir d'autres corpora, idéalement plus fournis.

Nous avons identifié plusieurs axes de recherche en lien avec nos expériences sur les mesures de corrélations entre des mesures de similarité et la qualité des modèles. Nos résultats ont montré une forte corrélation entre si-

milarité textuelle et qualité de modèle de thèmes et nous avons pu déterminer les mesures les plus pertinentes – i.e. les plus corrélées. Nous avons retenu trois cas distincts d'exploitation de ces métriques qui restent à expérimenter : prédire la qualité d'un modèle a priori, évaluer la qualité du modèle a posteriori, ou en usage interne dans l'algorithme choisi – seuil de convergence, limite, etc. De plus, si nous avons observé de très fortes corrélations en l'état, davantage de conflation de la variation dans l'espace de représentation devrait mener à un gain de performance. Nos résultats peuvent donc être améliorés en incorporant des traitements linguistiques plus sophistiqués pour l'identification des variantes graphiques d'une même lexie/d'un même concept.

D'un point de vue plus linguistique et concernant la structure même des graphies terminologiques, nos expériences ont porté sur l'extension de méthodes pré-existantes. Comme nous l'avons évoqué plus haut, le choix d'une méthode d'extraction n'est qu'assez peu consensuel. Nous nous sommes basés sur une étude comparative afin de sélectionner la plus performante, puis nous l'avons modifiée relativement à nos observations sur divers corpora. Nous avons tenté d'évaluer l'impact de nos modifications à partir d'une terminologie de référence, mais les biais rencontrés – du fait de la nature de la référence – ont rendu l'évaluation non pertinente. Il serait nécessaire de procéder à une nouvelle étude comparative à partir d'une autre terminologie de référence non biaisée afin d'estimer précisément notre apport, ce qui reviendrait à ouvrir la problématique de la construction d'une terminologie non biaisée.

Le biais dans la construction d'une terminologie peut prendre différentes formes, qui constituent fréquemment des obstacles à la réutilisation de la ressource. Le biais peut être sémantique, interprétatif – limites de l'aspect terminologique – mais également structurel. Quand seules certaines structures linguistiques sont acceptées, l'évaluation de nouvelles structures est rendue impossible. Plus simplement, les contraintes de structures font que la ressource annotée n'est pas exhaustive, la *vérité* n'est pas connue. Bien que justifiée sur le plan théorique et partiellement validée par nos expériences, notre proposition d'amélioration requiert une référence plus souple, plus complète, qui permette l'évaluation des nouveaux éléments extraits.

Enfin, il serait pertinent de prévoir une expérience sur l'intérêt des modèles de thèmes pour la désambiguisation de termes avec la même graphie. Un modèle flou comme LDA pourrait théoriquement construire une distribu-

tion de sens pour une graphie terminologique donnée à partir d'un corpus, et donc permettre la désambiguisation de termes polysémiques.

Bibliographie

[Aldous, 1985] Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

[Alonso et al., 2017] Alonso, H. M., **Delamaire, Amaury**, and Sagot, B. (2017). Annotating omission in statement pairs.

[Benavent and Parrilla, 2006] Benavent, P. and Parrilla, S. (2006). Análisis de la extracción automática de términos con el programa informático ExtraTerm. Technical report, Universitat Jaume I, Valencia.

[Blei and Lafferty, 2009] Blei, D. M. and Lafferty, J. D. (2009). Visualizing topics with multi-word expressions. *arXiv preprint arXiv :0907.1013*.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, 3 :993–1022.

[Boulaknadel et al., 2008] Boulaknadel, S., Daille, B., and Driss, A. (2008). Acabit : Un outil d'extraction des termes complexes. In *Acte du Colloque sur les Nouvelles Technologies d'Information : Opportunités Pour Lamazighe, Ircam*.

[Bourigault, 1992] Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pages 977–981. Association for Computational Linguistics.

[Bourigault et al., 1996] Bourigault, D., Gonzalez-Mullier, I., and Gros, C. (1996). Lexter, a natural language processing tool for terminology extraction. In *Proceedings of the 7th EURALEX International Congress*, pages 771–779.

[Bourigault and Jacquemin, 1999] Bourigault, D. and Jacquemin, C. (1999). Term extraction-i-term clustering : An integrated platform for computer-

aided terminology. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

[Brill, 1992] Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 152–155. Association for Computational Linguistics.

[Calberg-Challot, 2007] Calberg-Challot, M. (2007). Quand un vocabulaire de spécialité emprunte au langage courant : le nucléaire, étude de cas. *Cahier du CIEL*, 2008 :2007–2008.

[Cartier, 2017] Cartier, E. (2017). Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 95–98.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1) :37–46.

[Condamines and Rebeyrolle, 2000] Condamines, A. and Rebeyrolle, J. (2000). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. *Ingénierie des Connaissances, évolutions récentes et nouveaux défis. Paris : Eyrolles*, pages 225–242.

[Dagan and Church, 1994] Dagan, I. and Church, K. (1994). Termight : Identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language pr ocessing*, pages 34–40. Association for Computational Linguistics.

[Daille, 1994] Daille, B. (1994). Study and implementation of combined techniques for automatic extractio n of terminology. *The balancing act : Combining symbolic and statistical approaches to l anguage*.

[David and Plante, 1990] David, S. and Plante, P. (1990). De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence artificielle et sciences cognitives au Québec*, 3(3) :140–154.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent semantic analysis. *Journal of the American society for information science*, 41(6) :391.

[Delobelle et al., 2020] Delobelle, J., Cabrio, E., Villata, S., **Delamaire, Amaury**, and Ruti, R. (2020). Argument mining for fake news detection. In *Proceedings of the 8th International Conference on Computational Models of Argument*.

[Drouin, 2003] Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1) :99–115.

[Drouin and Doll, 2008] Drouin, P. and Doll, F. (2008). Quantifying termhood through corpus comparison. *Proceedings of Terminology and Knowledge Engineering (TKE-2008)*, pages 191–206.

[Dunn, 1973] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

[Enguehard and Pantera, 1995] Enguehard, C. and Pantera, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics*, 2(1) :27–32.

[Fort, 2014] Fort, K. (2014). Annotation collaborative de corpus : Dimensions de complexité.

[Fort, 2016] Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*. John Wiley & Sons.

[Francis and Kucera, 1964] Francis, W. N. and Kucera, H. (1964). Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, 1.

[Frantzi et al., 1998] Frantzi, K. T., Ananiadou, S., and Tsujii, J. (1998). The c-value/nc-value method of automatic recognition for multi-word terms. In *International Conference on Theory and Practice of Digital Libraries*, pages 585–604. Springer.

[Handschuh and QasemiZadeh, 2014] Handschuh, S. and QasemiZadeh, B. (2014). The acl rd-tec : a dataset for benchmarking terminology extraction and classification in computational linguistics. In *COLING 2014 : 4th international workshop on computational terminology*.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic Latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

[Justeson and Katz, 1995] Justeson, J. S. and Katz, S. M. (1995). Technical terminology : some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(01) :9–27.

[Kageura and Umino, 1996] Kageura, K. and Umino, B. (1996). Methods of automatic term recognition : A review. *Terminology. International Journal*

of Theoretical and Applied Issues in Specialized Communication, 3(2) :259–289.

[Laham, 1997] Laham, D. (1997). Latent semantic analysis approaches to categorization. In *Proceedings of the 19th annual conference of the Cognitive Science Society*, page 979.

[Lau et al., 2011] Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics.

[Lau et al., 2014] Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves : Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

[L'Homme, 2012] L'Homme, M.-C. (2012). Le verbe terminologique : un portrait de travaux récents. In *SHS Web of Conferences*, volume 1, pages 93–107. EDP Sciences.

[Liu et al., 2010] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916. IEEE.

[Lykke et al., 2010] Lykke, M., Larsen, B., Lund, H., and Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In *European Conference on Information Retrieval*, pages 627–630. Springer.

[Manning, 2011] Manning, C. D. (2011). Part-of-speech tagging from 97% to 100% : is it time for some linguistics ? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

[Manning et al., 2014a] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014a). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

[Manning et al., 2014b] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014b). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

[Mehrotra et al., 2013] Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM.

[Mei et al., 2007] Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM.

[Meyer and Mackintosh, 2000] Meyer, I. and Mackintosh, K. (2000). L'étirement du sens terminologique : aperçu du phénomène de la déterminologisation. *Le sens en terminologie*, pages 198–217.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mikolov et al., 2013c] Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

[Navigli and Velardi, 2002] Navigli, R. and Velardi, P. (2002). Semantic interpretation of terminological strings. In *Proc. 6th Int'l Conf. Terminology and Knowledge Eng*, pages 95–100.

[Newman et al., 2009] Newman, D., Karimi, S., and Cavedon, L. (2009). External evaluation of topic models. In *in Australasian Doc. Comp. Symp., 2009*. Citeseer.

[Newman et al., 2010] Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

[Nokel and Loukachevitch, 2016] Nokel, M. and Loukachevitch, N. (2016). Accounting ngrams and multi-word terms can improve topic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 44–49.

[Park and Byrd, 2001] Park, Y. and Byrd, R. J. (2001). Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 conference on empirical methods in natural language processing.*

[Park et al., 2002] Park, Y., Byrd, R. J., and Boguraev, B. K. (2002). Automatic glossary extraction : beyond terminology identification. In *COLING 2002 : The 19th International Conference on Computational Linguistics.*

[Paryzek, 2008] Paryzek, P. (2008). Comparison of selected methods for the retrieval of neologisms. 16 :163–181.

[Pazienza et al., 2005] Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M. (2005). Terminology extraction : an analysis of linguistic and statistical approaches. In *Knowledge mining*, pages 255–279. Springer.

[Peñas et al., 2001] Peñas, A., Verdejo, F., Gonzalo, J., et al. (2001). Corpus-based terminology extraction applied to information access. In *Proceedings of Corpus Linguistics*, volume 2001, page 458. Citeseer.

[Rand, 1971] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336) :846–850.

[Roche et al., 2004] Roche, M., Heitz, T., Matte-Tailliez, O., and Kodratoff, Y. (2004). Exit : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT*, volume 4, pages 946–956.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65.

[Sablayrolles, 2010] Sablayrolles, J.-F. (2010). Neologia : un dictionnaire néologique sous forme de base de données.

[Saneifar et al., 2009] Saneifar, H., Bonniol, S., Laurent, A., Poncelet, P., and Roche, M. (2009). Terminology extraction from log files. In *International Conference on Database and Expert Systems Applications*, pages 769–776. Springer.

[Schultz and Liberman, 1999] Schultz, J. M. and Liberman, M. (1999). Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of the DARPA broadcast news workshop*, pages 189–192. San Francisco : Morgan Kaufmann.

[Schütze et al., 2008] Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press.

[**Delamaire, Amaury** et al., 2019a] **Delamaire, Amaury**, Beigbeder, M., and Juganaru-Mathieu, M. (2019a). Exploitation de syntagmes dans la découverte de thèmes. In *CORIA*.

[**Delamaire, Amaury** et al., 2019b] **Delamaire, Amaury**, Juganaru-Mathieu, M., and Beigbeder, M. (2019b). Correlation between textual similarity and quality of lda topic model results. In *2019 13th International Conference on Research Challenges in Information Science (RCIS)*, pages 1–6. IEEE.

[**Delamaire, Amaury** et al., 2019c] **Delamaire, Amaury**, Juganaru-Mathieu, M., and Beigbeder, M. (2019c). Extension d'une méthode automatique d'extraction terminologique : étude de cas sur un corpus spécialisé. In *Conférence Toth 2019*.

[**Delamaire, Amaury** et al., 2020] **Delamaire, Amaury**, Juganaru-Mathieu, M., and Beigbeder, M. (2020). Contributions sur la structure morphosyntaxique des graphies terminologiques et sur l'hybridation entre terminologie et modèles de thèmes. In *Conférence Toth 2020*.

[Velardi et al., 2001] Velardi, P., Missikoff, M., and Basili, R. (2001). Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001*, page 5. Association for Computational Linguistics.

[Vergne, 2003] Vergne, J. (2003). Un outil d'extraction terminologique endogène et multilingue. *Actes de TALN*, 2 :139–148.

[Wallach et al., 2009] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM.

[Wang et al., 2007] Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams : Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 697–702. IEEE.