Traitement Automatique de la Langue Biomédicale

Aurélie Névéol

Mémoire présenté pour l'obtention de l'

HABILITATION À DIRIGER DES RECHERCHES

Spécialité Informatique

Traitement Automatique de la Langue Biomédicale

Présenté par :
Aurélie NÉVÉOL

Jury :
Marc Cuggia EXAMINATEUR
Cédrick Fairon RAPPORTEUR
Christine Froidevaux EXAMINATRICE
Emmanuel Morin RAPPORTEUR
Lynda Tamine Lechani RAPPORTRICE
Pierre Zweigenbaum PARRAIN

Soutenu publiquement le 26 Novembre 2018

Résumé

Ce mémoire aborde le Traitement Automatique de la Langue Biomédicale (ou TAL biomédical), un champ de recherche pluri-disciplinaire qui mobilise l'Informatique, la Linguistique ainsi que la Médecine. Cette thématique s'inscrit dans le champs du traitement automatique de la langue, tout en allant au delà du service rendu à la médecine. Ainsi je commence par présenter différentes facettes fondamentales et appliquées du TAL biomédical. J'ai ensuite choisi de développer dans ce mémoire trois thématiques qui ont fait l'objet de mon travail ces dernières années : la modélisation des informations, l'analyse de textes en langue de spécialité et des cas concrets d'application biomédicales. Je présente tout d'abord le développement de ressources en soutien du TAL biomédical, en particulier pour les langues autres que l'anglais comparativement peu dotées. Ce travail s'appuie sur une analyse des schémas de représentation des connaissances dans le domaine, qui a permis le développement de corpus annotés destinés à être partagés par la communauté à des fins d'évaluation. Ce mémoire aborde ensuite les méthodes d'analyse de textes médicaux. L'extraction d'entités et de relations montre l'importance de la question de l'adaptation en domaine et de l'adaptation cross-langue. Enfin, une dernière partie discute l'impact attendu du traitement automatique de la langue en épidémiologie, santé publique et sur les pratiques de recherche au delà de ces disciplines. L'un des défis du TAL biomédical est de réaliser pleinement ce potentiel en devenant un levier incontournable de la recherche translationnelle.

Abstract

My research addresses Biomedical Natural Language Processing (or bioNLP), a field of research that draws on multiple disciplines including Computer Science, Linguistics and Medicine. BioNLP as a field offers a contribution to Natural Language Processing that goes beyond mere applications to biomedicine. After introducing different fundamental and applied aspects of bioNLP, this thesis explores three topics that I have been particularly interested in over the years : information modeling, analysis of texts in specialized domains and concrete cases of biomedical applications. First, I present the development of resources in support of bioNLP, especially for the comparatively low resourced languages other than English. This work is based on an analysis of knowledge representation patterns in the field, which has contributed to the development of annotated corpora to be shared by the community for evaluation purposes. I next present some methods used for the automatic understanding of medical texts. The extraction of entities and relations between them shows the importance of domain adaptation and cross-language adaptation. Finally, I discuss the expected impact of natural language processing in epidemiology, public health and research practices beyond these disciplines. One of the challenges of bioNLP is to offer actionable tools and methods in order to become an essential component of translational research.

Acknowledgments

Writing this manuscript has been a challenge. It is also an opportunity to look back over the past fifteen years, reflect on accomplishments and milestones to look forward to. As stressful as it was, it has been a useful process which will help me shape the next stage of my research.

I would like to warmly acknowledge the computational linguistics and medical informatics communities where I found a professional home with opportunities to discuss exciting topics and grow as a researcher. My research has been conducted through collaboration with many colleagues at the National Library of Medicine, at LIMSI and other places - including talented post-doctoral fellows, PhD and master students I had the opportunity to co-supervise. Thank you for the fruitful collaborations and discussions about science, ethics and research. It has been a pleasure working with you all !

I have been fortunate to receive funding from the French *Agence Nationale pour la Recherche* (including Labex Digicosme) and European H2020 programme, fostering partnerships with institutes such as CEA, Inserm and hospitals.

I would like to express my appreciation and gratitude to the administrative and computer support staff who ensure things run smoothly on a day to day basis and help us navigate server access, recruiting, grant applications, travel and other purchase procedures.

Last but not least, thanks also go to my family for their loving support : my husband Pascal, my dear sons Maxime and William, my parents, my brother and my in-laws.

Table des matières

1 Introduction **1**
- 1.1 Contexte général et thématiques de recherche . . . 2
- 1.2 Spécificité du traitement automatique de la langue biomédicale . . . 5
- 1.3 Organisation du mémoire . . . 6

2 Modélisation des informations dans le domaine biomédical **7**
- 2.1 Représentation des connaissances . . . 8
  - 2.1.1 Méthodologie pour la modélisation . . . 9
  - 2.1.2 Modèles proposés . . . 9
- 2.2 Création de corpus annotés . . . 14
  - 2.2.1 Corpus de requêtes PubMed . . . 14
  - 2.2.2 Corpus clinique du français MERLoT . . . 16
  - 2.2.3 Corpus biomédical du français QUAERO . . . 16
- 2.3 Organisation de campagnes d'évaluation . . . 19
  - 2.3.1 Mesures d'évaluation . . . 21
  - 2.3.2 Extraction d'information multilingue à CLEF eHealth 2015-2018 . . . 22
  - 2.3.3 Codage des causes de décès en anglais, français, hongrois, italien. . . . 23
  - 2.3.4 Reproductibilité. . . . 25
  - 2.3.5 Traduction automatique de textes biomédicaux à WMT 2016-2018 . . . 25
- 2.4 Discussion . . . 26

3 Méthodes pour l'analyse de textes biomédicaux **28**
- 3.1 Revues de la littérature sur le traitement automatique de la langue clinique 30
- 3.2 Analyse linguistique . . . 33
- 3.3 Extraction d'entités . . . 37
- 3.4 Extraction de relations . . . 39
- 3.5 Discussion . . . 41

4 Impact en épidémiologie et santé publique **43**
- 4.1 Recherche documentaire . . . 43
  - 4.1.1 Contribution à la constitution de bases de données . . . 43
  - 4.1.2 Recherche d'information à partir d'un moteur de recherche . . . 46
  - 4.1.3 Recherche d'information à partir du dossier patient . . . 50
- 4.2 Analyse rétrospective des dossiers patients . . . 50
  - 4.2.1 Désidentification . . . 50
  - 4.2.2 Analyse temporelle . . . 54
  - 4.2.3 Extraction d'information précise et classification de documents . . . 57
  - 4.2.4 Discussion . . . 59

5 Conclusion et Perspectives **60**
5.1 Conclusion 60
5.2 Perspectives 61
5.2.1 Partage de ressources pour le français biomédical 61
5.2.2 Vers des modèles génériques ou adaptables ? 61
5.2.3 Intégration de l'expertise biomédicale 61

Annexes **76**

Annexe 1 **76**
1 Résultats des participants aux campagnes CLEF eHealth (2015-2018) 76
2 Résultats des participants à la tâche de traduction biomédicale de WMT (2016-2018). 76

Chapitre 1

Introduction

Ce document présente des travaux de recherche que j'ai effectués au sein de la Faculté de Médecine de Rouen, de la National Library of Medicine (Lister Hill National Center for Biomedical Communications - LHNCBC, puis National Center for Biomedical Informatics - NCBI) et du Laboratoire Interdisciplinaire pour la Mécanique et les Sciences de l'Information (LIMSI - UPR 3251) depuis la fin de ma thèse en Novembre 2005. Ces travaux ont abordé le traitement automatique de la langue biomédicale, et plus particulièrement la compréhension automatique de textes biomédicaux.

J'envisage la compréhension automatique de textes du domaine biomédical comme la traduction d'un texte en langue naturelle en une représentation structurée lisible par la machine, et exploitable pour des applications en recherche et en santé publique. Cette approche fait l'hypothèse qu'il est possible de définir un système de représentation structurée capable de modéliser les informations médicalement pertinentes contenues dans les textes biomédicaux. Cette représentation doit être à la fois suffisament détaillée pour capturer une grande quantité d'information et suffisamment générique pour pouvoir s'adapter à la diversité des textes du domaine. J'envisage une approche pragmatique afin de définir la représentation à utiliser, qui doit à la fois refléter les informations contenues dans les textes et être guidée par les applications concrètes afin d'être exploitable dans ce contexte.

Comme illustré par une chronologie de publications présentée en Figure 1.1, j'ai développé quatre grandes thématiques tout au long de cette période : il s'agit tout d'abord de la *modélisation* des informations contenues dans les textes d'un domaine de spécialité, le domaine biomédical. Ce travail débouche naturellement sur le *développement de ressources* qui illustrent l'adéquation des modèles avec les textes visés. Les *méthodes d'analyse de texte* s'appuient sur les ressources ainsi développées. Ces méthodes peuvent être mises en oeuvre et *évaluées en vue d'applications du traitement automatique de la langue en épidémiologie et santé publique*.

Finalement, une dernière thématique de recherche transverse a émergé depuis 2016 suite à une conjonction de circonstances que sont ma participation à l'organisation des campagnes d'évaluation CLEF [1] eHealth et WMT biomédical [2], au projet européen MIRoR (Methods in Research on Research) [3] et au collectif Ethique et TAL [4] : il s'agit de la reproductibilité en traitement automatique de la langue.

1. Cross Language Evaluation Forum `http://www.clef-initiative.eu`
2. Workshop on Machine Translation ; pour la dernière édition de la tâche biomédicale, voir `http://www.statmt.org/wmt18/biomedical-translation-task.html`
3. `http://miror-ejd.eu/`
4. `http://www.ethique-et-tal.org`

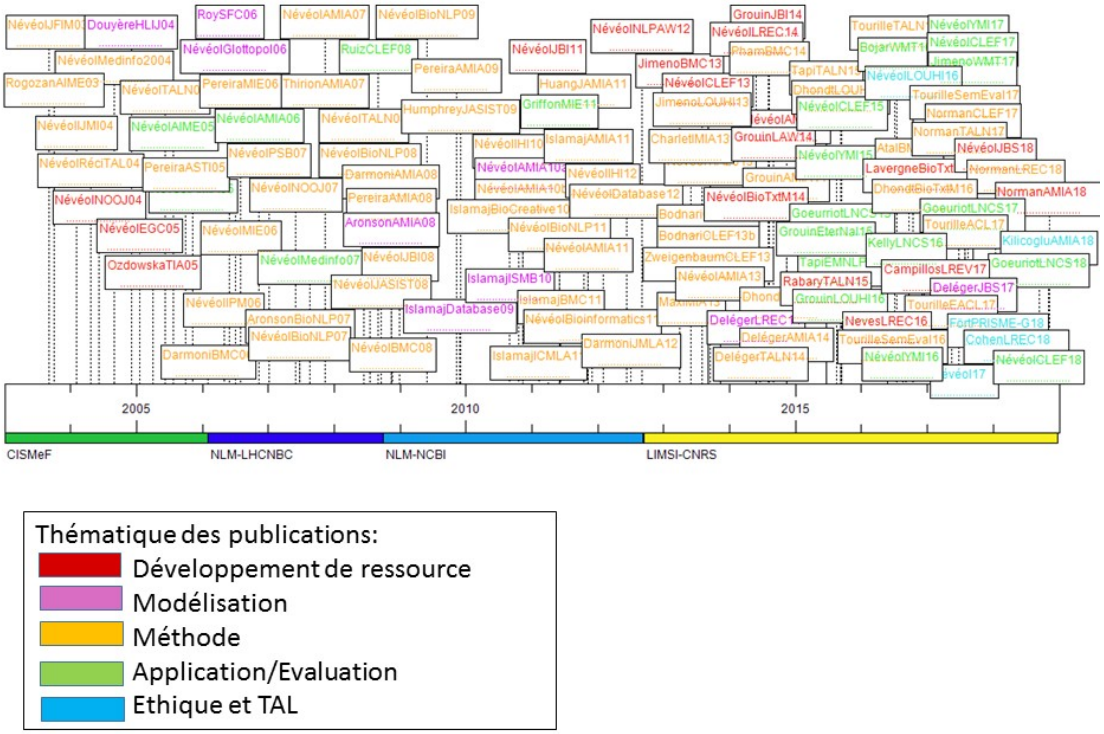


FIGURE 1.1 – Chronologie des publications par thématique (2003-2018)

1.1 Contexte général et thématiques de recherche

Dans tous les domaines de la connaissance, un nombre important d'informations sont contenues dans des textes libres et ne sont pas directement accessibles à des fins de traitement automatique. Dans le domaine biomédical, bien que de nombreuses données de biologie moléculaire soient disponibles dans des bases de données librement accessibles, nombre d'informations cruciales restent contenues dans du texte brut, par exemple dans des articles de la littérature. De même, une étude récente a montré l'importance du texte des comptes rendus hospitaliers comme source d'information clinique : certains phénotypes (des caractéristiques observables ou mesurables) sont renseignés exclusivement dans des textes libres pour 80 à 90 % des patients [Escudié et al., 2017]. Face à la complexité du langage naturel, la notion de domaine et de genre textuel a été largement étudiée [Biber, 1988], conduisant à l'hypothèse de l'existence de sous-langages [Harris, 1991]. Quelques travaux ont même cherché à établir un découpage de certains domaines de spécialité en divers sous-langages [Friedman et al., 2002, Lippincott et al., 2011], soulevant la question du degré de spécialisation de chacun des sous-domaines identifiés et par suite du degré de spécialisation nécessaire pour développer des outils d'analyse de texte performants en langue de spécialité.

Langue de spécialité. La définition de la notion de *langue de spécialité* est très liée à la notion de *domaine de spécialité* [Poibeau, 2005]. Ainsi, la langue de spécialité sert spécifiquement à véhiculer des informations relatives au domaine concerné. Ce constat ne

permet pas vraiment de distinguer la langue de spécialité de la langue générale, dans la mesure où des notions relevant d'un domaine de spécialité peuvent être discutées dans une grande variété de contextes. Pour la langue biomédicale par exemple, on peut envisager un discours scientifique précis et technique tout comme un discours relevant de la vulgarisation ou du dialogue médecin/patient. [Charnock, 1999] propose de laisser de coté la langue de spécialité pour s'intéresser plutôt à la notion de *langue technique* qui serait plus facilement distinguable de la langue générale par des caractéristiques linguistiques au delà du vocabulaire spécialisé. Globalement, l'ensemble des auteurs reconnaissent qu'il y a toujours une zone de recouvrement entre divers sous langages, ou entre un sous langage donné et la langue générale. En réalité, il existe donc un *continuum* entre la langue générale et les diverses formes de langue de spécialité. On retiendra cependant les éléments suivants pour caractériser la langue de spécialité :

— sur le plan sémantique, l'utilisation d'une terminologie propre au domaine qui met en évidence des catégories d'objets et concepts prévalents, l'omission fréquente d'informations et l'utilisation de présupposés (réalisés par exemple dans l'utilisation d'abréviations et d'acronymes) ;

— sur le plan distributionnel, les co-occurrences entre termes différents de la langue générale et les énoncés s'appuient sur des patrons syntaxiques particuliers ;

Informations véhiculées par les textes en domaine de spécialité. La grande disponibilité de ressources terminologiques et textuelles du domaine biomédical en fait un domaine de prédilection pour l'étude de la notion même de langue de spécialité et de sous domaine. Par ailleurs, le domaine biomédical fournit de nombreuses applications pratiques permettant au traitement automatique de la langue d'avoir un impact significatif sur la pratique clinique et la recherche biomédicale. Les informations cruciales de ce domaine sont contenues principalement dans des articles de la littérature ou dans les dossiers électroniques patient – définissant les deux sous genres identifiés par [Friedman et al., 2002] en combinant une analyse distributionnelle à la sagacité d'experts du domaine. Les textes rédigés par les patients notamment dans le cadre de leur activité sur les réseaux sociaux constituent une autre source d'information très étudiée [Gonzalez-Hernandez et al., 2017]. Pour l'ensemble de ces textes, je parlerai de *texte libre*, par opposition aux *données structurées* que peuvent constituer les autres informations contenues dans le dossier électronique patient (nom, dates de séjour, codage médico-économique, ...) ou les métadonnées concernant les articles scientifiques (titre de la revue, type de publication, mots-clés, ...). On peut néanmoins remarquer que certains de ces textes libres peuvent être rédigés de manière organisée ou structurée avec des types de contenus voire des titres de section plus ou moins standardisés.

La prévalence de ces textes crée un besoin fort en méthodes et outils de Traitement Automatique de la Langue Naturelle permettant d'extraire des données pertinentes de divers types de textes du domaine biomédical : littérature, dossiers patients, productions langagières des patients, ainsi que d'autres sources telles que les textes libres contenus dans certaines bases de données biologiques. Les données ainsi extraites doivent êtres formalisées et stockées dans un format accessible à la fois par la machine afin de permettre des traitements avancés et par l'homme afin de contribuer au partage des informations au travers de disciplines et de contextes différents. Ce type de partage de l'information est bénéfique à deux niveaux : 1/il permet d'améliorer la dissémination des connaissances auprès des chercheurs, des professionnels de santé et du grand public et 2/il améliore la disponibilité des informations auprès des chercheurs au-delà de leur strict domaine de spécialité, ce qui

permet d'accélérer les découvertes scientifiques et de faire avancer la recherche pluridisciplinaire, par exemple la médecine personnalisée [Altman, 2011].

Traitement automatique de la langue de spécialité. Depuis plusieurs décennies, de nombreuses méthodes issues du Traitement Automatique de la Langue Naturelle (TAL) ont été exploitées par les acteurs du domaine biomédical afin de répondre à un besoin croissant de traitement de documents spécialisés ([Collier et al., 2006, Zweigenbaum et al., 2007, Demner-Fushman et al., 2009, Chapman and Cohen, 2009] inter-alia). Le traitement automatique de la langue biomédicale s'appuie sur des ressources spécifiques de représentation des connaissances telles que l'UMLS (Unified Medical Language System). L'UMLS est une ressource terminologique développée par la National Library of Medicine qui contient plusieurs millions de termes du domaine biomédical, liés à 1.5 millions de concepts issus de plus de 60 familles de terminologies biomédicales [Lindberg et al., 1993].

Ce système de représentation des connaissances a donné lieu au développement de ressources terminologiques supplémentaires, au développement de bases de données et à divers types d'analyses de texte. Le projet « Indexing Initiative » de la National Library of Medicine a produit les premiers outils intégrant analyse syntaxique et sémantique pour le domaine biomédical disponibles librement : il s'agit de MetaMap [Aronson and Lang, 2010, Demner-Fushman et al., 2017] qui permet d'extraire des concepts UMLS à partir de textes libres et de MTI [Aronson et al., 2004, Mork et al., 2017] un outil d'indexation permettant d'associer des descripteurs MeSH à des articles de la littérature. Ces outils ont par la suite ouvert la voie au développement d'applications plus complexes, comme par exemple l'extraction de relations entre concepts UMLS réalisée par SemRep [Rindflesch and Fiszman, 2003], ou la détection d'informations contextuelles se rapportant aux concepts réalisée par NegEx [Chapman et al., 2001], maintenant intégré dans MetaMap et d'autres outils d'analyse de textes cliniques comme cTAKES [Savova et al., 2010] ou CLAMP [Soysal et al., 2017].

La création d'une section dédiée au traitement automatique de la langue clinique dans le *Yearbook of medical Informatics* de l'IMIA (Association Internationale d'Informatique Médicale) a permis de recenser systématiquement la littérature de ce domaine depuis 2014 [Névéol and Zweigenbaum, 2015, Névéol and Zweigenbaum, 2016, Névéol and Zweigenbaum, 2017, Névéol and Zweigenbaum, 2018]. Tout comme l'existence de nombreuses sessions dédiées au Traitement Automatique de la Langue dans le symposium annuel de l'Association américaine d'informatique médicale (AMIA) qui rassemble la communauté de l'informatique médicale, cette section du Yearbook de l'IMIA montre le statut reconnu du TAL en tant que méthodologie permettant des avancées médicales. Cependant, dans la communauté du traitement automatique de la langue, le TAL biomédical bénéficie d'une reconnaissance et d'une visibilité moindres. Les travaux y sont principalement présentés au travers du workshop BioNLP adossé annuellement à l'une des conférences de l'ACL (Association for Computational Linguistics) depuis 2002, dans d'autres workshops spécialisés comme LOUHI ou Biomedical Text Mining depuis 2008. Les quelques travaux de TAL biomédical présentés dans les conférences principales de l'ACL (ACL, EACL, NAACL, EMNLP) se positionnent par rapport à une problématique de TAL générale (par exemple, analyse temporelle, classification de texte, recherche d'information) qui se trouve être appliquée avec des corpus et données issues du domaine biomédical, voire de plusieurs domaines dont le domaine biomédical.

1.2 Spécificité du traitement automatique de la langue biomédicale

Le traitement automatique de la langue biomédicale est un champ de recherche pluridisciplinaire qui s'appuie sur l'informatique, la linguistique, et la médecine. Le TAL biomédical s'intéresse aux questions fondamentales du TAL sur la modélisation de langue que sont la syntaxe, la sémantique, l'analyse du discours et le traitement des signaux langagiers tels que la parole, le geste et l'écriture manuscrite. Il comporte un aspect applicatif important qui guide la méthodologie et les réalisations. En pratique, ce pilotage a des conséquences concrètes sur les problématiques de recherche et la manière de les aborder.

Il peut s'agir d'une redéfinition des objets d'étude. Ainsi dans le cadre de l'analyse temporelle, les expressions temporelles rencontrées dans le domaine biomédical sont de même nature que dans d'autres domaines bien que la distribution en soit différente [Tapi Nzali et al., 2015]. Cependant, la notion d'événement est dénotée par une réalisation linguistique différente du domaine général, et donné lieu à l'étude de relations temporelles également différentes (notion de "conteneur narratif" [Pustejovsky and Stubbs, 2011]). Le stage de M2 de Mike Tapi Nzali puis la thèse de Julien Tourille se sont appuyé sur ces travaux.

Le domaine biomédical peut être également vu comme un domaine d'application du TAL qui se caractérise par la disponibilité de ressources termino-ontologiques importantes, et des applications permettant d'explorer des problèmes particulier sur le plan méthodologique : classification extrême, recherche d'information exhaustive.

Par exemple, les applications que sont l'indexation MeSH (à laquelle je me suis intéressée pendant ma thèse et mon post-doc) ou le codage CIM10 (à laquelle je me suis intéressée dans le cadre des campagnes CLEF eHealth 2016-2018) peuvent être traitées comme un problème de classification automatique de texte. Cependant, il s'agit de classification complexe puisque les problèmes d'échelle sont présents à la fois au niveau du nombre de classes (18 000 codes CIM10, 25 000 mots clés MeSH, 600,000 termes d'indexation MeSH, plusieurs millions de combinaisons possibles à la fois pour la CIM10 et pour le MeSH.) que du volume de documents à traiter (environ 600 000 décès par an en France ; plus d'un million de documents par an ajoutés dans MEDLINE, soit 3 000 documents par jour ouvrable à traiter). En pratique, si des approches se focalisant sur quelques "classes" très fréquentes peuvent donner des performances acceptables en termes de précision et de rappel, leur intérêt est limité dans la pratique des indexeurs et des codeurs qui sont en demande d'un soutien intelligible sur leur cœur de métier, la multitude des "classes" peu fréquentes et régulièrement non présentes dans l'historique de codage ou d'indexation qui peut servir de corpus d'entraînement.

Autre exemple : en recherche d'information biomédicale, l'importance du rappel dans le cadre du filtrage de publications pour la constitution de revues systématiques est soulignée par les spécialistes[Manchikanti, 2008]. On pourrait donc penser qu'un calibrage des systèmes favorisant le rappel suffirait à apporter une solution satisfaisante (notion de "total recall"). Cependant, la pratique des experts auteurs de revues systématiques révèle que si le rappel est important pour eux, tous les documents n'ont pas forcément un impact équivalent sur les résultats de l'analyse effectuée dans une revue systématique. Ainsi, il importe de mieux comprendre l'implication d'un niveau de rappel donné sur les conclusions médicales d'une analyse issue du corpus de documents fourni par la recherche. Ce travail est en cours de réalisation dans le cadre de la thèse de Christopher Norman.

Ces quelques exemples permettent d'illustrer le caractère pluri-disciplinaire du TAL biomédical en tant que discipline. Une partie importante du travail consiste à identifier les problématiques de traitement automatique de la langue sous-jacentes dans les applica-

tions d'analyse de texte qui intéressent les praticiens hospitaliers et les biologistes, et à en proposer une approche qui intègre les caractéristiques métiers pratiques.

1.3 Organisation du mémoire

Pour résumer cette entrée en matière, on pourra retenir que de nombreuses connaissances et informations dans le domaine biomédical sont contenues dans des textes. Ces documents, qui constituent l'objet d'étude du traitement automatique de la langue biomédicale (TAL biomédical) sont nombreux et variés. Afin d'aborder l'analyse du contenu de ces documents d'une manière générique et systématique, un enjeu est de construire une représentation sémantique formalisée du "sens" (c'est à dire, des informations et des connaissances) contenu dans les textes médicaux.

Ce travail fait l'hypothèse qu'il est possible d'analyser une variété de textes médicaux (littérature, compte-rendus cliniques, réseaux sociaux) en s'appuyant sur une telle représentation et d'en exploiter le résultat dans le cadre de nombreuses applications : phénotypage à haut débit, recrutement de patients pour des essais cliniques, recherche de "cas" similaires ou de littérature pertinente, repérage d'éléments manquants dans un dossier, etc.

Par "analyser" des textes médicaux, nous entendons : convertir les connaissances et informations contenues dans ces textes vers la représentation sémantique choisie.

Ce mémoire d'Habilitation à Diriger des Recherches résume mes travaux dans les trois principales thématiques qui ont été le fil conducteur de ma recherche. Le chapitre 2 présente mes travaux en modélisation des informations, ainsi que les corpus annotés qui en ont découlé. Le chapitre 3 aborde les méthodes d'analyse de textes biomédicaux que j'ai contribué à développer, notamment pour l'extraction d'entités et de relations. Le chapitre 4 propose des cas concrets d'application de ces méthodes pour la recherche en épidémiologie et santé publique.

Chapitre 2

Modélisation des informations dans le domaine biomédical

Pour aborder la compréhension automatique des textes biomédicaux, je fais l'hypothèse que les informations contenues dans les textes peuvent être interprétées sous forme de représentation structurée, exploitable par la machine. La problématique qui se dégage est celle de la création de formalismes permettant cette représentation pour une grande variété de textes du domaine.

La modélisation des informations fait appel à la représentation des connaissances dans des constructions abstraites qui pourront ensuite être mises en correspondance avec une représentation contextuelle en corpus. Le triangle aristotélicien (Figure 2.1) présente les différentes notions mises en œuvre dans la représentation des connaissances : les choses, c'est-à-dire les objets ou actions du monde sont décrits de manière abstraite dans un système de représentation des connaissances par les concepts. Nous renvoyons le lecteur à [Hébert, 2010] pour une discussion détaillée de ces différents éléments. Finalement, les termes sont des signes permettant de désigner les concepts en langue naturelle [Wüster, 1981].

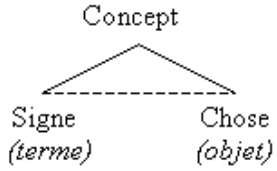


FIGURE 2.1 – Le triangle aristotélicien

Le formalisme que j'ai retenu pour cette modélisation est celui d'un réseau sémantique constitué de catégories qui peuvent être reliées à des *ressources termino-ontologiques*, c'est à dire des ressources lexicales avec divers degrés de formalisation, depuis le thésaurus jusqu'à l'ontologie. En outre, les entités du réseau sémantique peuvent également être caractérisées par des *attributs* et associées entre elles par des *relations*.

Dans ce chapitre, je présente plusieurs modèles de représentation des connaissances que j'ai contribué à développer en section 2.1, l'application de ces modèles dans des corpus annotés en section 2.2 et l'utilisation de corpus annotés dans des campagnes d'évaluation en section 2.3.

2.1 Représentation des connaissances

La modélisation des connaissances dans le domaine biomédical est une préoccupation de longue date qui fait l'objet de nombreux travaux. Par exemple, la première liste de termes destinés à catégoriser la littérature biomédicale a été publiée par la National Library of Medicine américaine dès 1954, ce qui donnera naissance au thésaurus MeSH actuel [1]. Ainsi, de nombreuses ressources sont créées pour répondre à différents besoins de représentation de l'information dans le domaine biomédical : le MeSH pour permettre la recherche d'information dans la littérature, la Classification Internationale des Maladies (CIM [2]) pour faciliter la production de statistiques de santé publique et la détermination du budget des hôpitaux grâce au codage médico-économique des dossiers patients, la SNOMED (Systematized Nomenclature of Medicine [3]) pour une description médicale systématique des dossiers patients...

Le développement de ces ressources termino-ontologiques dans les décennies qui suivent débouche sur la création de l'UMLS (Unified Medical Language System) à la fin des années 80. L'UMLS permet de regrouper un grand nombre de ces ressources et d'établir des *équivalences* entre les concepts dénotés dans les différentes sources. L'intégration de nouvelles ressources dans l'UMLS est un processus non trivial qui fait l'objet de travaux spécifiques (par exemple, [Lomax and McCray, 2004] pour Gene Ontology).

La disponibilité de ces nombreuses ressources, de nature complexe, représente à la fois une opportunité et un défi pour la modélisation des informations contenues dans les textes. Ainsi, de nombreux projets d'analyse de textes au niveau des énoncés commencent par une réflexion sur l'opportunité d'appliquer directement l'une des ressources existantes pour l'annotation de textes. Certains travaux sur l'analyse de la littérature ont utilisé le contenu de l'UMLS de manière très stricte : par exemple, [Kors et al., 2015] propose une annotation de documents biomédicaux dans des langues autres que l'anglais en projetant les concepts UMLS sur les textes en reflétant le contenu strict de l'UMLS. Ainsi, des concepts présents dans l'UMLS exprimés dans les textes par une expression non répertoriée dans la ressource ne sont pas annotés. D'autres travaux sur l'analyse du contenu des requêtes PubMed reposent sur une analyse automatique sans validation manuelle [Herskovic et al., 2007]. L'analyse de textes cliniques issue des dossiers électroniques patient a fait l'objet de différentes propositions, que nous avons passées en revue dans Deléger et al. [Deléger et al., 2017] dans le cadre de notre réflexion en amont de la proposition du schéma MERLoT.

Ces travaux explorent des questions fondamentales en représentation des connaissances et en analyse de textes :

— Quel doit être le contenu d'une ressource termino-ontologique ? On peut par exemple se demander s'il faut inclure des concepts, des termes et synonymes (c'est à dire des réalisations en langue naturelle des concepts), des relations formelles entre concepts ? Quelle doit être la granularité de la représentation ?

— Comment établir une correspondance entre le contenu d'une ressource et sa réalisation en corpus ? Faut-il orienter le contenu de la ressource en fonction de cette problématique ?

— Comment développer efficacement des corpus annotés de qualité à partir d'un modèle déterminé ?

1. `https://www.nlm.nih.gov/mesh/intro_hist.html`
2. `http://www.icd10.ch/index.asp`
3. `http://www.snomed.org/`

2.1.1 Méthodologie pour la modélisation

Mon premier contact avec le traitement automatique de la langue biomédicale s'est fait dans le cadre de ma thèse qui a porté sur l'indexation d'articles de la littérature à l'aide de descripteurs issus du thésaurus MeSH (Medical Subject Headings), une ressource termino-ontologique intégrée dans l'UMLS (Unified Medical Language System) [Lindberg et al., 1993]. Cette expérience m'a permis d'acquérir une bonne connaissance des ressources de modélisation des connaissances dans le domaine biomédical et de comprendre l'intérêt de ces outils pour la modélisation des informations.

Ainsi, après ma thèse, j'ai abordé la modélisation des informations contenues dans les textes biomédicaux dans le cadre de schémas structurés, s'appuyant sur des formalismes existants comme par exemple l'UMLS, les groupes sémantiques de l'UMLS [McCray et al., 2001] ou timeML [Pustejovsky et al., 2003] pour la modélisation temporelle.

Le principe de développement des schémas a été celui d'aller/retours entre les structures de représentation des connaissances biomédicales existantes, l'application envisagée pour le schéma développé et les textes visés par ces applications. Ainsi, le travail s'est appuyé sur des ressources existantes (par exemple, l'UMLS, la littérature proche), sur des experts du domaine d'application (par exemple, des spécialistes de recherche d'information biomédicale pour le schéma PubMed, des médecins pour le schéma MERLoT). Ce mode de développement itératif permet de concilier les fondements théoriques de représentation des connaissances du domaine biomédical, les aspects pratiques de l'application visée, et la confrontation de ces dimensions avec la réalité de terrain présente dans les textes choisis pour l'étude.

2.1.2 Modèles proposés

La table 2.1 présente un récapitulatif de trois schémas différents développés pour la modélisation d'informations biomédicales dans le cadre de mes travaux. Le schéma PubMed [Névéol et al., 2011] a été conçu pour caractériser le type d'information recherchée par les utilisateurs de PubMed afin de déterminer quelles bases de données de la NLM, au delà de PubMed, étaient susceptible de contenir ces informations. Le schéma QUAERO[Névéol et al., 2014c] visait à caractériser les concepts présents dans divers genres de textes médicaux et le schéma MERLoT [Deléger et al., 2017] avait pour but de proposer une modélisation fine et contextuelle des informations contenues dans les corpus cliniques.

Schéma / **Caractéristique**	PubMed	QUAERO	MERLoT
Entités	Oui (N=15)	Oui (N=10)	Oui (N=13)
Normalisation	Non	Oui	Non
Relations	Non	Non	Oui
Attributs	Abrévations	Non	Oui

Table 2.1 – Tableau récapitulatif des schémas et de leurs caractéristiques

Les trois schémas présentés s'appuient sur les types sémantiques et les groupes sémantiques de l'UMLS de différentes façons.

Le schéma PubMed comprend 6 catégories qui sont des groupes sémantiques ou des rassemblements de groupes sémantiques, 2 catégories qui résultent du partitionnement d'un groupe sémantique, 3 catégories qui sont des types sémantiques de l'UMLS et 4 catégories d'entités bibliographiques non couvertes par l'UMLS, et plutôt caractéristiques de la *recherche d'information* biomédicale. La table 2.2 présente et illustre les différentes catégories.

Le schéma QUAERO est le plus proche de l'UMLS. En effet, il utilise directement 10 groupes sémantiques de l'UMLS pour l'annotation de mentions et le choix de concepts pour la normalisation. La figure 2.2 présente ces concepts.

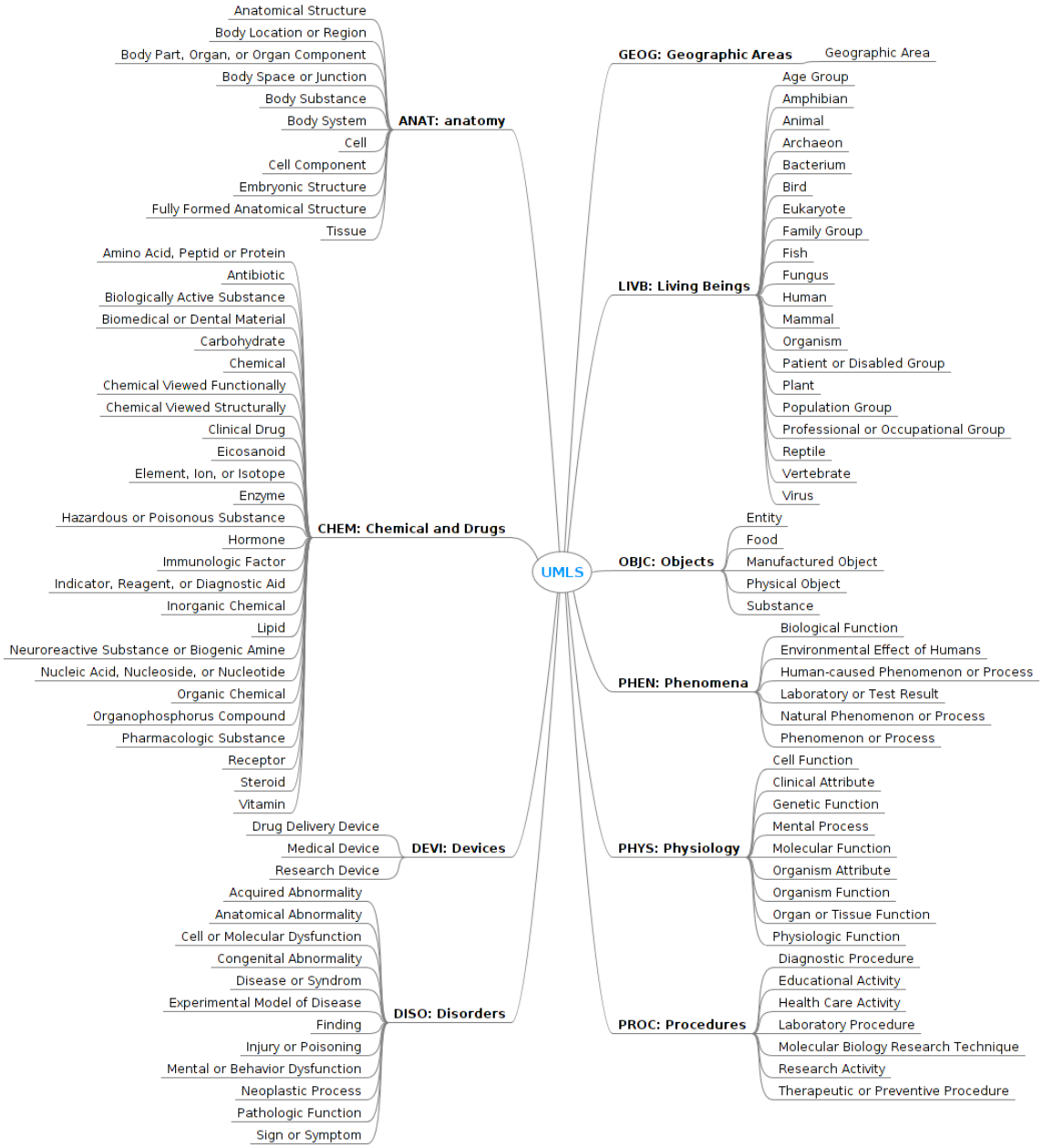


FIGURE 2.2 – Schéma d'annotation QUAERO fondé sur 10 groupes sémantiques de l'UMLS. Les codes à quatre lettres habituellement utilisés [4] pour désigner les groupes sémantiques sont explicités.

	Catégorie	Définition Courte	Exemples
⊂ ANAT	Body Part	Toute partie du corps humain, membre	finger, lung,...
	Cell or Cell Component	Type de cellule ou partie d'une cellule	Stem cell, membrane, nucleus
	Tissue	Groupe de cellules spécialisées	Abdominal muscle
	CHEM : Chemicals and Drugs	Antibiotique, médicament ou autre substance pharmacologique	Aspirin, methamphetamine, lithium, calcium
	DEVI : Devices	Objet utilisé à des fins de recherche, diagnostique ou thérapie	Adhesive bandage, insulin syringe
	DISO : Disorders	Maladie, syndrome blessure, ...	Diabetes, alcoholism, ankle fracture
	GENE : Genes, Proteins and Molecular Sequences	Nom de toute séquence moléculaire	P450, lck c-Myb transcription factor
	LIVB : Living Beings	Animal, humain, organisme	alfalfa, marine bacteria
PROC	Research Procedures	Activité de recherche, ou d'expérimentation	Real time PCR bibliometric analysis
	Medical Procedures	Activité diagnostique ou thérapeutique	appendectomy
PHEN ⋃ PHYS	**Phenomenon, Process or Function**	Fonction biologique, moléculaire, cellulaire, ou autre fonction de l'organisme	Mutation, protein interaction, apoptosis
	MEDLINE Title	Titre d'un article indexé dans MEDLINE	Author keywords in biomedical journal articles.
	Author name	Nom d'auteur	WJ Wilbur, Wendy Chapman
	Journal name	Nom d'une revue indexée dans MEDLINE	BMC Bioinformatics Science
	Citation information	Métadonnée liée aux articles : année de publication, pagination, ...	2008, 2009 Feb ;25(2)
	Abbreviations	Forme courte d'un un terme ou d'une entité	EEG, NEJM, UGT2B1

TABLE 2.2 – Schéma d'annotation PubMed. Les catégories correspondant à un groupe sémantique de l'UMLS sont en gras (en raison des contraintes de place, les codes désignant les groupes sémantiques introduits en figure 2.2 sont parfois utilisés). Les catégories correspondant à un type sémantique de l'UMLS sont soulignées.

Enfin, le schéma MERLoT comprend 4 catégories qui sont des groupes sémantiques, 8 catégories dont la définition s'appuie sur les types sémantiques et une catégorie qui ne peut

4. `https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt`

être directement définie grâce à l'UMLS (les hôpitaux). La table 2.3 présente ces catégories. En outre, le schéma MERLoT avait pour but de proposer une modélisation au delà des entités, en incluant le contexte de celles-ci ainsi que les relations entre les concepts ainsi dénotés. Pour cela, nous avons défini la notion d'*évènement* comme un regroupement de catégories susceptibles de dénoter un fait notable dans l'historique du patient. Ainsi, nous retenons comme événements : la catégorie *Chemicals and Drugs* dénotant la prise ou la prescription d'un médicament, les catégories *Disorder*, *Sign or Symptom* et *Biological Process or Function* dénotant l'expérience du patient ou l'établissement d'un diagnostic, la catégorie *Medical Procedure* dénotant la réalisation d'un soin et la catégorie *Concept or Idea*. Comme pour le travail de définition des concepts, le travail sur la définition des relations s'est appuyé sur le réseau sémantique de l'UMLS. Certaines relations ont été directement reprises du réseau sémantique (par exemple *Pharmacologic Substance/treats/Disease or Syndrome* s'est traduit par la relation *Chemical and Drugs TREATS Disorder*) et d'autres relations ont été définies sous la même forme, c'est à dire celle d'un triplet (concept1, relation, concept2). Les relations prenant en compte de multiples entités ont été ramenées à des relations sous forme de triplets, plutôt que d'être modélisées par des cadres sémantiques, ou des relations n-aires. Par exemple, les attributs d'une prescription médicamenteuse (dosage, forme, fréquence...) sont reliés individuellement au médicament concerné.

Ce schéma a été développé de manière itérative avec plusieurs phases de réflexions sur le schéma d'annotation, suivies de phases d'annotation concrète de textes cliniques variés, calcul d'accord inter-annotateur et révision du schéma et du guide d'annotation. Les principes et les premiers résultats de ce travail qui a duré environ deux ans sont décrits dans [Deléger et al., 2014]. L'application du schéma sur des textes issus d'une variété de spécialités médicales (fœtopathologie, hépato-gastro-entérologie, etc.) a permis de valider la pertinence générale du schéma et son applicabilité sur un large spectre de spécialités et de genres de textes cliniques (courriers, compte rendus de séjour, compte rendus d'actes...). Le schéma a également été présentés à des collègues médecins (Anita Burgun, Nicolas Griffon et Stéfan Darmoni) afin d'en valider la pertinence médicale.

Limites des modèles. Ces schémas comportent des limites, au-delà des caractéristiques dont l'absence est signalée dans La table 2.1.

Les limites principales de MERLoT sont au nombre de trois. Elles s'expliquent par un souci de simplification et de faisabilité des annotations à l'aide d'un schéma déjà relativement complexe. Il s'agit tout d'abord d'une légère divergence de l'annotation des expressions temporelles par rapport à la norme TimeML [Pustejovsky et al., 2003]. En effet, les expressions temporelles annotées dans MERLoT rassemblent les TIMEX et les "signaux", des déclencheurs qui signalent la présence d'expressions temporelles. La figure 2.3 illustre cette différence.

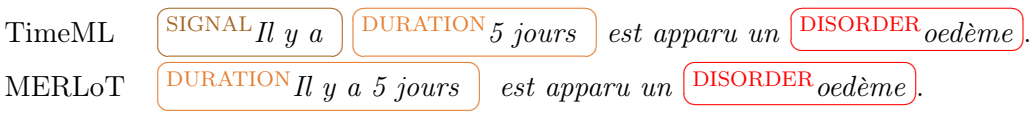


FIGURE 2.3 – Représentation des expressions temporelles

La suite de mon travail sur les expressions temporelles s'est appuyée sur la représentation timeML de manière plus fidèle [Tapi Nzali et al., 2015, Tourille et al., 2017a]. La deuxième limite concerne la représentation de l'anatomie et de la localisation dans l'espace. L'analyse spatiale est au moins aussi complexe que l'analyse temporelle, et mériterait une étude approfondie pour arriver à une représentation impliquant plusieurs entités

	Catégorie	Définition Courte	Exemples
	ANAT : Anatomy	Parties du corps	pied, artère fémorale droite
	CHEM : Chemicals and Drugs	Antibiotique, médicament ou autre substance pharmacologique	Questran, Insuline cortocoïdes
	DEVI : Devices	Objet utilisé à des fins de recherche, diagnostique ou thérapie	Pompe à insuline, sonde, tube
	GENE : Genes, Proteins and Molecular Sequences	Nom de toute séquence moléculaire	PTX1, POLYSERASE 3
LIVB	Living Beings (sauf humains)	Animal, bactérie	chien, e coli
	Personnes	Humain	patient, Dr. Durand
DISO	Disorders (sauf signes et symptômes)	Altération de l'état de santé	Diabète, MFIU, insuffisance mitrale
	Sign or Symptom	Signe ou symptôme	Fatigue, anneau fibreux
$\supset$ **CONC**	Concepts (sauf concepts temporels)	Terme dénotant une représentation mentale abstraite et générale, stable	poids, température longueur
	Expressions temporelles	Concepts temporels et leur instantiations	jeudi, 1987 trois semaines
$\subset$ **PROC**	Procédure Médicale	Activité relevant de la prise ouen charge des patients y compris l'ensemble des techniques diagnostiques et thérapeutiques	consultation, chimiothérapie, échographie
$\subset$ **(PHEN** $\cup$ **PHYS)**	Biological Process or Function	Processus ou état qui se produit naturellement ou résulte d'une activité physiologique.	transit
	Hôpital	Etablissement ou service de soins	CHU de Caen, Urgences Unité Henri Mondor

TABLE 2.3 – Entités comprises dans le schéma d'annotation MERLoT. Les catégories correspondant à un groupe sémantique de l'UMLS sont en gras (en raison des contraintes de place, les codes désignant les groupes sémantiques introduits en figure 2.2 sont parfois utilisés). Les catégories correspondant à un type sémantique de l'UMLS sont soulignées.

et relations. C'est une perspective de recherche ouverte, dont le résultat serait tout à fait compatible avec la modélisation simple proposée dans MERLoT : il suffirait d'y substituer une représentation plus complexe. Ce travail pourrait typiquement être amorcé par un travail de recherche de niveau M2 pour analyser la littérature sur l'espace en linguistique ([Aurnague et al., 2000] ou [Ligozat, 2010] inter alia) et en anatomie dans les ontologies médicales (par exemple [Schulz and Hahn, 2005]). La troisième limite concerne la représentation des relations complexes. Un premier aspect concerne la représentation des modalités dans les relations. En effet, seules des relations positives sont inclues dans le schéma. La représentation des relations niées ou hypothétiques se fait donc par le bias d'une modalité

(négation, hypothèse) appliquée sur l'une des entités impliquées dans la relation. Un autre aspect concerne les relations n-aires. En effet, la modélisation des relations par des triplets se traduit par la représentation de relations exclusivement binaire, entre deux concepts, alors qu'un plus grand nombre de concepts peuvent parfois intervenir. Par exemple, pour les prescriptions médicamenteuses, il peut arriver qu'un traitement soit prescrit à une première dose le matin, et une autre dose le soir. Dans ce cas, trois concepts entre en jeu dans l'évènement de prescription : le médicament, la dose, et la temporalité de la prise. Dans MERLoT, chaque aspect de la prescription est représenté par une relation binaire : *(Chemical_or_Drug, has_dosage, dosage1)*, *(Chemical_or_Drug, has_dosage, dosage2)*, *(Chemical_or_Drug, during, date1)*, *(Chemical_or_Drug, during, date2)*, ce qui fait que le lien entre les informations de temporalité et de dose n'est pas modélisé. Certains schémas, comme SHARP [Savova et al., 2012], pallient cette limite avec l'utilisation de représentation en cadres sémantiques (*semantic frames*).

Les limites du schéma QUAERO tiennent à sa proximité avec l'UMLS, et reflètent les limites de la représentation des connaissances mise en œuvre dans le métathésaurus. Bien que très complet, il n'est pas exhaustif. La mise en oeuvre du schéma dans une campagne d'annotation à un impact important sur la portée de ces limites, comme évoqué en section 2.2.3.

2.2 Création de corpus annotés

La création de corpus annotés peut être vue comme une instance de modélisation des connaissances par l'exemple, en vue d'apprentissage. Il existe un grand nombre de ressources disponibles pour l'anglais dans le domaine biomédical, mais moins pour les autres langues, en particulier pour le français. Ainsi, une partie de mon travail a été consacrée au développement et à la mise à disposition de telles ressources avec notament le corpus libre QUAERO français médical[Névéol et al., 2014c] et le corpus clinique MERLoT[Campillos et al., 2018]. Ces travaux se sont appuyés sur une première expérience d'annotation de de requêtes PubMed [Névéol et al., 2011], ainsi que sur ma participation à la campagne d'annotation pour le NCBI disease corpus[Doğan et al., 2014].

Les corpus annotés permettent également de caractériser le contenu d'un échantillon de textes. Par exemple, La table 2.4 met ainsi en évidence la prévalence des groupes sémantiques *GENE* et *DISO* dans les requêtes PubMed, la prévalence (attendue) du groupe *CHEM* dans les notices sur le médicament EMEA, la prévalence des groupes *DISO* et *PROC* dans les titres d'articles MEDLINE, et l'importance de la temporalité (groupe *CONC*) et des procédures (groupe *PROC*) dans les dossiers patients, des textes centrés autour de la personne du patient (groupe *LIVB*).

2.2.1 Corpus de requêtes PubMed

L'une des missions de la National Library of Medicine est d'accélérer les découvertes dans le domaine biomédical et de soutenir les avancées dans le domaine de la santé fondée sur les données. Pour ce faire, la bibliothèque s'appuie notament sur la base MEDLINE, qui indexe une grande partie de la littérature biomédicale (plus de 26 millions de documents en 2018). Le corpus de requêtes PubMed est une ressource développée dans le but de caractériser les besoins d'information des utilisateurs du moteur de recherche Pubmed, afin d'identifier des pistes de recherche pour améliorer l'accès à l'information de santé [Islamaj Dogan et al., 2009]. La figure 2.5 présente des extraits annotés du corpus. Le travail d'annotation réalisé dans ce cadre m'a permis d'avoir une réflexion sur la méthodologie de

Groupe Sémantique	**Requêtes PubMed**	**EMEA**	**Titres MEDLINE**	**MERLoT**
ANAT	877	583	1 499	4 446
CHEM	1 186	2 482	1 055	1 374
CONC	-	-	-	2 965 (+ 3 940 *Temporal*)
DEVI	104	170	128	1 068
DISO	2 247	1 510	2 843	3 329
GENE	2 263	-	-	5
GEOG		65	131	-
LIVB	767	817	941	3 921 (dont 3 862 *Personne*)
OBJC	-	174	100	-
PHYS∪PHEN	1 194	406	627	904
PROC	832	952	1 750	8 291

FIGURE 2.4 – Distribution des groupes semantiques dans les corpus biomédicaux

création d'annotations en corpus, et d'étudier en particulier l'apport de pré-annotations automatiques sur la qualité et l'efficacité du travail d'annotation ainsi que sur le confort de travail pour les annotateurs [Névéol et al., 2011].

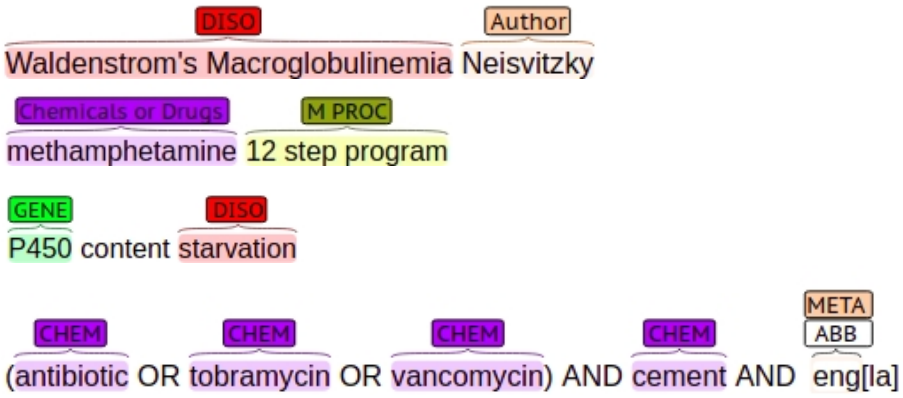


FIGURE 2.5 – Extraits du corpus de requêtes PubMed

2.2.2 Corpus clinique du français MERLoT

Le corpus MERLoT est une ressource développée dans le cadre du projet ANR CABeRneT comme support à la recherche en traitement automatique de la langue clinique en français. Ce travail a été réalisé dans le cadre des post-docs de Louise Deléger et Léonardo Campillos, avec la collaboration de Cyril Grouin, Anne-Laure Ligozat et Thierry Hamon.

Dans ce travail d'annotation, nous nous sommes attachés à marquer la complexité avec laquelle les entités médicales peuvent être dénotées en langue naturelle. Ainsi, nous avons annoté les entités discontinues, c'est à dire les cas où deux empans de textes séparés renvoient à un même concept, par exemple dans le cas de coordination ou d'élision : la mention *hépatite A et B* renvoie à deux entités distinctes, *hépatite A* et l'entité discontinue *hépatite B*. Nous avons également annoté les entités imbriquées afin de pouvoir étudier la compositionnalité et les principes de construction des termes médicaux du point de vue catégoriel. Par exemple, la mention *prévention du cancer du sein* renvoie à trois entités imbriquées de catégories différentes : l'anatomie (*sein*), les problèmes médicaux (*cancer du sein*) et les procédures (*prévention du cancer du sein*). La durée et la complexité de la campagne d'annotation a également soulevé la problématique de la consistance des annotations sur la durée et des moyens à mettre en œuvre pour la contrôler.

Ce corpus comprend une sélection de textes cliniques issus d'un groupe d'institutions hospitalières françaises [5]. Après un travail préliminaire appliquant le schéma d'annotation MERLoT à un large spectre de spécialités [Deléger et al., 2014], nous avons sélectionné pour un travail à plus grande échelle 500 documents issus du service d'hépato-gastro-nutrition que nous avons désidentifiés[Grouin and Névéol, 2014], et annotés en sections [Deléger and Névéol, 2014] avant de procéder à une annotation en entités, relations et attributs. La figure 2.6 présente des extraits annotés du corpus clinique du français MERLoT, illustrant la représentation des entités, de leurs attributs, et des relations.

La campagne d'annotation pour le corpus MERLoT s'est étendue sur deux années, et a impliqué la participation de six annotateurs. Ainsi, la campagne d'annotation a été suivie par un travail d'harmonisation des annotations afin d'améliorer la consistance globale sur l'ensemble du corpus. Des outils ont été développés afin de repérer systématiquement des éventuelles inconsistances afin de guider le travail d'harmonisation. Ce travail est décrit en détail dans un article de revue co-écrit par l'ensemble des contributeurs [Campillos et al., 2018].

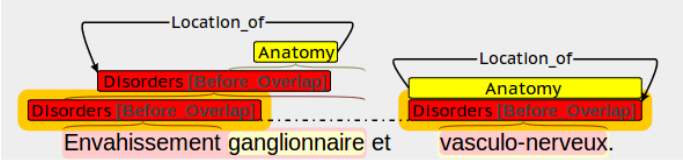
En marge de ce travail, une étude plus spécifique sur les aspects temporels a donné lieu à l'annotation de 360 documents supplémentaires en se focalisant sur les seules expressions temporelles, annotées selon le standard timeML et normalisées [Nzali et al., 2015]. Ce travail est illustré par la figure 2.7.

2.2.3 Corpus biomédical du français QUAERO

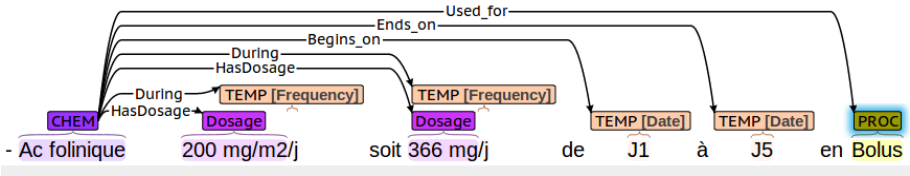
Le corpus QUAERO Médical du français est une ressource développée dans le cadre de la recherche en reconnaissance d'entité et normalisation dans un contexte multilingue [Névéol et al., 2014c]. En effet, les textes utilisés dans ce corpus ont été choisis pour disponibilité dans plusieurs langues : le français et l'anglais pour les titres d'articles indexés dans MEDLINE, 25 langues européennes pour les notices EMEA (European Medication Agency) [Tiedemann, 2009], le français, l'anglais et l'allemand pour les dépôts de brevets EPO (European Patent Office) [6]. La possibilité de transfert des annotations vers d'autres

5. Nous remercions le Service d'Informatique Biomédicale (SIBM) ainsi que l'équipe CISMeF du CHU de Rouen qui nous ont permis d'utiliser le corpus LERUDI pour cette étude.

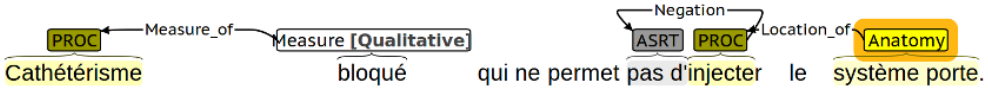
6. `https://www.epo.org/`



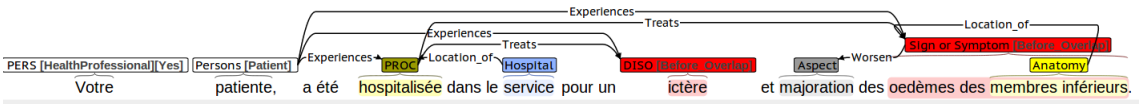
(a) Entités inbriquées et discontinues.



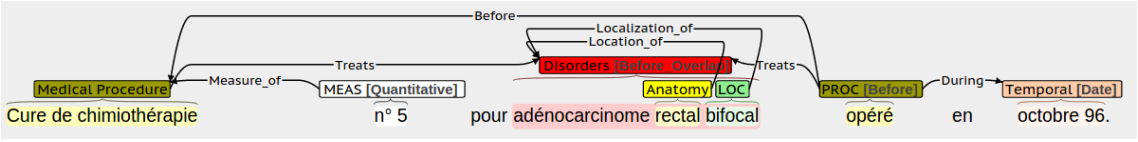
(b) Annotations autour du médicament



(c) Exemple de négation



(d) Exemple d'aspect



(e) Exemple de relations

FIGURE 2.6 – Extraits du corpus clinique du français MERLoT

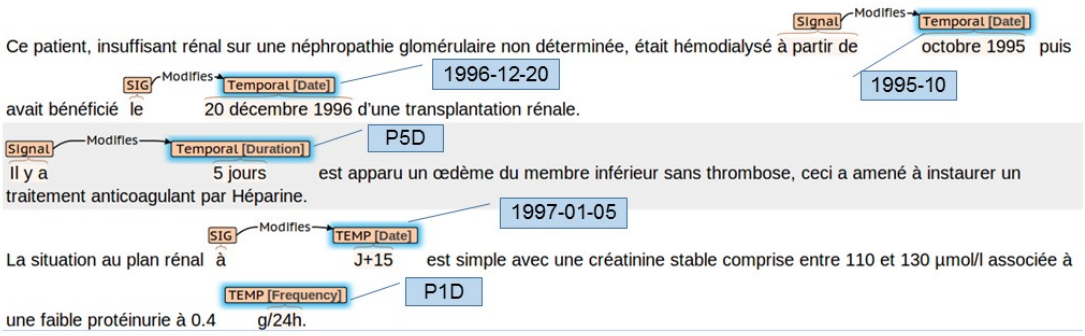


FIGURE 2.7 – Extraits du corpus clinique annoté en signaux et expressions temporelles normalisées.

langues reste ouverte, ainsi que la perspective d'une étude multilingue de la reconnaissance d'entités et de la normalisation.

Suite à la campagne d'annotation initiale, le corpus a ensuite été repris dans le but de

créer un jeu de référence normalisé pour les entités nommées dans domaine biomédical. Dans ce deuxième temps, la partie du corpus issue des brevets EPO a été mise de côté car elle représente un genre de texte tout à fait particulier qui semble moins généralisable au domaine biomédical que MEDLINE ou EMEA. Le travail sur le corpus a consisté d'une part en une conversion des annotations d'un format balisé vers un format déporté. Le nouveau format adopté améliorait la lisibilité des annotations pour une consultation manuelle, et présentait également l'avantage de pouvoir réaliser des annotations discontinues. Ce travail technique a néanmoins eu des conséquences sur le format du texte final, dont la segmentation n'est pas standard. Bien que l'algorithme de conversion se soit attaché à restituer le texte original, l'insertion d'espaces lors de l'annotation au format balisé n'a pas pu être complètement gommée. Une autre partie du travail a porté sur l'harmonisation des annotations afin d'améliorer la consistance globale sur l'ensemble du corpus.

MEDLINE

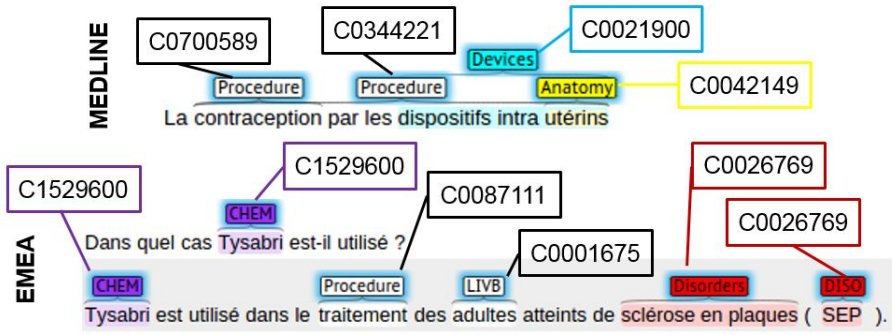
La<PROC CUI="C0344221"> <PROC CUI="C0700589"> contraception </PROC> par les <DEVI CUI="C0021900">dispositifs intra <ANAT CUI="C0042149">utérins </ANAT> </DEVI> </PROC>

EMEA

Dans quel cas <CHEM CUI="C1529600"> Tysabri </CHEM> est-il utilisé ?

<CHEM CUI="C1529600"> Tysabri </CHEM> est utilisé dans le <PROC CUI="C0087111"> traitement </PROC> des <LIVB CUI="C0001675"> adultes </LIVB> atteints de <DISO CUI="C0026769"> sclérose en plaques </DISO> (<DISO CUI="C0026769"> SEP </DISO>).

(a) Annotations balisées



(b) Annotations déportées visualisées à l'aide de BRAT

FIGURE 2.8 – Exemples d'annotations extraites du corpus QUAERO Médical du français

Ce corpus, librement disponible [7], a été utilisé dans le cadre des campagnes d'évaluation CLEF eHealth [Névéol et al., 2015, Névéol et al., 2016a] décrites dans la section 2.3.

Une sélection de titres MEDLINE et de documents EMEA ont été annotés manuellement. Les annotations s'appuient sur les concepts de l'UMLS comme illustrés en figure 2.2.

Les annotations ont été faites avec une couverture aussi large que possible, de sorte que des entités imbriquées sont marquées et une même entité peut renvoyer à plus d'un concept UMLS. Spécifiquement, de multiples annotations sont faites dans les cas suivants :

7. `https://quaerofrenchmed.limsi.fr/`

— Si une mention renvoie à plus d'un Groupe Sémantique, elle doit être annotée pour tous les Groupes Sémantiques pertinents. Par exemple, la mention "récidive" dans l'expression "prévention des récidives" doit être annotée avec les catégories "DISORDER" (CUI C2825055) et "PHENOMENON" (CUI C0034897) ;

— Si une mention renvoie à plus d'un concept à l'intérieur d'un même Groupe Sémantique, tous les concepts pertinents doivent être marqués. Par exemple, la mention "maniaques" dans l'expression "patients maniaques" doit être annotée avec les CUIs C0564408 et C0338831 (categorie "DISORDER") ;

— Les entités qui présentent un recouvrement avec d'autres entités doivent être marquées indépendamment. Par exemple, dans l'expression "infarctus du myocarde", la mention "myocarde" doit être annotée avec la catégorie "ANATOMY" (CUI C0027061) et la mention "infarctus du myocarde" doit être annotée avec la catégorie "DISORDER" (CUI C0027051)

A la différence de la méthodologie appliquée par [Kors et al., 2015], le mentions renvoyant à des concepts de l'UMLS pour lesquels aucun terme de l'UMLS n'est identique à l'expression rencontrée que ce soit en français ou en anglais sont tout de même annotés. Ainsi, la mention *français* pourra être annotée avec la catégorie "GEOG" (CUI C0016674) bien que ni les termes "français" ni French" ne soient dans l'UMLS.

Par contre, les entités auxquelles aucun concept ne correspond dans l'UMLS ne sont pas annotées. Cette limite pourrait être palliée par l'annotation d'une mention avec la normalisation "CUI-less" (c'est-à-dire, absence de concept correspondant dans la ressource termino-ontologique), comme cela a pu être fait pour la campagne SEMEVAL 2014 [Elhadad et al., 2015]. Des travaux plus récents ont également repris ce corpus pour proposer une normalisation plus fine des concepts annotés [Osborne et al., 2018]. Les auteurs ont proposé d'une part d'étendre la normalisation à l'ensemble de l'UMLS (comme c'est le cas pour QUAERO) plutôt que de se limiter à une seule ressource (SNOMED-CT pour le corpus SEMEVAL) et d'autre part d'autoriser également la normalisation à s'appuyer sur plus d'un seul concept à la fois, autorisant ainsi la *post-coordination* de concepts qui permet de pallier l'absence de concepts dont le sens serait défini en associant des concepts existants.

La figure 2.8 présente des exemples d'annotations extraites du corpus QUAERO Médical du français. La partie supérieure de la figure présente les annotations originales, réalisées avec un balisage inséré dans le texte. La partie inférieure présente les mêmes annotations, converties vers un format déporté et visualisées à l'aide de l'outil BRAT [8] [Stenetorp et al., 2012].

2.3 Organisation de campagnes d'évaluation

> « *J'entends et j'oublie. Je vois et je me souviens. Je fais et je comprends.* »
> Confucius (551 à 479 av. J.-C.)

L'une des applications directes des corpus annotés est leur utilisation dans le cadre de campagnes d'évaluation permettant de motiver la réflexion de la communauté sur des tâches précises définies par modélisation des informations instanciée dans les corpus. Ces campagnes proposent des conditions de travail contrôlées, partagées par la communauté, ce qui favorise la reproductibilité et la comparaison directe des méthodes et systèmes de traitement automatique du langage. Elles ont également un aspect pédagogique important car

8. `http://brat.nlplab.org/`

elles permettent à des étudiants de se confronter à des données et des tâches "réelles" dans un calendrier contraint (on pourra noter la participation de groupes d'étudiants à partir du niveau M2 à quasiment toutes les campagnes CLEF eHealth et WMT). Globalement, les campagnes fournissent des données précieuses sur l'état de l'art et les perspectives offertes par de nouvelles approches.

J'ai porté pendant quatre ans l'organisation d'une tâche d'extraction d'information multilingue à CLEF eHealth (2015-2018) et apporté mon soutien pendant trois ans à l'organisation d'une tâche sur la traduction automatique de textes biomédicaux dans le cadre du Workshop on Machine Translation (WMT, 2016-2018). Ma contribution à ces campagnes a permis d'animer la communauté autour d'une réflexion sur le traitement automatique des langues autres que l'anglais dans le domaine biomédical - en particulier le français, mais aussi l'espagnol, le hongrois, l'italien et le portugais.

Ce contexte évaluatif présente un intérêt fondamental sur le plan de la réflexion méthodologique conduite par l'ensemble des participants, mais aussi par les organisateurs sur le formalisme même de la campagne sur le plan de la modélisation des problèmes proposés. En collaboration avec Kevin Cohen et dans le contexte global du projet Européen MIROR, l'organisation des campagnes CLEF eHealth a ainsi permis d'aborder la question de la reproductibilité des travaux scientifiques en traitement automatique des langues. Après quelques études de cas[Névéol et al., 2016b, Cohen et al., 2017] nous proposons de distinguer trois dimensions : la reproductibilité d'une conclusion, d'une observation, et d'une valeur [Cohen et al., 2018].

Les campagnes présentent également un intérêt applicatif permettant aux acteurs du domaine d'apprécier quantitativement la performance effective des solutions proposées dans des conditions réelles. Les campagnes CLEF eHealth que nous avons organisées ont été conçues en ce sens. Ainsi, suite au travaux réalisés en collaboration avec Cyril Grouin (décrits en section 4) nous avons souhaité proposer une campagne d'évaluation des méthodes de désidentification des textes cliniques dans les langues autres que l'anglais. Malheureusement, nous n'avons pas réussi à obtenir un corpus clinique distribuable dans ce cadre. La disponibilité du corpus QUAERO médical du français nous a donc incités à proposer une campagne sur l'extraction et la normalisation d'entités cliniques en 2015 et 2016. Dans le cadre d'une collaboration entre le groupe ILES du LIMSI et l'équipe de l'Inserm-CépiDC, nous leur avons proposé de profiter de ce dispositif pour explorer la problématique du codage automatique des certificats de décès. En effet, le CépiDC est chargé de la production des statistiques relatives aux décés au niveau national et souhaitait évaluer l'apport potentiel de l'intégration d'outils de traitement automatique de la langue dans leur processus de codage. Le parallèle entre cet objectif, le dispositif d'indexation en place à la National Library of Medicine avec l'utilisation de MTI pour l'indexation MEDLINE et les avancées méthodologiques offertes dans ce domaine par les campagnes BioASQ [9] indiquaient un contexte tout à fait prometteur pour le codage des causes de décès. Ainsi, en collaboration avec le CépiDC et le réseau européen IRIS, nous avons également pu proposer un tâche de codage des certificats de décès dans plusieurs langues européennes à partir de 2016. Le travail effectué dans le cadre de l'ensemble de ces campagnes est résumé en section 2.3.2.

Les difficultés rencontrées pour la diffusion de corpus cliniques dans des langues européennes pour des campagnes de traitement automatique de la langue a par la suite motivé ma participation à l'organisation des campagnes WMT de 2016 à 2018 sur la traduction automatique de textes du domaine biomédical. En effet, les travaux actuels en TAL cliniques dans les langues autres que l'anglais se tournent vers des méthodes de production de corpus synthétiques afin de disposer d'un objet d'étude partagé par la communauté. La

9. `http://bioasq.org/`

traduction automatique m'a paru être une voie possible pour cette direction de recherche. Le travail effectué dans le cadre de ces campagnes est résumé en section 2.3.5.

La table 2.9 présente le nombre de participants pour chaque campagne au fil du temps. On observe un intérêt modeste mais constant voire en augmentation pour CLEF eHealth. Ce résultat est encourageant, car les tâches proposées représentent un défi et demandent de traiter des langues pour lesquelles peu de ressources sont disponibles. On peu noter que si certaines équipes montrent une participation récurrente au fil du temps, on observe toutefois un renouvellement des participants ; si plusieurs équipes ont participé à deux éditions sur trois (ou quatre), aucune équipe n'a participé à l'ensemble des éditions pour une même campagne. Cela s'explique par la diversité des langues proposées et par l'investissement en temps important pour mettre en œuvre une participation, même sur une campagne récurrente.

	Campagne				
	2015	**2016**	**2017**	**2018**	**total 2015-2018**
CLEF eHealth (entités)	7	5	-	-	9
CLEF eHealth (codage)	-	5	11	14	22
WMT biomédical	-	5	7	6	13

FIGURE 2.9 – Participation aux campagnes d'évaluation CLEF eHealth (tâche d'extraction d'information) et WMT biomédical.

2.3.1 Mesures d'évaluation

Pour CLEF eHealth , les performances des systèmes ont été évaluées avec les mesures habituelles dans le domaine de l'extraction d'information : précision, (Formule 2.1), rappel (Formule 2.2) et F-mesure (Formule 2.3 ; avec la valeur β=1.).

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{2.1}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2.2}$$

$$\text{F-measure} = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{2.3}$$

Les métriques ont été calculées au niveau du document puis micro-moyennées sur l'ensemble du corpus. Les performances des systèmes sont évaluées en comparant les prédictions soumises par les participants aux annotations manuelles de référence (gold standard). En pratique, pour le corpus QUAERO, l'outil `brateval` initialement développé par by Verspoor et al. [Verspoor et al., 2013], que nous avons étendu pour prendre en compte l'évaluation de la normalisation. Les mesures définies ci-dessous rendent compte de mesures *strictes*. Pour la reconnaissance d'entités, nous avons également utilisé une mesure secondaire qui évaluait la reconnaissance *inexacte*, c'est à dire autorisant une variation soit de frontière soit de catégorie pour les entités reconnues. En pratique, la mesure inexacte est généralement un supérieure à la mesure stricte, mais les performances respectives des systèmes

se positionnaient de manière similaire avec les deux mesures. Pour les causes de décès, les métriques sont obtenues à l'aide d'un programme Perl distribué aux participants.

Pour **la reconnaissance d'entités**, une entité proposée par le système était comptée comme un vrai positif quand le type et les frontières de l'entité correspondaient à la référence. En revanche, une entité extraite était considérée comme un faux négatif si le type ou les frontières de l'entité ne correspondaient pas exactement à la référence.

Pour **la reconnaissance d'entités normalisées**, une entité proposée par le système était comptée comme un vrai positif quand le type, les frontières, et le(s) CUI(s) associés à l'entité correspondaient à la référence. Un crédit partiel était attribué lorsque le et les frontières de l'entité proposée étaient identiques à la référence et que la liste de CUIs proposés correspondait partiellement à la référence.

Pour la **normalisation d'entités**, un vrai positif était compté lorsque le(s) CUI(s) associés à l'entité correspondaient à la référence. Si la référence ou le système proposait une liste de CUIs et non un CUI unique pour une entité donnée, un crédit partiel était attribué en cas de recouvrement et non d'adéquation complète. Cependant, les CUIs proposés par le système mais absent de la liste de référence étaient comptés comme des faux positifs.

Pour le **codage**, un code CIM10 proposé le système était compté comme un vrai positif s'il correspondait à la référence pour cette ligne, ou ce document.

Pour WMT , les performances ont été évaluées avec les mesures habituelle en traduction automatique. Le score BLEU est une mesure automatique qui cherche à évaluer la similarité entre une traduction produite par un système automatique et une traduction humaine de référence (il est admis que plusieurs traductions correctes d'un même texte sont possibles), en considérant que plus la traduction évaluée est proche de la référence, meilleure elle est. Le score BLEU repose sur le recouvrement entre les n-grammes tokens issus du texte évalué et de la référence. L'autre mesure utilisée est une évaluation qualitative issue d'une comparaison manuelle entre deux traductions candidates : soit une traduction automatique et une traduction humaine de référence, soit des traductions automatiques issues de différents systèmes. Dans cette évaluation, l'évaluateur n'a pas connaissance des types de traductions en présence et porte un jugement de préférence pour estimer si les deux traductions candidates sont équivalentes, ou si l'une des traductions est meilleure que l'autre. Pour mettre en œuvre cette évaluation qualitative, nous avons utilisé le système libre Appraise [10].

2.3.2 Extraction d'information multilingue à CLEF eHealth 2015-2018

Notre contribution à la plateforme CLEF eHealth a porté sur l'organisation de tâches d'extraction d'information biomédicale dans des langues autres que l'anglais et dans un contexte multilingue. Le français a été particulièrement représenté, mais nous avons également pu proposer des corpus en hongrois et italien en 2018. L'aspect multilingue a d'abord été abordé grâce à l'anglais en 2017, puis avec le hongrois et l'italien en 2018. Globalement, la tâche a reçu une participation soutenue aussi bien de la part d'équipes basées dans des pays francophones que non-francophones. Les résultats obtenus par les différentes équipes montrent que des méthodes variées sont mises en œuvre. Des performances très hétérogènes sont obtenues et montrent l'intérêt d'aborder une tâche dans la durée : d'une part pour surmonter les difficultés techniques liées au format des données et d'autre part pour donner plus de temps à la réflexion qui permet une maturation des approches proposées. Nous détaillons ci-dessous les modalités de déroulement des campagnes ainsi qu'un résumé des

10. `https://github.com/cfedermann/Appraise`

résultats. Une synthèse plus détaillée des performances obtenues par les systèmes soumis est présentée en Annexe.

Extraction et normalisation d'entités cliniques en français. En 2015 et 2016, nous avons proposé une tâche d'extraction d'information fondée sur le corpus QUAERO médical du français présenté ci-dessus. L'un des buts de ces campagne était de fournir un soutien au développement et l'évaluation d'outil d'analyse des textes biomédicaux en français, en dépit du manque de ressources terminologiques connu pour cette langue, par rapport à l'anglais. La table 2.4 présente des statistiques descriptives globales du corpus, découpé en trois parties : entraînement, développement (utilisé comme corpus de test en 2015) et test.

	EMEA			MEDLINE		
	Training	Development	Test	Training	Development	Test
Documents	3	3	4	833	832	833
Tokens	14,944	13,271	12,042	10,552	10,503	10,871
Entities	2,695	2,260	2,204	2,994	2,977	3,103
Unique Entities	923	756	658	2,296	2,288	2,390
Unique CUIs	648	523	474	1,860	1,848	1,909

TABLE 2.4 – Descriptive statistics of the QUAERO French Medical Corpus

La tâche de **reconnaissance d'entités nommées** consistait à analyser le texte brut des documents afin de marquer les dix types d'entités clinique d'intérêt définies dans le schéma d'annotation QUAERO. Les participants avaient la possibilité d'extraire des *entités simples* (c'est à dire de marquer les mentions renvoyant aux entités du textes) ou d'extraire des *entités normalisées* (c'est à dire de fournir pour chaque mention extraite le ou les identifiants de concept UMLS lié).

Les méthodes proposées par les participants pour la reconnaissances d'entités reposaient soit sur l'appariement entre le texte et une base de connaissance (lorsqu'il s'agissait de l'UMLS, les équipes proposaient alors des entités normalisées) soit sur des méthodes d'apprentissage statistiques telles que des Champs Aléatoires Conditionnels (CRF).

La tâche de **normalisation d'entités** consistait à effectuer la liaison référentielle entre les entités marquées dans le texte et les concepts UMLS pertinents. Cette tâche a finalement attiré peu de participations par rapport à la tâche d'extraction d'entités, de la part d'équipes différentes sur les deux éditions. Les comparaisons entre années sont indicatives du fait de la différence entre les jeux d'entraînement et de test utilisés. On peut observer que les systèmes offrant les meilleurs résultats pour les deux années reposaient sur une approche à base de connaissance. Au delà de la campagne CLEF eHealth, le corpus QUAERO médical continue d'être utilisé pour évaluer des méthodes de normalisation d'entités[Roland Roller and Leser, 2018].

2.3.3 Codage des causes de décès en anglais, français, hongrois, italien.

A partir de 2016, nous avons proposé une tâche de codage automatique des causes de décès fondée sur des corpus issus des organismes chargés de la collecte des statistiques

nationales de décès à l'aide de l'équipe de Grégoire Rey à l'Inserm-CépiDC. L'un des buts de ces campagnes était de stimuler le développement et l'évaluation d'outil de codage fondés sur la classification internationale des maladies (CIM10), afin de faciliter l'intégration de tels outils dans le processus de traitement des certificats de décès pour l'obtention des statistiques.

Pour l'ensemble des langues, les données reçues comprenaient le texte original des certificats de décès à l'issue d'une phase de désidentification (pour l'anglais, le français, et le hongrois) qui pouvait être suivie d'une ré-organisation du contenu (pour l'italien). Nous disposions également des codes CIM10 associés aux textes par les codeurs professionnels. Pour le français en particulier, nous avons proposé une méthode d'alignement des lignes de texte avec les codes CIM10 correspondants ; l'acteur principal de ce travail était Thomas Lavergne[Lavergne et al., 2016]. Ainsi, lors de la campagne, nous avons diffusé un jeu de données *aligné* pour le français, et un jeu de données dit *raw* pour les quatre langues.

L'utilisation de ces deux formats de données pour le français nous a permis d'une part d'évaluer l'impact du pré-traitement d'alignement sur les performances des systèmes. D'autre part, le format *raw* nous a permis de proposer une évaluation relativement comparable entre les différentes langues.

La Figure 2.10 présente l'évolution des performances sur le français aligné lors des campagnes 2016 à 2018. Les comparaisons d'une année à l'autre ne sont qu'indicatives en raison des différences entre les jeux d'entraînement et de test. Pour les trois équipes ayant participé à plus d'une campagne, on note une nette amélioration des performances d'une participation à l'autre.

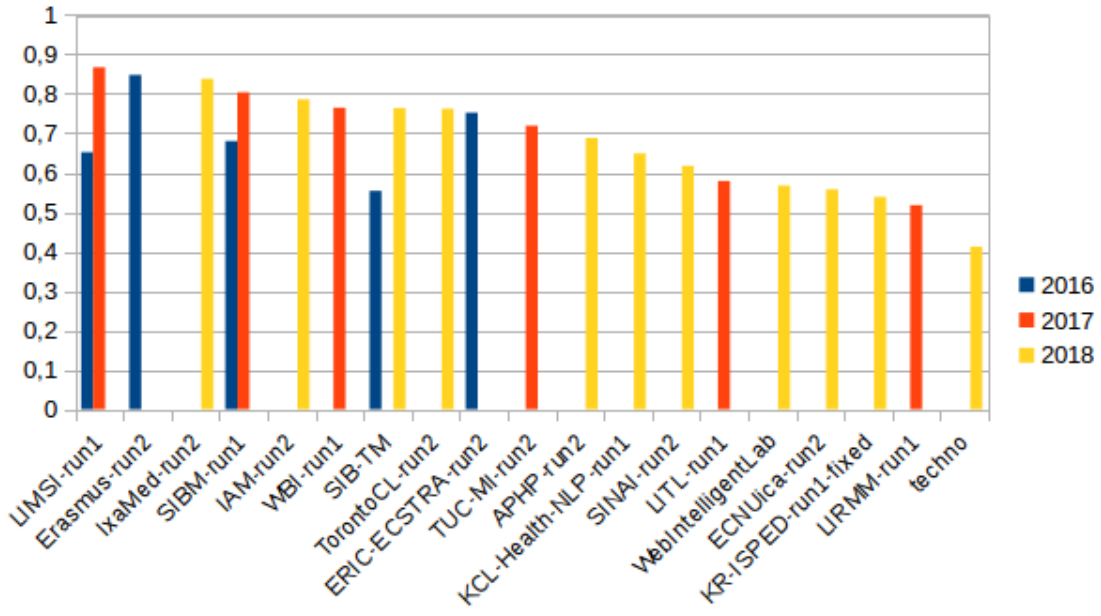


FIGURE 2.10 – Résultats obtenus lors des campagnes CLEF eHealth 2016-2018 sur le codage des causes de décès à l'aide la CIM10, pour les textes en français alignés.

Au cours des différentes campagnes, les méthodes proposées par les participants ont principalement reposé (1) sur des bases de connaissance telles que l'UMLS ou les dictionnaires CIM10 fournis par le CépiDC et exploré des méthodes d'appariement à divers niveau de variation lexicale, y compris la correction orthographique, (2) des algorithmes de recherche d'information et (3) des méthodes de classification mono ou multi-classes, en

particulier neuronales. Certaines équipes ont proposé des systèmes combinant plusieurs de ces méthodes. Lors de la campagne de 2018, la prise en compte des différentes langues proposées a été abordée soit par le biais de systèmes paramétrables pour chaque langue (par exemple, en spécifiant la langue des ressources à utiliser) soit grâce à un système unique conçu pour traiter l'ensemble des langues.

Les travaux entamés lors d'une campagne CLEF eHealth ont parfois été poursuivis afin d'être élargis et adaptés à d'autres cas d'usage comme l'analyse des réseaux sociaux [Tutubalina et al., 2018].

2.3.4 Reproductibilité.

En 2016 et 2017, nous avons proposé une tâche de reproductibilité à CLEF eHealth, afin d'encourager le partage des outils développé dans le cadre du challenge et d'expliciter les obstacles potentiels à la reproductibilité dans le contexte expérimental très contrôlé d'une campagne d'évaluation.

Lors de la première tâche de réplication[Névéol et al., 2016b], nous avons établi une grille d'évaluation afin de documenter les différentes étapes de l'expérience de réplication : configuration, installation, utilisation, résultats. Lors de la deuxième tâche de réplication, nous avons enrichi la grille d'évaluation pour recueillir plus d'informations sur le profil des analystes et les expérimentations et nous avons également inclus un script baseline dans la liste des systèmes étudiés.

La réplication des résultats des participants ayant soumis leur système pour les éditions de CLEF eHealth 2016 et 2017 s'est montrée faisable collectivement par une équipe de plusieurs analystes. On remarque cependant qu'aucun des analystes n'a été en mesure de reproduire l'ensemble des résultats seul. Nous avons observé en 2017 que la reproduction des résultats de la baseline, un système pourtant anticipé comme "simple" (un script Perl), ne présentait pas forcément moins de difficultés que les systèmes plus complexes.

Pour chacun des systèmes étudiés, nous avons été en mesure d'obtenir des résultats identiques ou très proches de ceux soumis par les participants eux-mêmes. La facilité de reproduction des résultats était néanmoins variable, en fonction de l'environnement de travail des analystes. Les principales difficultés rencontrées auraient pu être résolue grâce à une documentation de la part des auteurs des systèmes, qui ont parfois été contactés directement pour résoudre des problèmes à diverses étapes du processus. Le fait d'anticiper de telles difficultés et de les documenter en amont de l'utilisation des systèmes par des tiers demande un effort supplémentaire lors de la conception et du développement des systèmes, en ayant la reproductibilité comme objectif. Dans le cadre d'une campagne d'évaluation, l'utilisation d'un dispositif comme codalab [11] peut permettre de faciliter et de fluidifier le processus. Cette plateforme est par exemple utilisée dans les campagnes SemEval. Cependant, elle demande un effort de préparation supplémentaire de la part des organisateurs de la campagne, y compris la prise en main de l'outil.

2.3.5 Traduction automatique de textes biomédicaux à WMT 2016-2018

Les campagnes WMT sur la traduction automatique de textes biomédicaux ont été menées à l'initiative de Mariana Neves et Antonio Jimeno Yepes qui portent cette activité, et en collaboration avec Karin Verspoor et d'autres collègues sollicités pour leur expertise dans les langues de la campagne et/ou leur effort dans le développement de corpus utilisés dans les campagnes.

11. `http://codalab.org/`

Ma contribution principale à l'organisation de la tâche de traduction biomédicale à WMT à porté sur le développement et l'évaluation de corpus parallèles anglais/français utilisés dans la campagne : il s'agit notamment des corpus Scielo et EDP. Les discussions avec les collègues organisateurs de ces campagnes ont également débouché sur une revue des corpus parallèles disponibles pour le domaine biomédical [Névéol et al., 2018], et sur des discussions sur les mesures d'évaluation de la traduction biomédicale. Ce travail a fait suite à une collaboration avec Antonio Jimeno Yepes amorcée en 2012 à la National Library of Medicine où nous nous étions intéressés au développement de corpus parallèles pour les paires de langues anglais/français et anglais espagnol [Jimeno Yepes et al., 2013, Jimeno Yepes and Névéol, 2013]. La collaboration du LIMSI avec le centre Cochrane français sur la traduction de résumés de revues systématiques de la littérature [Max et al., 2013] a également motivé un investissement sur cette thématique.

Les résultats obtenus en termes de performance de traduction sont particulièrement intéressants pour l'espagnol et le portugais pour lesquelles un gain de plusieurs points BLEU a pu être obtenus (voir tableau 5.5, en annexe).

Cette amélioration est attribuée en partie au volume de données (corpus parallèles) mises à disposition. Des améliorations méthodologiques ont également été observées. Pour le français, le volume des corpus parallèles biomédicaux disponibles reste modeste et les performances offertes par la traduction biomédicale demande un soin particulier pour adapter les systèmes au domaine de spécialité.

Un certains nombre de problèmes particuliers comme la prise en compte des acronymes restent non résolus pour l'ensemble des langues.

Il est également dommage que le paradigme d'évaluation n'ait pas pu évoluer au fil des campagnes pour inclure des mesures au niveau du texte ou des mesures prenant en compte la correction médicale des traductions, en négligeant les aspects stylistiques ou grammaticaux - si tant est qu'il soit possible de les évaluer indépendamment. Cette lacune est due à la quantité de travail requise par l'organisation d'une campagne, en terme de préparation des corpus et de mise en place de l'évaluation.

2.4 Discussion

Les trois cas de modèles de représentation de l'information présentés à la section 2.1.1 montrent que les types et groupes sémantiques de l'UMLS sont globalement pertinents comme points de départ de la réflexion pour une diversité de textes biomédicaux. Dans la mesure du possible, il conviendra d'utiliser des schémas ou modèles existants, ou tout au moins des schémas qui permettent un positionnement ou la définition de correspondances avec l'existant.

Mes travaux ont contribué à une diversification de l'offre de corpus annotés du domaine biomédical, en particulier pour les corpus dans les langues autres que l'anglais. La question du partage libre de ces données reste problématique. Néanmoins, nos travaux ont également permis de dégager une méthodologie de développement de corpus assistée par le traitement automatique de la langue. Ce protocole a par exemple été formalisé pour une application à la désidentification de documents cliniques par repérage d'entités nommées identifiantes. Ce travail est décrit au chapitre 4.

Une question méthodologique importante soulevée par le développement de ressources termino-ontologiques et de corpus annotés est celle de la délimitation de la couverture de ces formes de représentation des connaissances. Il est bien sûr souhaitable de disposer de ressources avec une couverture très large en termes de concepts et de relations. Cependant, une ressource termino-ontologique a-t-elle vocation a être exhaustive sur le plan termino-

logique ? Est-il légitime d'y inclure l'ensemble des variations sémantiques et des variations de surface (par exemple, incluant l'ensemble des flexions voire même des fautes d'orthographe courantes...) que peuvent prendre les termes liés à un concept ? De même, lors du développement de corpus annotés, faut-il que les mentions annotées soient limitées par les connaissances strictement consignées dans les ressources termino-ontologiques ?

Pour ma part, je pense que ces deux types de représentations des connaissances sont complémentaires et qu'il faut exploiter cette complémentarité. La couverture des ressources termino-ontologique ne peut pas être exhaustive, et il est à mon sens plus efficace de relever le défi de la variation terminologique avec des outils méthodologiques de reconnaissance des variations. Par complémentarité, l'annotation de corpus doit être l'un des leviers permettant de recenser des réalisations contextuelles originales des concepts (par exemple, les entités discontinues), qui pourront ensuite être exploitées pour le développement de méthodes de reconnaissance des variations.

L'organisation de campagnes d'évaluation dans plusieurs langues sur des tâches d'analyse fine de texte est souhaitable ; en particulier, une telle campagne portant sur l'extraction d'information à partir de textes cliniques pourrait stimuler le développement de méthodes génériques permettant d'aborder plusieurs langues, ou de méthodes intégrant une adaptation à plusieurs langues.

Chapitre 3

Méthodes pour l'analyse de textes biomédicaux

L'objectif de mon travail est la compréhension automatique de textes du domaine biomédical, abordée sous l'angle de l'analyse de textes du domaine biomédical (et en particulier des textes cliniques) afin d'extraire les informations contenues dans ces textes sous forme de représentation structurée. Il convient donc tout d'abord de décomposer les différentes étapes de l'analyse d'un texte, depuis la sélection des documents à analyser en passant par l'extraction d'information et leur exploitation dans un cadre applicatif. La Figure 3.1 présente ces étapes avec un degré de complexité et de spécificité par rapport au domaine biomédical croissants.

FIGURE 3.1 – Étapes de l'analyse de textes cliniques

Les deux premières étapes sont certainement communes à tout travail s'appuyant sur des textes issus du monde réel. Cependant, il est important de souligner la difficulté et l'importance de ces étapes préliminaires d'acquisition des corpus. Il est d'autant plus important de ne pas négliger ces étapes qu'elles peuvent donner lieu à l'exploration de problèmes de recherche, comme par exemple le traitement des lignes coupées à l'aide de méthodes d'apprentissage sans données manuellement annotées [Zweigenbaum et al., 2016]. Dans le domaine biomédical, la confidentialité des données et des textes joue un rôle dès le début du processus, par exemple pour les textes issus des dossiers électroniques patient ou des réseaux sociaux. Ainsi, certaines étapes de l'analyse de contenu, placées en aval de la chaîne, peuvent contribuer à la mise à disposition des données pour l'ensemble des étapes d'analyse. La désidentification permet de contribuer à la préservation de la confidentialité. L'identification de la structuration des documents cliniques peut présenter un intérêt pour la confidentialité ou guider les analyses ultérieures, selon le type de structuration considéré. En effet, pour certains corpus il peut être utile de repérer les sections qui contiennent l'essentiel du contenu médical concernant un patient, à l'exclusion d'autres sections contenant des informations administratives et qui sont susceptibles de contenir également de nombreux éléments identifiants concernant le patient ou les professionnels de santé qui le prennent en charge [Deléger and Névéol, 2014]. Le contenu médical peut ensuite être analysé en sections selon un modèle international de présentation des informations cliniques appelé Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [1].

Mon travail s'est principalement concentré sur les trois étapes suivantes, qui comprennent l'extraction d'entités, de relations et de contexte. Dans le cadre du stage de M2 de Anne-Dominique Pham, en collaboration avec Thomas Lavergne et des collègues de l'Hôpital Européen Georges Pompidou (HEGP), nous avons étudié la contribution d'analyses de contenu approfondies pour la classification de compte-rendus de radiologie pour la détection de trois pathologies : la thrombose veineuse profonde, l'embolie pulmonaire et l'incidentalôme. Ce travail montre que des performances de classification satisfaisantes peuvent déjà être obtenues pour les deux premières pathologies avec une représentation du texte en sac de mots. L'extraction de sections (OMOP), d'entités, de relation et de modalités permettent néanmoins des améliorations de quelques points de F-mesure. La contribution de ces analyses est nettement plus significative en ce qui concerne la détection d'incidentalomes [Pham et al., 2014]. Ce résultat a motivé l'approfondissement d'un travail sur la définition d'un schéma d'annotation [Deléger et al., 2017] décrit en section 2.1.1. Par la suite, des travaux de nos collègues de l'HEGP ont également montré la contribution de la détection de la négation et du contexte familial pour améliorer la recherche d'information au sein du dossier électronique patient [Garcelon et al., 2017].

On peut noter que du côté application, l'analyse attendue doit fournir des résultats à différents niveaux de granularité : au niveau du document (par exemple, un article est-il pertinent pour inclusion dans une revue systématique ?), au niveau du patient (le patient présente-t-il un phénotype donné ?), au niveau d'une collection de documents (quelle est la prévalence d'un phénotype dans une cohorte ?). Quelle que soit la granularité finale attendue, le résultat peut s'appuyer sur une analyse au niveau de l'énoncé, suivie d'une agrégation au niveau du document puis de la collection si nécessaire. Ces aspects seront développés dans le chapitre 4.

1. `https://www.ohdsi.org/data-standardization/the-common-data-model/`

3.1 Revues de la littérature sur le traitement automatique de la langue clinique

Une partie importante de mon activité ces dernières années a été consacrée à une veille de la littérature en traitement automatique de la langue biomédicale, et en particulier en traitement automatique de la langue clinique. Au cours de ce travail de veille j'ai pu me rendre compte qu'aucune synthèse n'était disponible pour un certain nombre de sujets. La réalisation de revues de la littérature pour un sujet pluridisciplinaire comme le TAL biomédical soulève le problème méthodologique de la recherche de travaux pertinents pour le domaine : comment les recenser ? Comment définir le périmètre du domaine en termes de thématique, en termes de qualité des travaux examinés ?

TAL biomédical dans les langues autres que l'anglais. Les manques observés dans la littérature ont donné lieu à la rédaction de synthèses sur la disponibilité des ressources terminologiques biomédicales pour le français [Névéol et al., 2014b] ainsi que sur la disponibilité des corpus parallèles dans le domaine biomédical [Névéol et al., 2018]. J'ai également proposé et animé des tables rondes sur le traitement automatique de la langue clinique dans les langues autres que l'anglais [Névéol et al., 2014a, Névéol et al., 2017], qui ont débouché sur la rédaction d'un article de revue co-écrit avec les collègues ayant participé aux tables rondes [Névéol et al., 2018]. Le Tableau 3.1 donne une vue d'ensemble du résultat de cette revue et présente les méthodes abordées dans chaque langue.

L'analyse de ces publications a permis de formuler des recommandations afin de poursuivre les avancées de la recherche en TAL clinique pour les langues autres que l'anglais, et d'identifier les défis du domaine. Il est non trivial d'identifier les publications concernant des travaux en TAL clinique dans des langues autres que l'anglais. Ainsi, un effort collectif de cartographie des travaux dans ce domaine pourra permettre à la communauté de suivre les progrès grâce à une revue systématique de la littérature mise à jour régulièrement. Par ailleurs, on note un décalage important dans le volume et la quantité de ressources disponibles par rapport à l'anglais (ressources terminologiques et corpus annotés). Le défi le plus important à l'heure actuelle reste de soutenir les efforts de création de suites d'outils de TAL modulaires et multilingues afin de faciliter l'analyse des textes. Un moyen de parvenir à cet objectif peut être d'encourager également l'organisation de campagnes dans des langues autres que l'anglais à l'instar des campagnes CLEF eHealth ou BARR [Intxaurrondo et al., 2017].

TAL clinique. En parallèle, j'ai également assuré avec Pierre Zweigenbaum le rôle de co-éditeur de la section sur le TAL clinique du Yearbook de l'IMIA (International Medical Informatics Association) de 2015 à 2018. Cette tâche consiste à effectuer une revue systématique de la littérature sur la thématique de la section au cours de l'année écoulée afin d'en faire une synthèse et de sélectionner les "meilleurs articles" de l'année. Ces articles sont présentés dans le Yearbook afin de proposer à la communauté une vue d'ensemble de travaux récents. Ils peuvent être vus comme une sélection de suggestions à commenter en groupe de lecture.

Ce travail a été l'occasion d'une réflexion sur des stratégies de recherche pour identifier les publications en traitement automatique de la langue clinique, ce qui n'est pas une tâche simple du fait de la nature pluridisciplinaire du domaine. Nous avons commencé par définir le "TAL clinique" comme un ensemble de travaux en traitement automatique du langage naturel appliqué à des textes cliniques ou visant un résultat clinique. Cela englobe clairement le TAL appliqué aux textes des dossiers électroniques patient ; une

grande partie des travaux en extraction d'information pour l'aide à la décision rentre dans ce cadre. Nous avons également considéré comme des applications cliniquement pertinentes l'analyse automatique de textes ou de discours produits par des patients à des fins de diagnostic ou l'analyse de la communication axée sur le patient, ce qui conduit à fournir de meilleures informations sur la santé au public [Névéol and Zweigenbaum, 2015]. Nous avons utilisé PubMed et l'ACL Anthology [2] comme points d'entrée vers la littérature. Notre méthode de sélection des articles a évolué au fil des années afin d'augmenter le nombre de documents examinés. En 2015, nous utilisions une requête PubMed orientée vers la précision qui utilisait des termes MeSH. Afin d'éviter tout biais par rapport à l'avancement de l'indexation MEDLINE, dès 2016 nous avons modifié la requête PubMed afin de n'utiliser que des termes en langue naturelle. Nous avons également commencé à utiliser l'outil BibReview [3] afin de garder une trace des évaluations et notes prises sur les articles [Névéol and Zweigenbaum, 2016]. A partir de 2017, nous avons introduit dans BiBreview une présentation ordonnée des articles en utilisant les résultats du travail de Christopher Norman sur la sélection d'articles pour les revues systématiques, décrit au chapitre 4 [Névéol and Zweigenbaum, 2017]. Finalement, en 2018, nous avons étendu notre requête PubMed pour inclure le mot clé "text mining" [Névéol and Zweigenbaum, 2018]. Sur cette dernière année, notre revue de la littérature nous a conduit à examiner un total de 709 articles (dont 4,3% concernaient une étude sur une langue autre que l'anglais). La figure 3.2 présente la distribution de ces articles en fonction des sources, ainsi que leur pertinence.

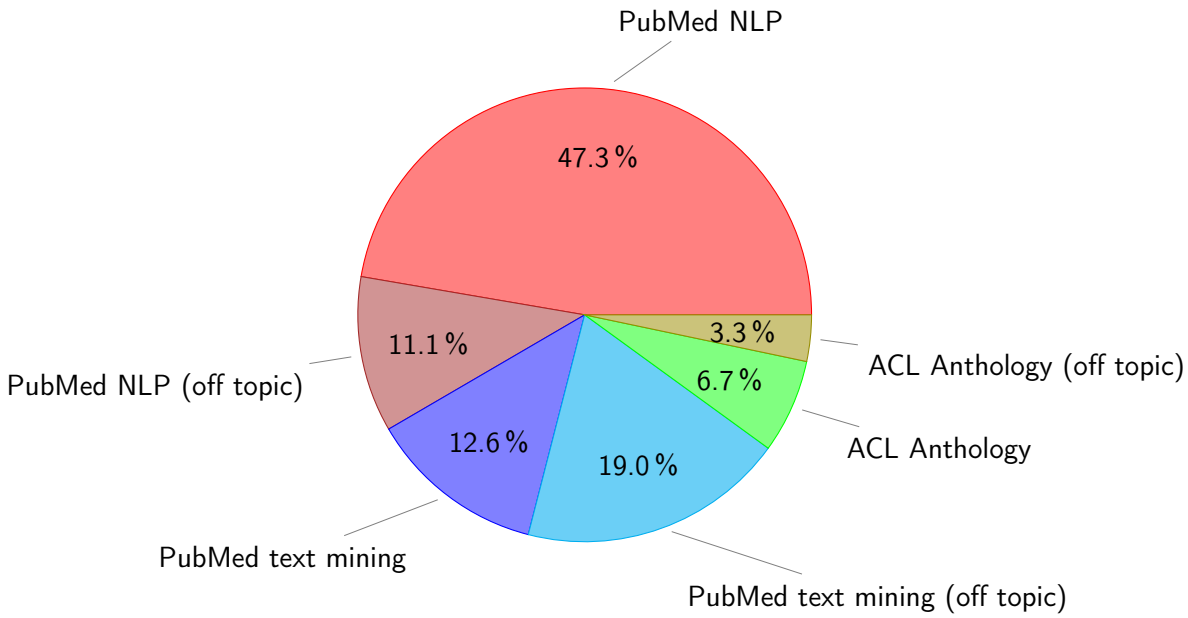


FIGURE 3.2 – Caractérisation des articles examiné en 2018 pour la section clinical NLP du Yearbook de l'IMIA.

On peut observer que le volume d'articles recueillis à l'aide de PubMed, principalement des articles publiés dans des revues médicales, est largement supérieur au volume d'articles recueillis dans l'ACL Anthology.

Il est intéressant de constater que si les requêtes "PubMed NLP" et "ACL Anthology"

2. http://aclasb.dfki.de/)
3. https://pypi.org/project/BibReview/

sont globalement plus précises (65% d'articles portent bien sur le TAL clinique), la requête "PubMed Text Mining" permet de retrouver un nombre non négligeable d'articles pertinents (N=90). L'accord inter-annotateur mesuré lors de ces revues était relativement élevé (de l'ordre de 85% en fonction des années et des catégories d'articles). Ainsi, la définition du domaine, ainsi que la notion d'article "intéressant" pour le domaine semble stable.

Campagnes d'évaluation. Comme indiqué au chapitre 2, les campagnes d'évaluation sont un levier important des avancées méthodologiques pour la communauté scientifique. Ainsi, mon travail s'est appuyé sur ce dispositif à plusieurs reprises afin d'explorer de nouvelles problématiques, ou de catalyser un effort méthodologique en utilisant des ressources disponibles et profiter de l'émulation de la collectivité sur une thématique particulière.

La figure 3.3 présente une chronologie des campagnes d'évaluation en TAL biomédical sur les vingt dernières années, en mettant en évidence ma contribution en tant que participant ou organisateur. On peut remarquer que les campagnes offrant des données cliniques ont démarré quelques années après celles portant sur les données de la littérature, en raison du délai nécessaire pour mettre en place un dispositif de partage de ces données. On peut également constater que la plupart des campagnes portent sur l'anglais, même si d'autres langues viennent diversifier l'offre depuis quelques années.

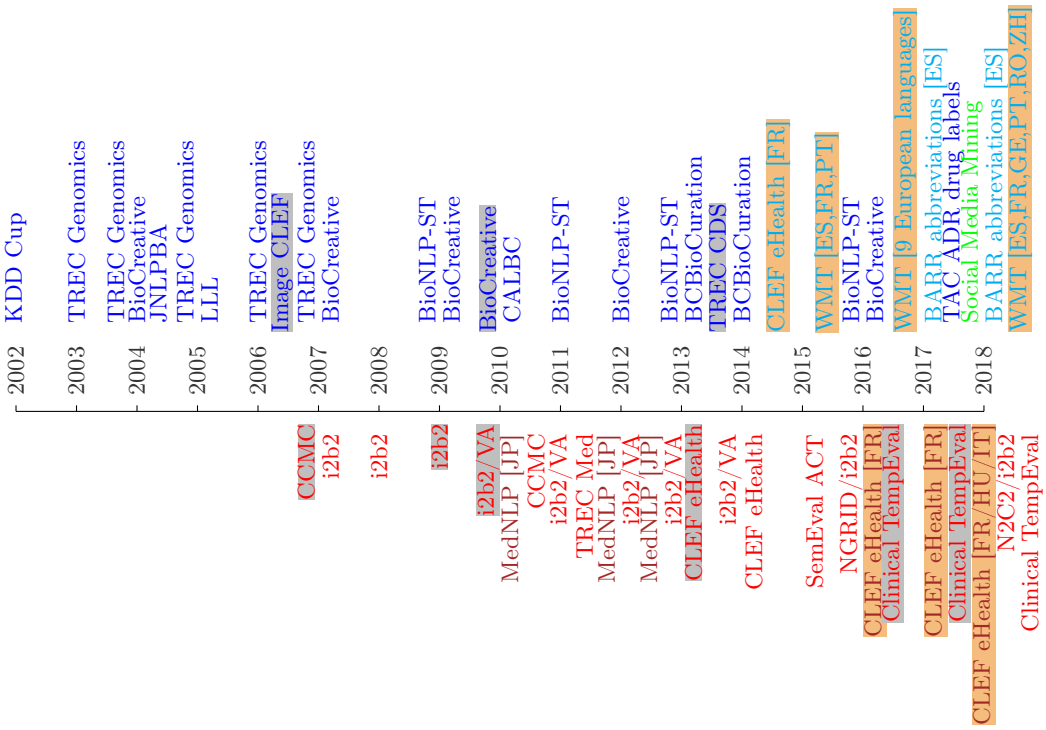


FIGURE 3.3 – Chronologie des campagnes d'évaluation dans le domaine biomédical. La partie haute de la figure présente les campagnes diffusant des textes issus de la littérature en anglais (en bleu foncé) ou dans d'autres langues (en bleu turquoise). Une campagne fondée sur des textes issus des réseaux sociaux figure en vert. La partie basse de la figure présente les campagnes diffusant des textes cliniques en anglais (en rouge vif) ou dans d'autres langues (en rouge Bordeaux). Les campagnes auxquelles j'ai participé sont grisées, celles que j'ai contribué à organiser sont sur fond jaune.

3.2 Analyse linguistique

Une rapide analyse des méthodes d'extraction d'entité (présentée à la section 3.3) montre que des étapes d'analyse linguistique telles que le découpage en phrase ou l'analyse morphosyntaxique sont des préliminaires nécessaires à l'extraction d'entités. Ainsi, je résume ici le travail accompli sur ces étapes pour les textes biomédicaux en français. D'autres éléments identifiés dans la troisième étape de la figure 3.1 sont également présentés. Chaque paragraphe reprend des éléments des publications associées à la conférence TALN et aux ateliers LOUHI et BioTxtM.

Identification de redondances. La question de la redondance dans les documents cliniques pose des problèmes sur le plan de l'analyse linguistique et de l'interprétation clinique des dossiers. La redondance dans les documents cliniques a un effet négatif sur les performances des outils de traitement automatique du langage naturel utilisés pour analyser les documents cliniques [Cohen et al., 2013]. De plus, la présence de différentes versions d'un même document dans un dossier patient demande de pouvoir identifier la version la plus complète et la plus à jour afin prendre en compte les informations les plus pertinentes concernant le patient.

Par *redondance*, nous désignons des portions de texte qui se retrouvent à l'identique ou quasiment à l'identique à différents endroits d'un même corpus clinique. Il y a tout d'abord une dimension opérationnelle pour repérer les documents doublons qui ne sont pas toujours des copies identiques. Dans le cadre d'une campagne d'annotation, un repérage plus fin des redondances internes au documents peut s'avérer utile afin de maximiser la diversité du contenu annoté (nous nous sommes appuyé sur le repérage de phrases redondantes pour maximiser la diversité des annotations morpho-syntaxiques, comme décrit dans le paragraphe ci-dessous). Dans le cadre d'une analyse temporelle, il peut également être utile de s'appuyer sur les redondances partielles pour identifier les versions mises à jour et complétées de certains documents pour reconstituer l'historique du patient. L'ampleur des redondances identifiées peut également avoir un impact sur l'analyse automatique des textes. Il est donc d'autant plus important de caractériser tout nouveau corpus en ce sens afin de le traiter de manière adaptée. Pour détecter ces redondances, nous avons exploré des méthodes d'alignement de séquence comme BLAST (Basic Local Alignment Search Tool) et, à l'occasion d'un séjour invité au LIMSI d'Efstathios Stamatatos (University of the Aegean, Grèce), des méthodes de détection de plagiat.

Pour comprendre l'origine possible des redondances dans les dossiers électroniques patient (DPE), il faut revenir aux modalités de création de ces documents, qui sont issus soit de la création directe de documents électroniques dans les systèmes d'information de santé des hôpitaux (c'est par exemple le cas du corpus MERLoT), soit par la numérisation de documents papier historiques (c'est par exemple le cas du corpus utilisé dans le projet ANR ACCORDYS). Le format électronique des notes cliniques présente des avantages évidents en termes de stockage et de partage, mais facilite également la duplication d'informations d'un document à un autre par un simple clic.

Ainsi, nous nous sommes intéressés à la détection de deux types de redondances correspondant à la réutilisation de textes dans les DPE [D'hondt et al., 2016] : 1) la détection de versions mises à jour du même document et 2) la détection de doublons de documents qui comportent des différences de surface dues au traitement OCR de numérisation, où au traitement de désidentification des documents pour une utilisation secondaire à des fins de recherche. Nous présentons une méthode robuste de détection de réutilisation de texte pour identifier automatiquement la réutilisation de texte dans des paires de documents

dans deux corpus français de DPE qui atteignent une macro F-mesure globale de 0,68 et 0,60 respectivement et identifient correctement toutes les paires de documents redondantes.

Par ailleurs, des études antérieures ont montré que la redondance (induite par le copier-coller) peut atteindre des niveaux élevés dans les dossiers électroniques patient américains [Cohen et al., 2013]. Dans ce contexte, nous avons réalisé une étude préliminaire sur le niveau de redondance dans les DPE français [D'hondt et al., 2015]. Nous étudions l'évolution de la redondance dans le temps et son apparition en fonction des différents types de documents et sections d'un petit corpus composé de trois dossiers de patients (361 documents). Nous trouvons que les niveaux de redondance moyens dans notre sous-ensemble sont inférieurs à ceux observés dans les corpus américains (respectivement 33% contre 78%), ce qui peut indiquer des pratiques médicales différentes entre ces deux pays. De plus, nous ne trouvons aucune preuve de l'augmentation progressive (heures supplémentaires) du texte redondant dans les notes cliniques qui a été trouvée dans les DPE américains. Ces résultats suggèrent que des stratégies d'atténuation de la redondance peuvent ne pas être nécessaires lors du traitement des DPE français.

Identification de sections. Dans le cadre du développement du corpus MERLoT, une étape de pré-traitement des documents a consisté à séparer le contenu médical des documents et un contenu non médical constitué d'informations administratives contenues par exemple dans les entêtes et pieds de page [Deléger and Névéol, 2014]. En effet, de nombreux documents cliniques du corpus contiennent des passages indiquant les coordonnées du service hospitalier ayant produit le document, ainsi qu'une liste des médecins susceptibles d'intervenir dans le service. Ces informations ne sont pas spécifiques au patient ni aux éléments de la prise en charge décrits dans le document.

Nous avons proposé une méthode d'identification des zones de haut niveau (en-tête administratif, en-tête patient, pied de page, contenu médical) des textes cliniques, qui s'appuie sur les travaux de la littérature en détection de sections. Nous présentons un système d'identification automatique de zones dans les documents cliniques via la classification supervisée de lignes à l'aide d'un modèle CRF utilisant l'outil Wapiti. Le système offre une F-mesure de 0,97, équivalente à l'accord inter-annotateur de 0,98. Notre étude montre que le contenu médical ne représente que 60% du contenu total de notre corpus, ce qui justifie la nécessité d'une segmentation en zones.

Découpage en phrases. Dans le cadre du projet CABerNeT, nous nous sommes intéressés à la segmentation en phrases de textes du domaine biomédical. L'essentiel de ce travail a été décrit dans [Boyer and Névéol, 2018]. La segmentation en phrases, aussi appelée "détection de frontière de phrase" est l'une des premières étapes des chaînes de traitement du langage naturel, sur laquelle repose les étapes suivantes telles que la segmentation en mots (ou *tokenisation*), l'étiquetage morpho-syntaxique, la reconnaissance d'entités nommées. Les performances élevées obtenues pour les corpus journalistiques en anglais font que la segmentation en phrases est globalement considérée comme un problème résolu [Kiss and Strunk, 2006]. Cependant, les bons résultats obtenus sur le domaine général ne se maintiennent pas toujours sur des domaines spécialisés, ce qui a des répercutions sur l'ensemble des étapes postérieures dans une chaîne de traitement. Certaines spécificités de la langue biomédicale peuvent en effet présenter des difficultés pour l'identification des frontières de phrase : l'emploi de termes spécialisés comme les noms d'organismes peuvent contenir des marques de ponctuation (e. coli), une forme hétérogène dans les documents cliniques où les marqueurs classiques de début et fins de phrases sont souvent omis (majuscules, signes de ponctuation). Dans ce travail, nous avons évalué cinq outils de segmentation en phrases sur

trois corpus issus de différents domaines. Nous ré-entraînons l'un de ces outils sur le corpus clinique MERLoT pour étudier l'adaptation en domaine. La détection de frontières de phrase à l'aide d'un modèle OpenNLP entraîné sur un corpus clinique offre une F-mesure de .73, contre .66 pour la version standard de l'outil, entraîné sur le French Tree Bank.

Etiquetage morpho-syntaxique. L'étiquetage morpho-syntaxique dans le domaine biomédical, et plus particulièrement clinique, a fait l'objet de plusieurs stages (niveau L3, puis M1) de Christiane Rabary, co-encodrée par Thomas Lavergne et moi-même [Rabary et al., 2015]. Le but de ce travail était de mener une réflexion sur l'adaptation en domaine pour l'étiquetage morpho-syntaxique et de proposer une ressource annotée pour le français clinique afin de mettre en œuvre des méthodes statistiques d'étiquetage séquentiel. Nous nous sommes appuyés sur des travaux utilisant des corpus cliniques en anglais, qui montrent le potentiel d'adaptation à ce domaine lorsque des corpus spécialisés, même de taille modeste, sont disponibles [Pakhomov et al., 2006, Liu et al., 2007, Ferraro et al., 2013]. Une étude avec l'outil statistique MedPOST caractérise plus spécifiquement l'apport des données annotées et des ressources lexicales spécialisées riches pour l'adaptation d'un étiqueteur au domaine biomédical [Smith et al., 2006]. A partir de ces résultats, un gain de performance significatif peut être obtenu dans un système de traduction automatique de documents biomédicaux utilisant un étiqueteur morpho-syntaxique adapté au domaine [Pécheux et al., 2014]. Pour le français, le développement du corpus Sequoïa [4] a abordé l'adaptation en domaine pour des outils d'analyse syntaxique. Le corpus comporte notamment deux rapports publics européens discutant la mise sur le marché d'un médicament. Cependant, la principale ressource annotée morpho-syntaxiquement pour le français reste le corpus journalistique French Tree Bank [Abeillé et al., 2003], à la fois en termes de taille du corpus et de complexité du jeu d'étiquettes utilisé.

Pour cette étude, nous avons utilisé des documents du corpus MERLoT afin de l'enrichir au niveau morpho-syntaxique. Des travaux sur le développement de corpus clinique en anglais annoté morpho-syntaxiquement ont montré qu'il était possible de minimiser la taille du corpus annoté grâce à des heuristiques de fréquence simples [Liu et al., 2007]. Par ailleurs, les études précédentes menées sur notre corpus ont montré que certaines parties des documents étaient redondantes et pouvaient présenter un intérêt limité pour l'analyse clinique et morpho-syntaxique (par exemple, en-têtes des documents). Ainsi, afin d'optimiser les efforts d'annotation dans notre étude, nous avons procédé à une sélection de phrases à l'intérieur des documents afin de concentrer le travail sur des phrases pertinentes et non-redondantes entre elles. La sélection a été opérée à l'aide d'un outil libre développé par Cohen et al. [Cohen et al., 2013]. Un jeu de 60 documents a été préannoté avec un modèle CRF [5] entrainé sur les données du French Tree Bank et les données médicales du corpus Sequoïa. Les annotations disponibles pour ces corpus ont été converties vers le jeu d'étiquettes *Multitag*, plus simple à appréhender pour notre travail. Seules les phrases sélectionnées ont été pré-annotées, et présentées à un linguiste pour correction à l'aide de l'outil BRAT [Stenetorp et al., 2012].

Le système CRF atteint 0.80 de taux d'erreur sur l'ensemble des étiquettes ce qui est relativement faible pour une tâche d'analyse morpho-syntaxique générale mais est raisonnable pour un système non adapté. L'analyse des erreurs d'étiquetage faites par l'outil générique a mis en évidence deux principales caractéristiques du domaine clinique : notre évaluation indique qu'environ 41% des erreurs de pré-annotation sont dues à des particularités de tokenisation et 22% sont liées au vocabulaire médical.

4. `http://deep-sequoia.inria.fr/` [Candito and Seddah, 2012]
5. Grâce à l'outil Wapiti [Lavergne et al., 2010]

Ainsi, le *vocabulaire spécialisé* dénote des connaissances à apporter à l'étiqueteur grâce à des lexiques spécialisés et des données étiquetées. Les difficultés rencontrées sont de différentes natures. On trouve naturellement des termes spécialisés traités par le système comme des mots inconnus. On rencontre également le problème de l'homographie lorsque la catégorie d'un terme connu dans le domaine général change dans le domaine médical. Par exemple, le token "patient" qui est majoritairement étiqueté comme un adjectif dans les textes du domaine général sera plus souvent étiqueté comme un nom dans les textes cliniques. Enfin, on rencontre également des termes inconnus du système dont la graphie induit une erreur d'interprétation. Par exemple, dans le syntagme *confection d'un Hartmann*, le dernier token doit être annoté comme un nom commun contrairement à ce que la majuscule suggère car il s'agit d'une marque de dispositifs médicaux (cas similaire à l'emploi du terme *Frigidaire* dans la langue générale).

On note également un besoin d'une *tokénisation particulière* pour des phénomènes linguistiques particulièrement prévalents dans les textes cliniques, tels que les posologies, mesures et abréviations. On trouve deux types d'abréviations. D'une part des abréviations partielles, telles que *chir.* pour *chirurgie*, qui ne font pas partie des listes classiques d'abréviations et sont donc inconnues à la fois du tokeniseur et du modèle CRF. La ponctuation qui doit être considérée comme faisant partie du token est généralement séparée car reconnue comme une ponctuation finale. Le token se trouve donc incorrectement étiqueté et le marqueur de fin de phrase a tendance à propager cette erreur aux mots suivants. D'autre part, de nombreux termes médicaux tels que *anesthésie générale* ou *sérum glutamooxaloacétate transférase* sont complètement abrégés en *A.G.* et *SGOT*. Même si la morphologie de ces tokens permet de les reconnaître plus simplement, leur regroupement sous une seule étiquette « abréviation » est ici peu approprié, ces termes étant souvent porteurs d'une information sémantique importante pour l'analyse des documents médicaux. Le choix le plus approprié est de réaliser une tokenisation assurant un découpage en tokens similaire à ceux du terme non abrégé ainsi qu'un étiquetage complet de la séquence. L'abréviation *A.G.* est donc annotée *A.*`:NC` *G.*`:ADJq` au contraire de *A.G.*`:NP` suggéré par le système de pré-annotation. Une deuxième particularité du domaine médical est l'abondance de quantités et de mesures. Si certaines sont simples, comme *3mm*, et sont correctement analysées par un système non-adapté au domaine, ce n'est pas le cas pour les plus complexes telles que *3x/j* ou *5,4 mmoles/l*. De plus, une même mesure peut-être abrégée de manière différentes, pour *3 fois par jour* par exemple, nous avons observé les abréviations suivantes dans notre corpus : *3 fois/jours*, *3 fois/j*, *3x/j*, *3/j*...

On trouve aussi des termes ressemblant à la fois à des abréviations et des mesures tels que *T2N+* dans *lésion du rectum T2N+* qui indique le degré d'évolution d'un cancer. Il s'agit de la classification TNM (« tumor, nodes, metastasis » en anglais) qui prend en compte la taille et la localisation de la tumeur primitive (notée parfois pT), le nombre et le site des ganglions lymphatiques régionaux qui contiennent des cellules cancéreuses, et la propagation du cancer, ou métastases, vers une autre partie du corps. Ces deux types de termes : *3x/j* et *T2N+*, ont tendance à être considérés comme un seul token par la chaîne de traitement non-adaptée au domaine. Comme pour les abréviations simples, il est pourtant pertinent ici de les décomposer afin de les étiqueter de manière similaire à leur écriture non-abrégée. On annotera donc *3*`:Det` *x*`:NC` */*`:Prep` *j*`::NC` et *T*`:NC` *2*`:ADJc` *N*`:NC` *+*`:ADV`. Cette annotation complète bien que plus coûteuse et demandant des connaissances médicales dans certains cas permet de faciliter les étapes suivantes de l'analyse automatique de ces documents.

Une contribution importante de ce travail est le développement d'un corpus du domaine clinique annoté morpho-syntaxiquement, dans le but d'entraîner et d'évaluer un outil d'éti-

quetage spécialisé. Cependant, notre analyse des données a mis en évidence la nécessité d'approfondir la problématique de l'étiquetage morpho-syntaxique des textes biomédicaux avant qu'un outil d'étiquetage puisse effectivement être développé. Il faut tout d'abord mettre en œuvre une tokénisation spécifique au domaine, définir un guide d'annotation adapté à cette tokénisation afin de proposer des ressources annotées correspondantes.

3.3 Extraction d'entités

L'extraction d'entités nommées est l'une des tâches élémentaires de l'extraction d'information à partir de textes. Elle consiste à reconnaître des mentions textuelles dénotant des objets ou concepts appartenant à des catégories prédéfinies. Dans le domaine général il s'agira par exemple de noms de personnes, de lieux ou d'entreprises. En domaine de spécialité, par exemple dans le domaine biomédical, les entités reflètent les catégories des taxonomies du domaine, comme par exemple décrit dans les modèles présentés au chapitre 2. Il faut également remarquer qu'en pratique des entités définies de manière similaire peuvent être annotées différemment en corpus. Par exemple, dans une étude comparant la détection de la négation dans cinq corpus en anglais, Wu et al. indiquent que les frontières d'entités sont envisagées différemment selon les corpus étudiés, par exemple pour ce qui est de l'inclusion de pronoms : *her shortness of breath* vs. *shortness of breath* [Wu et al., 2014].

Méthodes d'extraction d'entités. Ainsi, dans le domaine biomédical, les méthodes d'extractions d'entités s'appuient souvent sur des bases de connaissances telles que l'UMLS afin d'effectuer un appariement entre les mentions rencontrées dans un texte et les instanciations des concepts répertoriées dans la base de connaissance. Ces méthodes ont un degré de complexité croissant qui va de l'appariement exact à la prise en compte de variations terminologiques de difficulté croissante telles que la casse, l'orthographe, la morphologie, la syntaxe, la sémantique (synonymes), voire la langue. Ainsi, le BioAnnotator [Jonquet et al., 2009] apparie les termes et synonymes de concepts recensés dans l'UMLS à des textes biomédicaux en anglais. Récemment, de nouvelles fonctionnalités permettent à cet outil de prendre en compte certains éléments de contexte comme la négation et de traiter des textes en français [Tchechmedjiev et al., 2018]. MetaMap, un outil développé pour l'annotation des textes de la littérature biomédicale en anglais à l'aide de concepts de l'UMLS [Aronson and Lang, 2010] met en oeuvre des méthodes de traitement automatique de la langue plus complexes. La chaîne de traitement procède à un découpage en phrases, suivi de l'étiquetage morpho-syntaxique des phrases, puis de chunking qui permet d'identifier des segments sous-phrastiques susceptibles de constituer ou de contenir des entités nommées. Ces segments sont ensuite traités pour être appariés à des termes ou synonymes associés à des concepts de l'UMLS. L'appariement peut être direct, ou s'appuyer sur des règles linguistiques de variation terminologique. cTAKES [Savova et al., 2010] est un outil open source dédié à l'analyse de textes cliniques en anglais qui met en œuvre une chaîne de traitement similaire à MetaMap, adaptée au contexte clinique. L'extraction d'entités nommées repose néanmoins sur un appariement plus direct. MedLEE [Friedman et al., 1995] est le premier système à avoir démontré le potentiel de méthodes de traitement automatique de la langue pour l'analyse de textes cliniques. Il s'agit d'un outil propriétaire qui met en œuvre une analyse fondée sur le principe des cadres sémantiques (*frames*), et intègre des sources terminologiques au delà de l'UMLS, comme par exemple Bi-RADS. Pour le français, on peut enfin mentionner le système ECMT (Extracting Concepts with Multiple Terminologies) [Sakji et al., 2010] qui met en œuvre un algorithme de recherche d'information pour

effectuer l'appariement entre texte et ressource terminologique [Tutubalina et al., 2018].

Un autre type de méthode de reconnaissance d'entité est la reconnaissance de séquences fondée sur l'apprentissage statistique. Ces méthodes reposent sur des corpus annotés dans lesquels un grand nombre de mentions d'intérêt sont annotées. Dans ce groupe de méthodes, on trouve notamment les modèles de Markov Cachés (HMM), les champs aléatoires conditionnels (CRF) et plus récemment les réseaux de neurones (par exemple, RNN). Les HMM et CRF reposent sur l'affinage de traits fournis au modèles, tandis que les réseaux neuronaux s'appuient principalement sur des plongements lexicaux issus de gros corpus non annotés. Les systèmes développés à partir de ces méthodes s'appuient sur des outils génériques tels que Wapiti ou MALLET [6]. Dans le cadre de la thèse de Julien Tourille, les besoins de nos expérimentations ont conduit Julien à développer l'outil libre YASET [7], qui propose une implémentation de l'algorithme neuronal bi-LSTM à l'aide de la librairie TensorFlow. Cet outil offre des temps de traitement très inférieurs à NeuroNER [Dernoncourt et al., 2017], qui était le seul outil de ce type disponible au début de la thèse.

On peut également noter l'émergence de méthodes neuronales d'encodage/décodage [Sutskever et al., 2014] dans le domaine de la traduction automatique. Dans le cadre de la reconnaissance d'entités, ces méthodes s'appuient sur les bases de connaissances, tout comme les méthodes d'appariement. Les modèles construits cherchent à "traduire" une séquence textuelle (typiquement, une mention) en l'un des termes ou synonymes recensés dans la ressource terminologique.

Contributions. Mes travaux se sont appuyés sur ces différents types de méthodes, en fonction du contexte d'étude spécifique. Par exemple, dans le cadre de la participation de la NLM aux campagnes i2b2 2009 et 2010 pour l'extraction de médicaments [Mork et al., 2010] puis de test, problèmes et traitements[Demner-Fushman et al., 2010], l'extraction d'entité réalisée reposait sur MetaMap associé à des règles de filtrage et d'ajustement des frontières d'entités. Dans le cadre du développement du corpus MERLoT, nous avons utilisé deux méthodes d'extraction d'entité afin de fournir une pré-annotation aux annotateurs : il s'agissait d'abord d'une méthode d'appariement simple à base de dictionnaire, puis d'une méthode statistique supervisée (CRF) [Campillos et al., 2018].

Ces différentes expériences montrent que l'exploitation de méthodes existantes, parfois simples (projection de dictionnaire) peuvent donner des résultats satisfaisants lorsqu'elles sont correctement adaptées au corpus et au besoin d'analyse. Cependant, cela soulève la question de l'adaptation : quel est le coût de l'adaptation, en termes de temps d'ajustement des outils, temps de développement de corpus annoté ? Ce sont des questions auxquelles nous avons essayé de répondre dans le cadre du stage de M2 de Mike Tapi Nzali, co-encadré avec Xavier Tannier, sur l'analyse temporelle des documents cliniques. Nous nous sommes tout d'abord intéressés à l'extraction d'expressions temporelles dans des textes cliniques en français. Pour ce faire, nous avons comparé une méthode à base de règle, Heideltime, et une méthode d'apprentissage supervisé, un modèle CRF implémenté à l'aide de Wapiti. Notre meilleur système statistique offre une performance de 0,91 de F-mesure, surpassant pour l'identification d'expressions temporelles le système état de l'art Heideltime [Nzali et al., 2015]. Ce travail a également montré qu'une douzaine de règles suffisent à adapter la version d'Heideltime développée pour traiter les textes journalistiques aux textes cliniques. Pour la suite de ce travail nous avons donc comparé l'effort d'adaptation et les performances obtenues pour l'extraction d'expressions temporelles dans plusieurs domaines (presse, textes historiques, textes cliniques) [Tapi Nzali et al., 2015]. Nous avons

6. `http://mallet.cs.umass.edu`
7. Yet Another Sequence Tagger `https://github.com/jtourille/yaset`

observé que la méthode à base de règle implémentée par Heideltime offre globalement des performances plus stables sur l'ensemble des domaines étudiés. Concernant la méthode statistique, on constate comme on pouvait s'y attendre que les meilleures performances sont obtenues pour des modèles entraînés sur des corpus de domaine. Néanmoins, nous avons également observé que les plus gros écarts de performance sont obtenus par des modifications de la méthode de segmentation des textes plutôt que grâce à une augmentation de la taille des corpus d'entraînement. Par ailleurs, la comparaison des annotations en expressions temporelles dans trois domaines met en évidence que si les mêmes types d'expressions temporelles sont présents dans l'ensemble de ces textes, la distribution des types d'expression est très différente.

Récemment, avec le développement de l'outil YASET, nous avons continué d'explorer la problématique de la généricité des méthodes en validant l'application du modèle LSTM implémenté dans YASET sur une variété de corpus du domaine biomédical [Tourille et al., 2018]. La sélection de corpus utilisée dans l'étude offre une couverture de plusieurs sous-domaines (littérature, clinique) de plusieurs langues (anglais, français) et de plusieurs types de séquences (étiquettes morphosyntaxiques, entités identifiantes, maladies, ...). Notre étude montre que le modèle offre en moyenne des performances proches de l'état de l'art sur l'ensemble des corpus, sans effort d'adaptation particulier.

3.4 Extraction de relations

L'extraction de relations est une des tâches d'extraction d'information intervenant en aval de l'extraction d'entités. Elle consiste à déterminer s'il existe une association typée entre plusieurs entités. De manière générale, on parle de relation *n-aire* en fonction du nombre n d'entités impliquées dans l'association. Dans le cas d'association entre deux entités, on parle de relation *binaire*. L'extraction de relations peut être générique, c'est à dire, s'intéresser à déterminer simplement s'il existe une relation entre deux entités ou pas, sans chercher à préciser la nature de la relation. L'extraction peut également être spécifique et s'intéresser à déterminer la nature de la relation, si elle existe. Par exemple, dans l'énoncé "Douleurs abdominales intermittentes soulagées par le PARACETAMOL", le prérequis pour l'extraction de relation est l'extraction des entités "Douleurs abdominales" (à laquelle on pourra associer la catégorie *Disorder*), "abdominales" (à laquelle on pourra associer la catégorie *Anatomy*) et "PARACETAMOL" (à laquelle on pourra associer la catégorie *Chemicals and Drugs*). L'extraction générique de relations pourra déterminer l'existence d'une relation entre les couples d'entités (Douleurs abdominales et abdominales), (Douleurs abdominales et PARACETAMOL) et l'absence de relation entre les entités du couple (abdominales et PARACETAMOL). L'extraction spécifique de relations pourra préciser que la relation (Douleurs abdominales et abdominales) est une relation de *localisation* alors que la relation (Douleurs abdominales et PARACETAMOL) est une relation de *traitement*. On peut remarquer que dans le premier cas, la catégorie des entités et l'existence d'une relation permet d'inférer la nature de la relation, si l'on dispose d'une ressource qui recense la liste des relations possible, comme les modèles présentés au chapitre 2. Dans le deuxième cas, le contexte est nécessaire pour déterminer la nature de la relation qui pourrait être *traitement* ou *cause* dans le cas d'effet secondaire du médicament.

Méthodes d'extraction de relations. Dans le domaine biomédical, une première catégorie de méthodes d'extraction de relations s'appuient sur des connaissances telles que le réseau sémantique de l'UMLS qui définit au niveau des catégories de concepts quelles sont les relations pour lesquelles il est possible de rencontrer une réalisation en corpus. L'ex-

traction des relations se fonde ensuite sur l'analyse syntaxique et sémantique des textes pour identifier les prédicats reliant les entités qui correspondent aux relations possibles. Par exemple, SemRep [Rindflesch and Fiszman, 2003] intègre MetaMap pour extraire des propositions en trois parties, appelées prédications sémantiques, à partir de phrases dans un texte biomédical en anglais. SemRep utilise les entités repérées par MetaMap et repère les prédicats en s'appuyant sur l'hypothèse qu'ils sont réalisés en anglais grâce à l'une des trois stratégies syntaxiques que sont les verbes, les appositions ou les modifieurs nominaux.

Un autre type de méthode de reconnaissance de relations, indépendante du domaine dans son principe, consiste à aborder l'extraction de relations comme une tâche de catégorisation : étant donné deux entités en présence, il s'agit de déterminer s'il existe une relation entre ces entités (classification binaire) ou de déterminer quelle est la nature de la relation entre les entités. Dans ce type de classification multi-classe, chaque relation correspond à une classe et on considère généralement une classe supplémentaire correspondant à l'absence de relation. La classification est effectuée à l'aide de méthodes supervisées construisant une représentation vectorielle des entités potentiellement en relation et de leur contexte à l'aide de divers traits syntaxiques, morphologiques, sémantiques. Plus récemment, les méthodes neuronales ont permis de réduire l'effort nécessaire à l'optimisation des traits en s'appuyant sur une représentation distributionnelle non supervisée du contexte à l'aide de plongements lexicaux. L'inconvénient de ces méthodes est qu'elles nécessitent cependant un volume important de données d'entraînement (corpus annotés) pour offrir des performances satisfaisantes.

Afin de s'abstraire de la nécessité de données annotées, des méthodes de supervision distante ont été exploitées afin de construire des corpus annotés de grande taille en faisant l'hypothèse que le bruit inhérent à l'absence de validation des annotations serait compensé par le volume global d'annotations correctes[Mintz et al., 2009]. Finalement, des travaux récents explorent des méthodes de transfert, qui permettent d'exploiter des données hors du domaine d'application néanmoins considérées comme suffisamment proches[Legrand et al., 2017].

Contributions. Dans le cadre du développement de ressources autour de la maladie et du médicament au NCBI, je me suis intéressée aux indications médicamenteuses. L'indication d'un médicament dénote la ou les maladies qu'un médicament peut traiter - un type d'informations fréquemment recherché par les utilisateurs de PubMed. Afin d'offrir une source unifiée d'information structurées sur les indications médicamenteuse, nous avons abordé ce problème sous l'angle de l'extraction de relations entre maladies et médicaments à partir de multiples ressources telles que les notices DailyMed et les *Scope Notes* du MeSH. L'extraction des relations est effectuée à l'aide de l'outil SemRep après un pré-traitement des textes pour la résolution des ellipses et des anaphores, deux stratégies destinées à augmenter le rappel. Le résultat de ce travail d'extraction et d'intégration de relations à partir de quatre sources différentes est une base de 7 670 relations TREATS uniques entre 4 666 médicaments et 1 293 maladies avec une précision globale estimée à 77% et une spécificité de 84% [Névéol and Lu, 2010]. Nous avons ensuite souhaité appliquer cette ressource pour guider l'extraction de relations entre médicaments et indications dans le cadre du challenge i2b2 2009. Malheureusement, nous avons constaté que la couverture de la ressource n'était pas suffisante pour avoir un impact sur le corpus de la campagne [Mork et al., 2010].

Dans le cadre de la campagne i2b2 2010 puis de Clinical Tempeval 2017, je me suis intéressée à l'extraction de relations dite de "de bout en bout" qui consiste à enchaîner l'extraction d'entités puis des relations entre ces entités. Cet enchaînement engendre généralement une baisse de performance mécanique lorsque l'extraction de relations est réalisée

sur des entités extraites automatiquement par rapport aux entités de référence. Lors de la campagne i2b2 2010, nous avons néanmoins pu montrer que grace à la sélection automatique de traits, il était possible de maintenir des performances élevées pour l'extraction de relations même fondée sur des entités extraites automatiquement [Islamaj Doğan et al., 2011].

Les campagnes Clinical Tempeval 2016 et 2017 ont également fourni une opportunité d'explorer deux aspects particulièrement intéressants de l'extraction de relations : la prise en compte d'un contexte au dela de la phrase et l'adaptation en domaine [Tourille et al., 2016] [Tourille et al., 2017b].

Dans le cadre de la thèse de Julien Tourille, un travail est en cours sur l'extraction de relations de coréférence dans des textes cliniques en anglais. Une difficulté de cette tâche est la taille du contexte en prendre en compte, puisque les chaînes de coréférences sont susceptible de couvrir l'ensemble d'un document et donc de dépasser largement le cadre de la phrase. Julien explore en particulier l'apport potentiel des relations temporelles pour la détection de coréférence.

3.5 Discussion

Les revues de la littérature effectuées mettent en évidence un déséquilibre dans les publications en TAL biomédical, qui paraissent majoritairement dans la communauté médicale, comparé à la communauté de traitement automatique de la langue. Ce constat peut s'expliquer en partie par le paradigme d'évaluation des chercheurs en TAL médical. Dès lors que ces collègues sont en poste dans des hôpitaux ou des laboratoires Inserm (en France) ou financés par un organisme médical (par exemple, le NIH aux États-Unis), l'évaluation de leur activité est fondée sur les publications indexées dans la base MEDLINE. Cela crée un biais de publication en faveur des revues et conférences médicales. On peut également remarquer dans les revues de la littérature effectuées que les travaux portent sur les méthodes et les applications du TAL biomédical. La communauté de Traitement Automatique de la Langue est principalement intéressée par les aspects méthodologiques, d'autant plus qu'ils peuvent se généraliser au delà du domaine biomédical. Ainsi, cela réduit le spectre des travaux publiables dans cette communauté.

La variété des textes et des sous-domaines d'application met en évidence la nécessité d'une réflexion sur des approches capables de gérer cette variété. Faut-il proposer des méthodes génériques, permettant d'analyser des textes situés à n'importe quel niveau du continuum entre langue générale et langue de spécialité ? Faut-il au contraire proposer des méthodes paramétrables en fonction de la langue et du sous-domaine concerné ? Dans ce cas, comment définir et régler les paramètres pertinents ?

Method/Task	**Number of references cited in [Névéol et al., 2018]**, per language
Core NLP	
- Morphology	FR 1 PL 1
- Part of Speech tagging	PT 1 ES 1
- Parsing	FI 3 FR 2 GR 1 JA 2
- Segmentation	DE 1 HE 1
Resource development	
- Lexicons	BG 1 EL 1 EU 1 FR 4 HE 1 JA 1 SV 1 ZH 2
- Corpora and annotation	EL 1 EN-{FR,ES} 1 EN-{ES,FR,PT} 1 ES 1 FR 2
- Models, methods	DE 1 FR 1
De-identification	FR 4 KO 1 SV 2
Information extraction	
- Medical Concepts	BG 2 ZH 2 DE 4 DU 1 ES 1
	IT 3 PL 1 SV 1
- Findings/Symptoms	DE 1, SV 2 ZH 1
- Drugs/Adverse events	BG 2 DA 1 ES 1 FR 2 SV 1
- Specific characteristics	EN-{ZH,FR,DE,JA,ES} 1 FR 1 ZH 1 DU 1
- Relations	BG 1 DE 2 IT 2
Classification	
- Phenotyping from EHR text	BG 1 ES 1 FI 1 FR 2 KA 1 PT 1 SV 1 ZH 1
- Indexing and coding	EN-FR 1 FI 1 FR 1 JA 2
- Patient-authored text	JA 1
- Cohort stratification	DA 1 DE 1
Context Analysis	DU 1
- Negation detection	BG 1 DA 1 DE 1 DU 1 ES 2 FR,DE,SV 1 SV 2
- Uncertainty/Assertion	SV 1 ZH 1
- Temporality	FR 1 SV 1
- Abbreviation	DE 1 SV 1
- Experiencer	DU 1
Multilingual tasks	
- Translation	EN-ES 1 EN-FR 1 EN-{KO,RU,ES,ZH} 1
	EN-{FR,DE,HU,PL,ES,TU} 1, FR-DE 1
- Information Retrieval	AR 1 FR 1, EN-{CZ,DE,FR} 1 EN-{ES,FR} 1
- Cultural analysis	DE 1, EN-ZH 1, FI-SV 1
Shared tasks	
- CLEF-ER2013	DE,DU,FR,ES-
- CLEF eHealth 2015, 2016	FR 2
- NTCIR 2014, 2016	JA 2

TABLE 3.1 – Overview of NLP methods used in studies and language(s) addressed according to [Névéol et al., 2018] ; ISO 639-1 language codes are used. When multiples languages are addressed in one paper we provide a comma separated list ; dashes mark language pairs in multilingual work.

Chapitre 4

Impact en épidémiologie et santé publique

La compréhension automatique de textes du domaine biomédical s'envisage naturellement dans le contexte applicatif de la recherche en biologie et en santé publique. Ainsi, les informations extraites automatiquement des textes biomédicaux doivent être exploitées pour répondre à des besoins applicatifs de la recherche biomédicale afin d'accélérer les découvertes scientifiques. Un premier défi est celui de l'accès à l'information de santé parmi la masse de documents disponibles pour le public et les professionnels de santé ou chercheurs. Un deuxième défi est l'utilisation secondaire du contenu des dossiers patients [1]. Ce chapitre va décrire ce qu'attendent les chercheurs et les médecins du TAL biomédical et faire un état des lieux de ce qui est réalisé ou réalisable à court terme, et ce qui reste une perspective. Deux thématiques sont particulièrement abordées : la recherche d'information et l'analyse rétrospective des dossiers patients.

4.1 Recherche documentaire

La masse de documents disponibles dans la littérature biomédicale fait de la gestion de la documentation un défi pour les bibliothèques médicales dont la mission est d'assurer la diffusion d'information de santé de qualité auprès du public, des professionnels de santé et des chercheurs. Ainsi, je me suis intéressée à diverses problématiques permettant d'une part d'organiser l'archivage des informations de santé telles que les articles scientifiques ou les données brutes de la recherche et d'autre part d'améliorer l'accès à l'information de santé par divers profils d'utilisateurs. Cette expertise a plus récemment été mise à contribution dans le cadre du projet Européen MIROR avec la thèse de Christopher Norman sur l'automatisation du processus de création de revues systématiques de la littérature.

4.1.1 Contribution à la constitution de bases de données

Développement d'outils d'indexation fine : extraction automatique de paires de descripteurs MeSH à partir de textes du domaine biomédical. Mon travail de thèse s'est concentré sur la caractérisation de documents du domaine biomédical par le biais de leur catégorisation [Névéol et al., 2004] et d'une indexation fine à l'aide de paires

1. Les dossiers patients sont avant tout créés dans le cadre de la prise en charge des patients. On parle d'*utilisation secondaire* des dossiers patients lorsque que les documents sont exploités après le passage du patient à l'hôpital pour des analyses dans le cadre de la recherche clinique, par exemple des analyses rétrospectives.

de descripteurs du thésaurus MeSH (Medical Subject Headings) [Névéol et al., 2006]. Dans ce domaine, j'ai notamment proposé lors de ma thèse une méthode innovante permettant d'extraire des paires mot clé/qualificatifs en plus des mots clés isolés pour l'indexation automatique de documents en français. J'ai poursuivi cette thématique de recherche lors de mon stage de post-doctorat au Lister Hill National Center for Biomedical Communications a ensuite été d'adapter cette méthode à l'analyse de textes en anglais, puis de la développer dans le contexte de l'indexation automatique de documents de la littérature pour la base MEDLINE. Le résultat de ce travail, détaillé dans [Névéol et al., 2007a, Névéol et al., 2007b, Névéol et al., 2008, Névéol et al., 2009], permet au logiciel d'indexation MeSH de la NLM (Medical Text Indexer [2], utilisé quotidiennement par 120 indexeurs pour inclure près d'un million de notices par an dans la base MEDLINE) de proposer des recommandations de qualificatifs affiliés ou non à des mots clés en plus des recommandations de mots clés isolés. Différentes approches issues du TAL, de la fouille de données et de l'apprentissage (collaboration avec V. Claveau, CNRS [Névéol et al., 2008]) ont été évaluées puis combinées pour cette tâche.

Évaluation des mots-clés proposés par les auteurs comme source d'indexation pour la base MEDLINE et comme enrichissement pour le thésaurus MeSH. En complément des méthodes d'indexation automatique, je me suis également intéressée aux mots-clés attribués par les auteurs comme source possible de termes d'indexation. Dans une étude réalisée en collaboration avec Rezarta Islamaj Dogan et Zhiyong Lu au NCBI, nous avons analysé le corpus PubMed Central Open Access afin de caractériser l'évolution de la disponibilité de mots-clés attribués par les auteurs pour les articles indexés dans MEDLINE, ainsi que leur proximité sémantique avec des termes d'indexation MeSH d'une part, et avec les termes MeSH attribués par les indexeurs dans MEDLINE[Névéol et al., 2010].

Du point de vue de l'indexation MEDLINE, les résultats de cette étude ont montré que la majorité (62%) des mots clés proposés par les auteurs dénotent des concepts déjà couverts par les termes d'indexation MeSH choisis par les indexeurs. Par exemple, lorsqu'un auteur proposait le mot clé *thiamethoxam*, qui fait partie des "Supplementary Concepts" du MeSH, le descripteur correspondant, *Thiazoles* était sélectionné par les indexeurs MEDLINE. Un grand nombre de cas de divergence entre les mots clés auteurs et les descripteurs choisis par les indexeurs sont attribuables à la variabilité inter-indexeur décrite dans des travaux antérieurs[Funk and Reid, 1983]. Par ailleurs, d'autres travaux offrant une analyse des annotations réalisées par les auteurs rapportent que les auteurs d'articles n'ont pas une connaissance aussi approfondie que les indexeurs des terminologies et des règles d'indexation, ce qui doit conduire à considérer leurs recommandations de termes d'indexation avec prudence [Hahn et al., 2007]. Une expérience de la revue *FEBS Letters* avait sollicité des annotations des auteurs pour les interactions entre protéines. L'évaluation des annotations fournies par les documentalistes spécialisés avait été globalement négative et débouché sur la conclusion que les auteurs n'étaient pas forcément compétents pour réaliser des résumés à base de mots-clés de leurs propres recherches [Lok, 2010].

L'une des raisons principales de l'absence d'équivalence entre un mot clé auteur et un terme d'indexation MEDLINE peut être l'absence de couverture du concept dénoté par le mot clé auteur dans le MeSH. Ainsi, du point de vue de l'enrichissement terminologique, notre étude a également montré que 49% des mots clés auteur sont soit absents du MeSH (33%) soit non relié au descripteur équivalent dans le MeSH (16%). Ce résultat suggère que les mots clés auteur peuvent avoir une contribution pour l'enrichissement terminolo-

2. `https://ii.nlm.nih.gov/MTI/`

gique. Ainsi, les mots clés auteur avec une distance sémantique faible avec des descripteurs MeSH pourraient être de bon candidats synonymes (par exemple, *CD8+ t-cells* pourrait être introduit comme synonyme du descripteur *CD8-Positive T-Lymphocytes*). De même, certains mots clés auteurs dénotant des concepts qui ne sont pas couverts dans le MeSH peuvent donner lieu à la création de nouveaux descripteurs (par exemple, *non-alcoholic steatohepatitis* a été introduit comme descripteur en 2015 ; *Systematic review as Topic* reste un candidat d'actualité). Cependant, d'autres termes utilisés par les auteurs dénotent des concepts qui peuvent être jugés comme hors sujet pour un thésaurus qui à pour vocation d'indexer la litérature biomédicale (par exemple, le mot-clé *robustness* dans le contexte de simulation par ordinateur). Un autre argument en faveur de l'examen des mots clés auteurs comme candidats termes pour l'enrichissement d'un thésaurus comme le MeSH est que les auteurs choisissent souvent ces termes sur la base de leur propre expérience de recherche d'information : les mots clés qu'il choisissent sont probablement des termes qu'ils utiliseraient dans leurs propres requêtes. Ainsi, l'inclusion de mots clés proposés par les auteurs dans une terminologie comme le MeSH pourrait améliorer les résultats de recherche d'information des utilisateurs de la base MEDLINE.

En conclusion, cette étude a établi dès 2010 que les mots clés proposés par les auteurs devenaient disponibles pour un nombre croissant d'articles indexés dans MEDLINE. Nous avons pu montrer l'intérêt de ces termes pour la recherche d'information, l'enrichissement de terminologie et proposer une méthode efficace pour établir leur proximité sémantique avec des descripteurs MeSH existants. Les résultats de cette étude ont contribué à la décision d'inclure les mots-clés des auteurs dans les notices MEDLINE en Janvier 2013 [3].

Découverte automatique de liens entre termes, concepts, documents, bases de données Les progrès réalisés au cours des dernières années dans le domaine de la détection de termes et de concepts lors de l'analyse de documents permettent d'aller plus loin en s'appuyant sur ces travaux et de prendre en charge la détection automatique de relations entre termes ou concepts au sein d'un même document, voire d'exploiter l'extraction de concepts pour inférer automatiquement des liens entre documents partageant des concepts proches. Ainsi, dans le cadre du projet PubMed Health puis du challenge i2b2 2010, je me suis intéressée à la détection automatique de relations spécifiques entre maladies et médicament dans des ressources institutionnelles [Névéol and Lu, 2010] puis de relations entre problèmes de santé, traitements et examens dans des dossiers électroniques patient [Islamaj Doğan et al., 2011]. Plus récemment, dans le cadre du PubMed Disease Sensor, j'ai contribué au développement d'un outil permettant d'extraire des termes de maladies rapportés aux concepts pertinents dans l'UMLS [Névéol et al., 2009b]. Appliqué à des ressources institutionnelles d'information sur les maladies, cet outil a permis d'établir automatiquement des liens fiables entre diverses ressources sur une même maladie [Névéol et al., 2012]. Dans l'optique de faciliter le partage des données et résultats issus de la recherche en biologie moléculaire, j'ai également développé un outil permettant l'extraction automatique de passages d'articles de la littérature rapportant le dépôt de données dans des bases données publiques [Névéol et al., 2011]. Ce travail permet d'évaluer la prévalence du dépôt de données biologiques et de contribuer à la création de métadonnées liant la littérature avec les bases conservant les données produites.

3. `https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ot`

4.1.2 Recherche d'information à partir d'un moteur de recherche

Recherche d'Information dans MEDLINE, CISMeF Une première piste de recherche explorée à la suite de ma thèse a porté sur la recherche d'information cross-lingue, afin de faciliter l'accès des francophones à des informations de santé en anglais, telles que les articles indexés dans MEDLINE. Cette première étude faisait l'hypothèse que les utilisateurs avaient une compétence en anglais permettant de comprendre des documents dans cette langue mais néanmoins pas suffisante pour formuler des requêtes d'information précises sur un sujet spécialisé comme la santé. Cette étude présente une méthode de recherche d'information translangue qui a pour but de permettre au grand public et en particulier aux patients d'accéder à une information de santé en anglais ou en français, quel que soit la langue utilisée pour formuler la requête [Névéol et al., 2006]. Deux terminologies reliées avec MeSH sont exploitées dans ce travail : les synonymes de la terminologie CISMeF, des termes français choisi par l'équipe CIMSeF pour enrichir le thésaurus MeSH, et les termes patients de MedlinePlus. Ces deux ressources ont été automatiquement reliées grâce à leur liens respectifs avec le thésaurus MeSH. Les 129 topics de MedlinePlus ont été reliés automatiquement à 142 synonymes patient de la terminologie CISMeF. Ces liens ont été ajoutés dans le portail terminologique et le moteur de recherche CISMeF et peuvent être exploités pour traduire les requêtes des patients. La méthode d'alignement de termes proposée peut également être exploitée pour d'autres paires de langue, dans la mesure où une terminologie patient est disponible dans les langues d'intérêt.

Une deuxième piste de recherche explorée lors de mon post-doctorat en collaboration avec les collègues de l'équipe CISMeF a également porté sur l'amélioration de l'accès à la littérature biomédicale. Cette étude s'appuyait sur la structuration du MeSH en concepts. Introduit en 2000 dans la terminologie, ce paradigme offre une granularité plus fine que les descripteurs qui constituaient jusqu'alors l'unité d'information primaire du MeSH. Malgré cela, la recherche d'information a continué de s'appuyer sur les descripteurs. Nous faisons l'hypothèse que les concepts MeSH ont un rôle fondamental dans la structuration du Thésaurus MeSH [Darmoni et al., 2000]. Par ailleurs, il a été montré que la méthode utilisée pour interpréter les requêtes des utilisateurs et les aligner avec le MeSH a un effet sur la spécificité et l'efficacité des résultats obtenus [Gault et al., 2002]. Ainsi, on peut s'attendre à ce que l'expérience des utilisateurs de MEDLINE bénéficie de méthodes d'indexation et de recherche d'information qui mette à profit l'intégralité de la structuration du MeSH. Nous avons donc proposé d'explorer l'impact de l'utilisation des concepts MeSH plutôt que des descripteurs pour la recherche d'information en lien avec les maladies rares et chroniques [Darmoni et al., 2012]. Pour cela, nous avons construit trois jeux de requêtes ciblant trente-deux maladies rares et vingt-deux maladies chroniques afin de mettre en œuvre trois stratégies de recherche : La première stratégie repose sur la recherche d'information PubMed standard, appelée *PubMed Automatic Term Mapping* (ATM). La deuxième stratégie repose sur l'algorithme de recherche d'information mis en œuvre dans le Catalogue et Index des Sites Médicaux Francophones (CISMeF), qui repose sur l'indexation à l'aide de concepts MeSH. La troisième permet d'extrapoler l'indexation MEDLINE pour simuler une indexation à l'aide de concepts MeSH.

Dans nos expériences, la troisième stratégie de recherche ramène un nombre significativement inférieur de documents par rapport aux deux autres (environ 18 000 notices versus 200 000 pour les maladies rares ; environ 300 000 notices versus 2 000 000 pour les maladies chroniques). La précision offerte par la stratégie de recherche CISMeF ATM est également plus élevée que la précision de la stratégie PubMed ATM pour les deux types de maladies étudiés.

Ces résultats suggèrent que l'utilisation de concepts MeSH (implémentée dans la stra-

tégie de recherche CISMeF et simulée dans la troisième stratégie de recherche proposée) présente des avantages pour la recherche d'information sur les maladies rares et chroniques. Cependant, notre étude s'est focalisée sur la précision des résultats et n'a pas mesuré le rappel. La précision a été estimée en faisant l'hypothèse que l'ensemble des articles pour lesquels un concept MeSH apparaissait dans le titre ou le résumé seraient indexés avec le concept correspondant. En pratique, ce ne serait pas forcément le cas, en particulier pour les articles où les termes n'apparaissent que dans le résumé. Cette expérience portant sur cinquante-quatre maladies rares et chroniques (Concepts MeSH) montre qu'une précision plus élevée peut être obtenue avec des requêtes basées sur les concepts MeSH plutôt que des descripteurs MeSH, qui demeure le comportement par défaut. Ceci illustre la conclusion de Lipscomb dans son aperçu historique du MeSH après l'introduction des concepts MeSH en 2000 : «le MeSH a un rôle important à jouer dans l'organisation de l'information d'une manière précise et puissante» [Lipscomb, 2000].

En pratique, la stratégie d'interrogation spécifique utilisée dans cette expérience (stratégie 3) pourrait être utilisée pour modifier la requête PubMed ATM pour des concepts pertinents (c.-à-d., des concepts MeSH non préférés plus étroits que le concept préféré du descripteur MeSH pertinent). Bien que cette stratégie offre l'avantage de ne nécessiter aucune modification de la politique d'indexation actuelle, l'utilisation d'une indexation des concepts combinée à certaines règles d'indexation appliquées aux concepts supplémentaires MeSH (les substances chimiques ne sont pas des descripteurs MeSH, et les maladies rares faisaient également partie des supplementary concepts jusqu'en 2012) serait une approche fondamentalement meilleure. Cette amélioration pourrait être facilement intégrée dans l'interface PubMed pour augmenter la précision lors de l'interrogation de la base de données bibliographiques MEDLINE, en particulier pour les maladies rares où il existe plusieurs concepts MeSH pour un descripteur MeSH. Pour ce faire, les auteurs suggèrent fortement de créer un Concept Supplémentaire MeSH pour chaque Concept MeSH subordonné qui n'est pas un concept préféré (n = 90 736) et de les utiliser pour l'indexation et la recherche d'informations, élargissant ainsi Liste de concepts supplémentaires introduite dans MeSH en 2011. Ce changement pourrait être transparent pour les utilisateurs. Une simple requête mappée automatiquement sur le concept MeSH pertinent produirait de meilleurs résultats sans nécessiter de connaissances avancées du MeSH, ce qui s'est avéré être un défi pour les non spécialistes en recherche d'information médicale [Delozier and Lingle, 1992].

Aide à la recherche d'études à inclure dans les revues systématiques Une *revue systématique*[4] répond à une question relative à l'efficacité d'une intervention en soins ou médicale, en ayant pris en compte toutes les études effectuées sur ce sujet au cours du temps. Une revue systématique est le fruit d'une démarche scientifique rigoureuse constituée de plusieurs étapes bien définies, incluant une recherche de littérature systématique, une évaluation de la qualité de chaque étude, une synthèse, quantifiée ou non, des résultats obtenus. Les différentes étapes de ce processus sont illustrées dans la figure 4.1

Les revues systématiques de la littérature dans le domaine biomédical reposent essentiellement sur le travail bibliographique manuel d'experts. Le travail de thèse de Christopher Norman explore la contribution du traitement automatique de la langue biomédicale afin de faciliter le travail des auteurs de revues systématiques, et en particulier les revues consacrées aux études de tests diagnostiques (*DTA studies*). Une première partie de la thèse a été consacrée à l'étude de méthodes de classification supervisée pour la découverte

4. Je reprends ici la définition proposée par le centre Cochrane, un organisme international qui réalise de nombreuses revues systématiques chaque année. `https://swiss.cochrane.org/fr/les-revues-syst%C3%A9matiques-systematic-reviews`

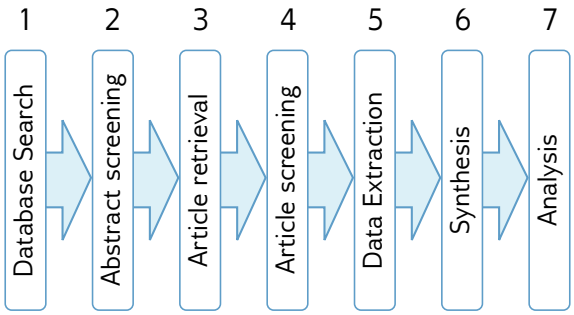


FIGURE 4.1 – Etapes de réalisation d'une revue systematique ; d'après [Tsafnat et al., 2014]

automatique d'articles (étapes 2 à 4). Ce processus, détaillé en figure 4.2, consiste à extraire les études susceptibles d'être inclues dans une revue systématique à partir des résultats d'une requête très large, soumise à des moteurs de recherche tels que Embase et Pubmed.

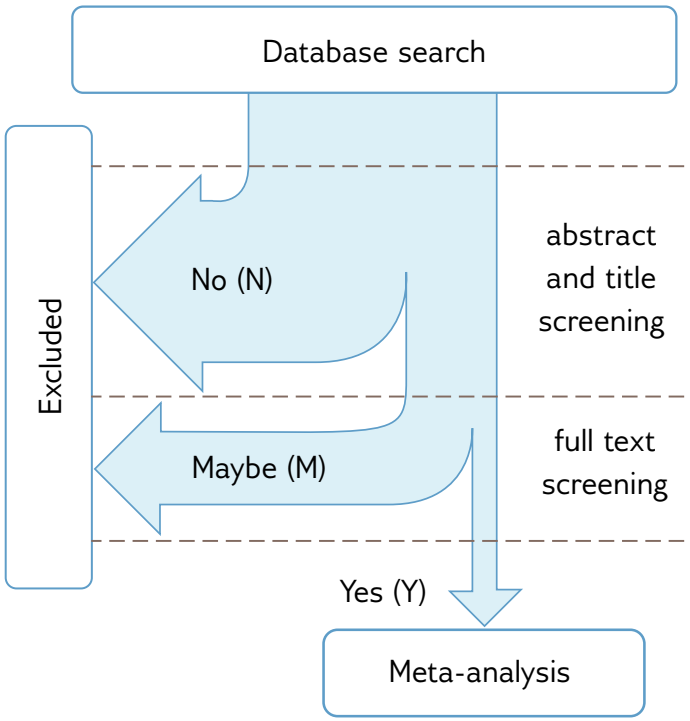


FIGURE 4.2 – Sélection d'articles à inclure dans une revue systématique.

Christopher a abordé ce problème sous l'angle de la classification en considérant que les articles candidats doivent être classés selon les catégories utilisées par les auteurs de revue systématique : l'article n'est pas pertinent pour la revue (classe N), l'article semble pertinent d'après son résumé mais n'est pas inclus après consultation du texte intégral (classe M), l'article est pertinent d'après son résumé et est inclus après consultation du texte intégral (classe Y). L'une des difficultés de cette tâche est le déséquilibre important entre les classes. La classe N est très majoritaire et la classe Y ne comporte parfois que très peu d'exemples. Un premier travail a consisté en une exploration de l'utilisation de la classe M afin d'optimiser les résultats de classification [Norman et al., 2018c]. Un modèle de régression logistique est appliqué sur deux corpus issus de revues systématiques conduites

dans le domaine du traitement automatique de la langue et de l'efficacité des médicaments. La classification offre une aire sous la courbe moyenne (AUC) de 0.769 si le classifieur est construit à partir des jugements experts portés sur les titres et résumés des articles, et de 0.835 si on utilise les jugements portés sur le texte intégral. Ces résultats indiquent l'importance des jugements portés dès le début du processus de sélection pour développer un classifieur efficace pour accélérer l'élaboration des revues systématiques à l'aide d'un algorithme de classification standard.

Ces résultats ont ensuite été appliqués dans le cadre de la participation de Christopher à la campagne d'évaluation CLEF eHealth 2017 [Norman et al., 2017] pour la tâche TAR " Technologically Assisted Reviews in Empirical Medicine" [Kanoulas et al., 2017]. L'approche proposée a été classée dans le tiers supérieur pour les 14 systèmes soumis et a obtenu les meilleurs résultats parmis les systèmes n'utilisant pas de méthode de recherche d'information et en particulier le principe de retour de pertinence. La complémentarité des deux approches (classification et retour de pertinence) nous a conduit à les combiner (en utilisant un classifieur neuronal) lors de la campagne d'évaluation CLEF eHealth 2018 [Norman et al., 2018b]. L'approche de classification offre de bonnes performances une fois qu'un nombre critique d'exemples positifs (classe Y) a été rencontré. Le retour de pertinence permet d'extraire ces exemples rapidement. Ainsi, le système résultant permet d'obtenir des résultats meilleurs que chacune des méthodes prises indépendamment.

Sur le plan pratique, la méthode de classification par régression logistique permet d'accélérer la sélection d'articles à inclure dans des revues de la littérature. Cette méthode a été mise en œuvre pour la réalisation de revues de la littérature pour la section "Clinical NLP" du Yearbook de l'IMIA (International Medical Informatics Association) en 2017 et 2018. Dans le cadre de collaborations avec plusieurs partenaires du projet Européen MIROR, une évaluation de la méthode est également en cours afin de l'intégrer dans le processus de mise à jour de certaines revues systématiques.

Une deuxième piste de recherche explorée dans le cadre de la thèse de Christopher porte sur la notion d'exhaustivité d'une recherche documentaire, et sur l'impact d'un éventuel défaut d'exhaustivité sur les résultats d'une méta-analyse. Pour effectuer ce travail, Christopher a constitué de manière semi-automatique un corpus issu des revues systématiques Cochrane sur les tests diagnostiques [Norman et al., 2018a] qui comporte outre les liens entre les revues systématiques et les études inclues ou exclues, l'ensemble des résultats des tests ayant fait l'objet d'une méta-analyse dans la revue. Ainsi, après un travail visant à reproduire le résultat des méta-analyses à l'aide de l'ensemble des études qui étaient à disposition des auteurs des revues systématiques, Christopher a mis en oeuvre une étude dite d'ablation dans laquelle il a reproduit le processus conduisant aux résultats de la méta-analyse à chaque étape du processus de sélection des articles. Ainsi, à l'aide de la méthode de classification des articles candidats décrite ci-dessus, les études ont été ajoutées à la méta-analyse au fur et à mesure de leur sélection. Nous avons ainsi pu étudier la différence entre les résultats des méta-analyses partielles à chaque sélection d'article et les résultats de la méta-analyse finale. Christopher a également proposé différents critères d'arrêt afin de déterminer automatiquement quand la revue de la littérature pouvait être arrêtée pour avoir une estimation proche du résultat de la méta-analyse complète. Globalement, les résultats de cette étude montrent que l'utilisation de méthodes automatiques (classification des études candidates à l'inclusion, utilisation d'un critère d'arrêt automatique de la revue de littérature) permet d'obtenir une approximation résultat des méta-analyses avec une marge d'erreur de 2% en examinant seulement 10% de l'ensemble des études candidates. Cependant, en pratique, ces caractéristiques ne s'appliquent qu'à un nombre limité de tests diagnostiques pour lesquels le nombre d'études primaires est suffisant pour construire une

analyse solide. La rédaction de cette étude est en cours en vue d'une publication.

4.1.3 Recherche d'information à partir du dossier patient

Les méthodes et les outils développés pour l'indexation MeSH peuvent servir de point de départ pour traiter des problèmes apparentés, tel que l'indexation d'articles de biologie moléculaire ou le codage automatique de documents cliniques aussi bien en anglais [Aronson et al., 2007] qu'en français [Pereira et al., 2006]. Ce dernier projet a donné lieu à l'implémentation d'un "info button", c'est à dire un lien contextualisé intégré dans les dossiers patients du système d'information du CHU de Rouen permettant un accès direct à la littérature pertinente pour le dossier patient indéxée dans le catalogue CISMeF [Darmoni et al., 2008].
Lors de mon stage de post-doctorat, j'ai également eu l'opportunité d'étudier le traitement de documents multimédia (texte et image radiographiques issus des dossiers patients) dans le cadre de l'indexation automatique ou de la recherche d'information translangue dans le dossier patient. Ces deux points ont donné lieu à des collaborations avec des collègues de l'université d'Aachen (T. Deserno, équipe IRMA) [Névéol et al., 2009a] et de l'Université de Buffalo (M. Ruiz) [Ruiz and Névéol, 2007].
Plus récemment, dans le cadre du projet ANR Accordys, j'ai contribué à la participation du LIMSI à la campagne TREC Clinical Decision Support Track qui avait pour but de proposer des méthodes permettant d'identifier de la littérature médicale pertinente pour un rapport de cas clinique. La description du cas clinique combiné un aspect clinique spécifique (diagnostic, examen, traitement) constituait une expression du besoin d'information. Nous avons alors proposé d'interroger la base PubMed avec une requête constituée des termes MeSH extraits des rapports de cas, combinés avec les termes MeSH des cinq principales hypothèses de maladie générées pour les rapports de cas et filtré par rapport à la dimension clinique souhaitée [D'hondt et al., 2014].

4.2 Analyse rétrospective des dossiers patients

A l'occasion de mon retour en France et de mon arrivée au LIMSI, je me suis intéressée à l'analyse de documents biomédicaux dans les langues autres que l'anglais. Mon travail s'est porté en particulier sur les documents cliniques présents dans les dossiers patients en Français, grâce à des collaborations avec des hôpitaux français, par exemple à l'APHP (Assistance Publique, Hôpitaux de Paris) et au CHU de Rouen.

La Figure 4.3 illustre le travail d'analyse des documents cliniques, qui est guidé par des applications concrètes définies par les besoins des professionnels de santé notamment en ce qui concerne la recherche d'information à partir du dossier patient et l'analyse rétrospective de cohortes.

Ce paradigme fait abstraction de la dimension de confidentialité des documents, que nous présentons dans la section 4.2.1. Les analyses présentées dans les sections 4.2.2 et 4.2.3 font l'hypothèse que des textes désidentifiés ou directement consultables sont disponibles.

4.2.1 Désidentification

L'un des obstacles majeurs au développement de telles méthodes de TAL est la difficulté d'accéder à des corpus de textes cliniques en raison de la confidentialité qui les entoure. Récemment, quelques corpus cliniques annotés ont été mis à disposition de la communauté scientifique dans le contexte des challenges i2b2 ([Uzuner, 2007, Sun et al., 2013] inter alia) et SHARE/CLEF [Pradhan et al., 2015], ce qui a permis des avancées méthodologiques

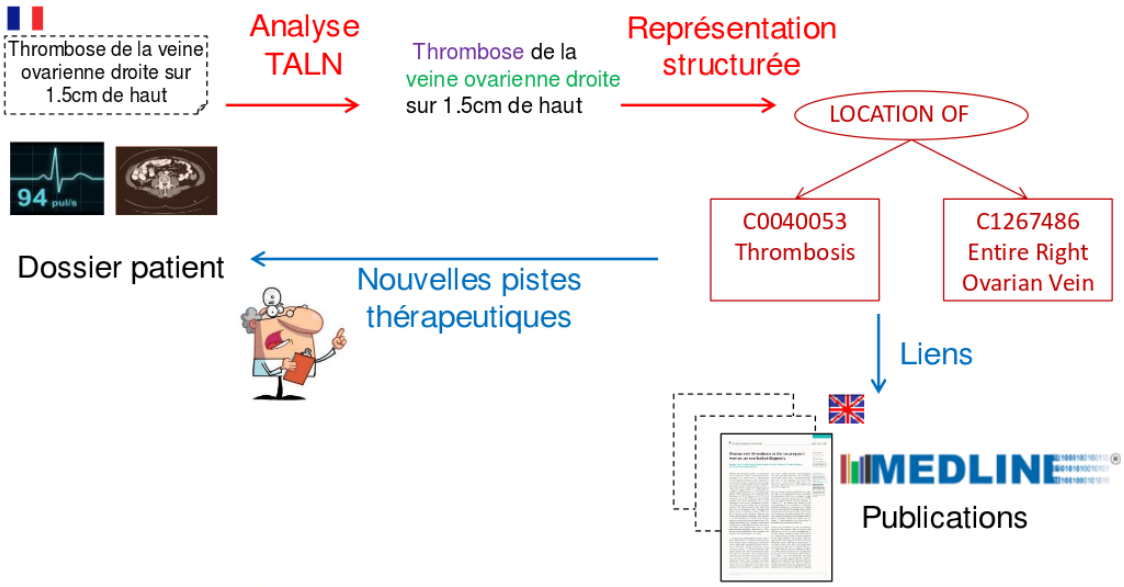


FIGURE 4.3 – Analyse automatique du contenu des dossiers électroniques patient et leurs applications.

considérables sur l'anglais [Chapman et al., 2011]. Cependant, aucun corpus équivalent n'est disponible dans une langue autre que l'anglais. Cette restriction perdure pour le français, malgré un effort de dialogue avec des partenaires hospitaliers et la CNIL. Ainsi une partie de mon travail, en collaboration avec Cyril Grouin a porté sur le développement et l'évaluation de méthodes de désidentification des textes cliniques en français.

Méthodologie de désidentification et de production de corpus désidentifiés. Dans le cadre du projet CABeRneT et du développement du corpus MERLoT, nous avons étudié différentes méthodes de désidentification de textes cliniques en français. Nous avons abordé la problématique de désidentification comme un problème de reconnaissance d'entités et focalisé notre analyse sur une dizaine de type d'entités dites identifiantes issues des critères américains HIPAA [5] (noms, prénoms, adresse, numéros de téléphones, numéros identifiants...). Ainsi, ce travail s'est appuyé sur l'outil à base de règles issu des travaux de thèse de Cyril (MEDINA [Grouin, 2013] [6]) que nous avons appliqué à un corpus de textes cliniques en français afin d'obtenir une référence pour l'évaluation de méthodes de désidentification en français [Grouin and Névéol, 2014]. Nous avons ainsi pu comparer les performances d'une méthode à base de règles et d'une méthode statistique utilisant les champs aléatoires conditionnels (CRF) grâce à l'outil Wapiti[Lavergne et al., 2010] [7]. L'observation globale à l'issue de ce travail a été que si une méthode à base de règle permet d'obtenir des performances intéressantes sur un nouveau corpus, la méthode statistique peut égaler et même dépasser ces performances dès lors qu'une petite quantité de données annotées du nouveau corpus sont disponibles.

Au cours de ce travail, nous avons également mené une réflexion sur la méthodologie

5. La législation "Health Insurance Portability and Accountability Act" entrée en vigueur aux Etats-Unis en 1996 `https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html`

6. `https://medina.limsi.fr/`

7. `https://wapiti.limsi.fr/`

d'annotation de corpus afin d'obtenir une ressource de qualité qui permette d'évaluer des méthodes, d'entraîner un modèle statistique sur un nouveau corpus adapté à une application pratique de désidentification, et de procéder efficacement à la désidentification de nouveaux textes cliniques. En collaboration avec Thomas Lavergne, nous avons exploré trois pistes possibles d'optimisation des efforts d'annotation en corpus, visant en particulier à réduire le temps d'annotation et l'intervention d'experts annotateurs. Nous partons du principe qu'un corpus de référence de qualité est établi en faisant intervenir au moins deux annotateurs sur l'ensemble du corpus qui travaillent d'abord indépendamment puis se concertent afin de résoudre leurs divergences pour arriver à un consensus. Le temps de travail est donc réparti entre : (1) la création de multiples annotations par les annotateurs, (2) la résolution des conflits pour arriver au consensus et (3) l'expertise des annotateurs qui ont pour tâche de créer toutes les annotations nécessaires, sans erreurs. Nos résultats indiquent que l'effort d'annotation peut être optimisé en limitant l'effort de double annotation à une petite partie du corpus. Une fois un accord inter-annotateur satisfaisant obtenu, les annotateurs peuvent se répartir la tâche sur le reste du corpus afin d'obtenir une large couverture. Un modèle statistique offre des performances sans différence significative s'il est entraîné sur un corpus ainsi constitué en annotation simple avec consensus [Grouin et al., 2014b]. Nous avons appliqué cette méthodologie dans le développement du corpus MERLoT.

La suite de ce travail, en collaboration avec des collègues médecins de l'HEGP et du CHU de Rouen, a conduit à la formalisation d'un protocole pour parvenir à la désidentification rapide d'un corpus clinique, présenté en figure 4.4 [Grouin et al., 2014a]. Ce protocole et les outils permettant de le mettre en œuvre, a été implémenté à l'HEGP et est maintenant utilisé en routine pour la désidentification de documents.

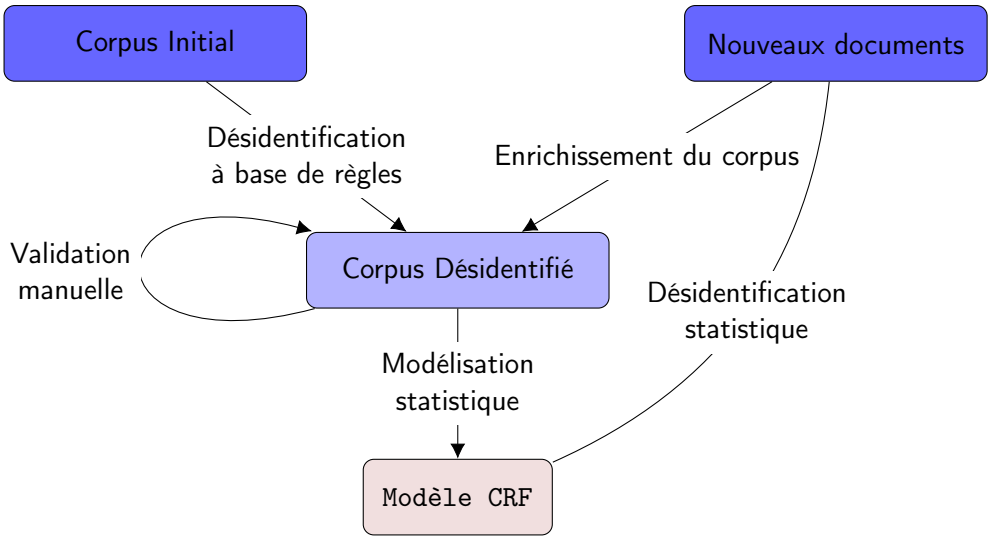


FIGURE 4.4 – Protocole de désidentification.

Etude des risques de réidentification. En parallèle, en collaboration avec Nicolas Griffon, nous nous sommes intéressés à l'évaluation pratique des risques de ré-identifications à partir de documents désidentifiés automatiquement dans lesquels les informations sensibles repérées par l'outil étaient ensuite automatiquement remplacées par des substituts plausibles [Grouin et al., 2015]. Cette étude a été réalisé sur un corpus pour lequel la

désidentification automatique offrait une performance état de l'art de 0,93 de F-mesure et n'avait pas été validée manuellement. Nous avons volontairement constitué un corpus de travail comportant des documents pour lesquels il est fortement probable que des informations personnelles n'aient pas été identifiées, et restent visibles dans les documents transformés. Ainsi, notre étude permet d'évaluer les risques de ré-identification dans un contexte où ce risque peut être considéré comme élevé. Nous avons extrait du corpus désidentifié 60 documents pour lesquels nous savons que des informations ont échappé à l'outil de désidentification. Cette extraction repose sur différents critères jugés difficiles à appréhender pour un outil de désidentification automatique, La figure 4.5 présente un exemple de texte original dans lequel les informations identifiantes sont repérées puis remplacées.

Texte original	Je vois en consultation Mr. Durand, né le 18/05/1954 à la Ciotat.
Marquage	Je vois en consultation Mr. NOM *Durand* , né le DATE *18/05/1954* à la Ciotat.
Texte pseudonymisé	Je vois en consultation Mr. Dupont, né le 26/11/1952 à la Ciotat.

FIGURE 4.5 – Désidentification et pseudonymisation automatique d'un texte clinique.

Dans cet exemple, deux informations identifiantes sont correctement repérées par le système (le nom et la date de naissance du patient, encadrés en couleur dans la partie "marquage") et une information n'est pas repérée (le lieu de naissance, souligné dans la partie "marquage"). Ainsi, le texte désidentifié comporte des informations identifiantes substituées lorsque le système les a correctement repérées, et des informations identifiantes originales lorsque le système ne les a pas extraites. Les documents ainsi désidentifiés de manière imparfaite sont présentés à six évaluateurs avec une connaissance variable des documents et de la méthode de désidentification employée, afin qu'ils repèrent les informations identifiantes originales et réidentifient les patients.

En dépit de l'absence de désidentification de certains éléments par notre outil de désidentification d'une part, et de la connaissance des faiblesses de l'outil par les développeurs d'autre part, jamais la protection de la vie privée des patients n'a été remise en cause sans disposer d'un accès privilégié au système d'information patient (SIP) de l'hôpital d'où sont issues les données, accès strictement réservé au personnel médical de l'établissement. Lorsqu'un accès au SIP est possible, les patients peuvent être réidentifiés par le biais d'un recoupement d'informations trouvées dans plusieurs documents et par la mobilisation de connaissances médicales sur le codage des actes médicaux. Le respect de la vie privée des patients semble néanmoins respecté lorsque n'est fourni qu'un seul document par patient, y compris pour le personnel médical disposant des accès au SIP.

Les résultats de cette étude montrent qu'une désidentification même imparfaite présente un risque de réidentification très faible. Cependant, l'étude montre également que ces risques sont non nuls, et rendent la méthodologie incompatible avec la définition de l'anonymisation introduite dans la législation européenne [8]. La conclusion de ce travail semble être l'anonymisation d'un texte clinique au sens de la réglementation n'est pas possible. En effet, la définition retenue pour l'anonymisation d'une base de données (dans ce contexte, un corpus clinique est vu comme une base de données textuelle) s'appuie sur trois critères à l'issue du traitement : i) est-il toujours possible d'isoler un individu ? ii)

8. CNIL, GROUPE DE TRAVAIL « ARTICLE 29 ». Avis 05/2014 sur les techniques d'anonymisation. `https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr.pdf`

est-il toujours possible de relier entre eux les enregistrements relatifs à un individu ? et iii) peut-on déduire des informations concernant un individu ? Notre expérience a mis en évidence que la possibilité de réidentification des patients s'appuie sur des informations qui ne sont pas masquées par la substitution d'informations considérées comme identifiantes : les réidentifications opérées dans notre expérience ont été fondées sur une analyse médicale des documents dans leur ensemble, ce qui consiste à *déduire des informations concernant un individu.*

4.2.2 Analyse temporelle

La chronologie des évènements est particulièrement importante dans le domaine médical. Avec le projet Digicosme COT (Coréférence événementielle cross-document dans les dossiers électroniques patient) nous nous sommes fixé comme objectif d'analyser des textes cliniques issus des dossiers électroniques patient du point de vue temporel et chronologique. À partir d'un ensemble de documents contenus dans le dossier d'un patient, le but de ce travail est de repérer les événements saillants de l'historique médical du patient ainsi que les marqueurs temporels associés afin de les agréger dans une chronologie synthétique (la figure 4.6 présente un exemple d'une telle chronologie), qui aura pour vocation d'être analysée par des cliniciens et comparée à la prise en charge de référence.

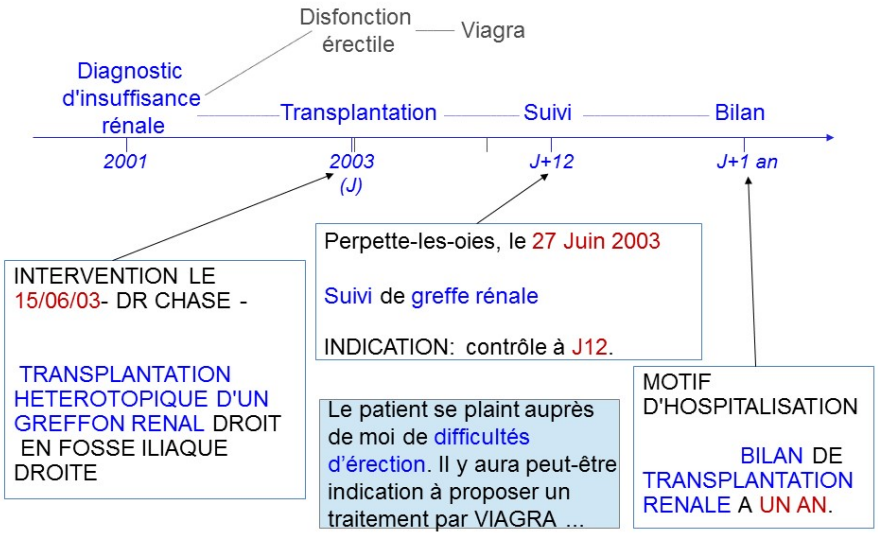
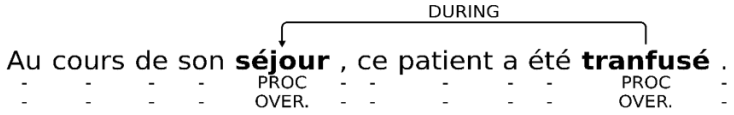


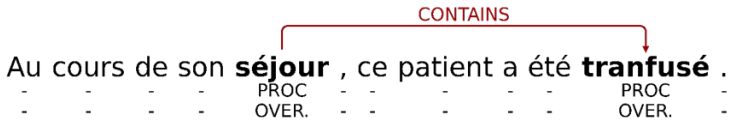
FIGURE 4.6 – Exemples de documents issus d'un dossier patient et agrégé en une chronologie des évènements saillants.

Dans ce cadre, la thèse de Julien Tourille a porté sur une analyse évènementielle des documents cliniques, et en particulier sur l'analyse des relations entre évènements : relations temporelles et coréférence. L'aspect méthodologique de l'extraction d'évènement, d'expressions temporelles et de relations a été exploré principalement sur des corpus en anglais, comme détaillé au chapitre 3. En pratique nous avons également étudié le potentiel d'adaptation au français de ces méthodes, grâce aux corpus THYME pour l'anglais et MERLoT pour le français.

Ce travail d'adaptation, décrit dans [Tourille et al., 2017c], a porté sur la détection des relations d'inclusion temporelle (X contient Y) intra-phrastiques entre les événements et/ou expressions temporelles (CR pour Container Relation) et des relations temporelles entre les événements et la date de création du document (DR pour Doctime Relation). Nous avons considéré pour cela que les événements et les expressions temporelles étaient déjà connus. Afin de partir de données équivalentes pour le français et l'anglais, nous avons pour cette étude restreint les relations du corpus THYME aux relations intra-phrastique (les relations inter phrases sont donc exclues), et effectué une conversion des annotations du corpus MERLoT afin d'expliciter les relations d'inclusion temporelles. La figure 4.7 illustre cette adaptation. Dans l'exemple, une relation TimeML "contains" est convertie vers une relation "contains". On peut noter que des relations du corpus MERLoT contenant un présupposé temporel comme "reveals" ont également été exploitées pour inférer des relations "contains".



(a) Annotations originales dans le corpus MERLoT



(b) Annotations modifiées pour intégrer la relation "conteneur"

FIGURE 4.7 – Exemples d'adaptation des relations temporelles du corpus MERLoT vers le schéma du corpus THYME

La figure 4.8 donne une vue synthétique des différents processus mis en œuvre pour traiter l'analyse temporelle de documents cliniques en français et en anglais.

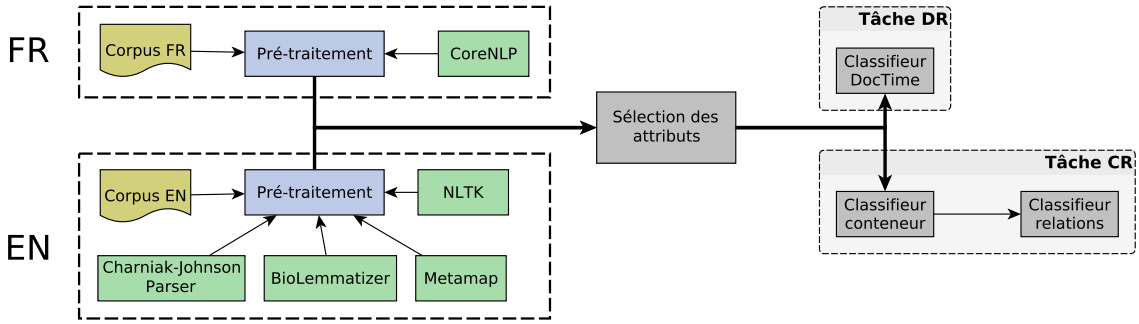


FIGURE 4.8 – Architecture du système d'extraction de relations temporelles.

Le pré-traitement des corpus repose sur les outils de traitement automatique de la langue disponibles dans les deux langues. Pour l'anglais, des outils adaptés au domaine biomédical sont disponibles. par contre, pour le français, nous avons dû utiliser des ressources du domaine général. La première étape consiste à segmenter, tokeniser et étiqueter en parties du discours. Pour l'anglais, nous avons utilisé NLTK pour la segmentation [Loper and

Bird, 2002] et le BLLIP Reranking Parser [Charniak and Johnson, 2005] en association avec un modèle pré-entrainé sur un corpus biomédical [McClosky, 2010] pour la tokenisation et l'analyse morpho-syntaxique. Pour le français, nous avons utilisé CoreNLP [Manning et al., 2014] et le modèle pré-entrainé sur le français pour le domaine général pour la segmentation, la tokenisation et l'analyse morpho-synatxique. Enfin, concernant l'identification des événements dans les textes, nous avons adopté des stratégies différentes selon les deux corpus considérés : dans le cas du corpus français, nous nous sommes appuyés sur une catégorisation déjà existante des entités pour identifier les événements tandis que pour le corpus anglais, nous avons exploité les résultats de l'outil Metamap [Aronson and Lang, 2010] pour déterminer les types des événements présents dans le corpus. Pour les différents classifieurs mis en place dans les tâches DR et CR, nous avons extrait des attributs structurels, contextuels et lexicaux. Les tailles optimales des fenêtres pour les contextes gauche et droit ont été calculées par validation croisée sur le corpus d'entraînement.

Les résultats principaux de ce travail sont présentés dans le tableau 4.1.

	MERLOT (fr)			THYME (en)		
	P	R	F1	P	R	F1
baseline	0.67	0.67	0.67	0.47	0.47	0.47
bef./over.	0.68	0.69	0.69	0.73	0.60	0.66
before	0.81	0.60	0.69	0.88	0.88	0.88
after	0.79	0.69	0.73	0.84	0.84	0.84
overlap	0.88	0.92	0.90	0.88	0.90	0.89
micro-average	0.83	0.84	0.83	0.87	0.87	0.87

(a) DR task results over the test corpus. We report precision (P), recall (R) and F1-Measure (F1) for all relation types.

	MERLOT (fr)			THYME (en)		
	P	R	F1	P	R	F1
baseline	0.43	0.15	0.22	0.55	0.06	0.11
no-relation	0.99	1.00	0.99	0.96	0.98	0.97
contains	0.75	0.57	0.65	0.61	0.47	0.53
micro-average	0.98	0.98	0.98	0.93	0.94	0.93

(b) CR task results over the test corpus. We report precision (P), recall (R) and F1-Measure (F1) for all relation types.

TABLE 4.1 – Résultats des expériences d'extraction de relations temporelles en deux langues.

Nous obtenons des résultats satisfaisants dans les deux tâches CR et DR. Pour la tâche DR, on observe un certain écart entre les F-mesures observées pour l'anglais (0,84) et le français (0,76). On note par ailleurs que les résultats par catégorie ne sont pas homogènes pour les deux langues. Pour le français, on observe un écart d'environ 0,30 entre les F1-mesures des catégories *Before-Overlap*, *Before* et *After* d'un côté et *Overlap* de l'autre. En ce qui concerne l'anglais, le modèle ne parvient pas à obtenir de bons résultats dans la catégorie *Before-Overlap* (0,59 de F1-mesure) mais parvient à des résultats homogènes autour de 0,80 pour les autres catégories. En ce qui concerne la tâche CR, les résultats obtenus sont proches avec une F1-mesure de 0,61 pour le corpus français sur la catégorie *contient* et une F1-mesure de 0,54 pour l'anglais. On observe que la précision pour les deux langues est identique. Le rappel en revanche est en dessous de la moyenne pour l'anglais. A titre de comparaison le meilleur score obtenu lors de l'édition 2016 de Clinical TempEval est 0,57 (F1-Mesure).

En ce qui concerne les différences de performance observées entre les deux langues, plusieurs hypothèses peuvent être formulées. Tout d'abord, l'utilisation de ressources spécialisées pour le traitement de la langue biomédicale en français permettrait d'obtenir de

meilleurs résultats quant au pré-traitement des textes et pourrait donc améliorer les résultats finaux. Ensuite, les corpus anglais et français sont déséquilibrés en termes de volume et d'entités annotées. Les performances sur le français peuvent ainsi être touchées par le faible nombre de textes annotés. Enfin la qualité des annotations, notamment celle des annotions en événements, qui est plus formelle et plus fine pour le corpus français que pour le corpus anglais, peut influencer les performances du système, notamment pour le modèle *Relation*.

En conclusion, cette étude semble indiquer que le traitement des relations temporelles dans des textes cliniques peut se généraliser au-delà de ces deux langues.

4.2.3 Extraction d'information précise et classification de documents

Prévalence des incidentalômes. Au sein de la direction informatique des hôpitaux, de nombreux cas d'usage et desiderata d'application du TAL sont récoltés. Ainsi, Anita Burgun de l'HEGP nous a fait part de la problématique de détection des incidentalômes. Il s'agit de cancers qui sont découverts fortuitement alors que le patient bénéficie d'un examen dans le cadre de l'exploration d'une autre pathologie, suspectée ou avérée. L'aspect fortuit de la découverte fait que le dossier du patient n'est pas codé formellement (à l'aide de la CIM10) par rapport à l'incidentalôme, mais par rapport à la raison initiale de l'examen réalisé. Cette situation rend difficile d'une part l'évaluation de la prévalence des incidentalômes et d'autre part l'identification systématique des patients concernés afin de leur proposer un suivi approprié. Le stage de M2 d'Anne-Dominique Pham, que j'ai co-encadré avec Anita Burgun et Thomas Lavergne, nous a permis d'évaluer l'apport de méthodes de TAL et de classification automatique pour apporter une réponse à ces problèmes. Ce travail a été publié dans un article de revue résumé ci-dessous [Pham et al., 2014].

Il a été démontré que le traitement automatique du langage naturel permet d'analyser le contenu des rapports de radiologie et d'identifier le diagnostic ou les caractéristiques du patient. Nous évaluons la combinaison du TAL et de l'apprentissage automatique pour détecter le diagnostic des maladies thromboemboliques et des incidentalômes à partir des rapports d'angiographie et de phlébographie rédigés en français. Nous modélisons le diagnostic thromboembolique et les découvertes fortuites sous la forme d'un ensemble de concepts, de modalités et de relations entre concepts pouvant être utilisés comme caractéristiques par un algorithme d'apprentissage automatique supervisé. Un corpus de 573 rapports de radiologie a été désidentifié et annoté manuellement avec l'aide des outils de TAL par un médecin pour les concepts, les modalités et les relations pertinents. Un classifieur a été entraîné sur les données et évalué par rapport au gold standard diagnostique réalisé par un médecin pour la thrombose veineuse profonde, l'embolie pulmonaire et les incidentalômes. Les modèles utilisés ont pris en compte le déséquilibre des classes et ont exploité la structure OMOP des rapports.

Le meilleur modèle offre une F-mesure de 0,98 pour l'identification de l'embolie pulmonaire, de 1,00 pour la thrombose veineuse profonde et de 0,80 pour les incidentalômes. L'utilisation de concepts, de modalités et de relations permet une amélioration des performances par rapport à modèle sac-de-mot pour l'ensemble des diagnostics.

Cette étude démontre les avantages de développer une méthode automatisée pour identifier les concepts médicaux, la modalité et les relations à partir des rapports de radiologie en français. Un système automatique d'annotation et de classification de bout en bout pouvant être appliqué à d'autres bases de données de rapports radiologiques serait utile pour la surveillance épidémiologique, le suivi des performances et l'accréditation dans les hôpitaux français.

Cartographie de la recherche clinique. La thèse d'Ignacio Atal [Atal, 2017] a porté sur l'élaboration d'une cartographie de la recherche clinique afin de la caractériser et d'identifier d'éventuels manques par rapport au fardeau observé des maladies [9]. Pour une partie de ce travail, résumée ci-dessous, des méthodes de traitement automatique de la langue semblaient pouvoir apporter une contribution à l'analyse des descriptions d'essais cliniques en anglais. A ce titre, j'ai pu apporter un éclairage sur les méthodes de TAL qui pouvaient répondre au besoin applicatif de la thèse et proposer une solution appropriée pour les mettre en œuvre.

Ce travail, décrit dans [Atal et al., 2016], repose sur l'hypothèse que les registres d'essais cliniques contiennent des informations susceptibles d'être utilisées pour dresser une cartographie globale de la recherche en santé. Cependant, les pathologies étudiées dans les essais ne sont pas indexées dans les registres à l'aide de taxonomies standardisées. Des travaux antérieurs ont étudié les registres d'essais cliniques dans l'optique d'améliorer l'appariement des patients avec des essais cliniques pertinents pour leur profil en vue d'une inclusion. Cependant, aucune étude n'a abordé la classification des essais cliniques par pathologie à l'aide d'une taxonomie standardisée qui permette une analyse globale alignant l'effort de recherche et le fardeau des maladies. Dans ce travail, nous proposons un classifieur à base de connaissances qui permette d'apparier les essais cliniques répertoriés dans les registres cliniques avec les catégories de pathologies définies par l'étude de l'Organisation Mondiale de la Santé sur le fardeau des maladies [Murray et al., 2012]. Le classificateur s'appuie sur l'UMLS et sur des algorithmes heuristiques pour l'analyse des données. Il propose des liens entre les essais cliniques et un groupe de 28 classes rassemblant des catégories du GBD en extrayant automatiquement les concepts UMLS de la description des essais en texte libre puis et effectuant une projection des concepts entre terminologies médicales. Le classifieur s'appuie sur le traitement automatique du langage naturel et sur la représentation des connaissances médicales pour créer des liens entre les registres d'essais cliniques et des catégories GBD candidates. Une sélection finale est effectuée en fonction de règles de priorisation définies par des experts. Nous avons comparé les classifications automatiques et manuelles pour un ensemble de test externe de 2 763 essais. Nous avons automatiquement classé 109 603 essais interventionnels enregistrés avant février 2014 dans le registre ICTRP de l'OMS. Lors de la validation externe, le classificateur a identifié les catégories exactes de GBD pour 78% des essais. Il offre de très bonnes performances pour la plupart des 28 catégories, en particulier les "néoplasmes" (sensibilité 97,4 %, spécificité 97,5%). La sensibilité était modérée pour les essais ne relevant d'aucune catégorie de GBD (53%) et faible pour les essais de blessures (16%). Pour les 109 603 essais enregistrés sur le registre ICTRP de l'OMS, le classificateur n'a attribué aucune catégorie de GBD à 20,5% des essais. Les catégories GBD les plus fréquemment extraites étaient "néoplasmes" (22,8%) et "diabète" (8,9%). Nous avons développé et validé un classificateur à base de connaissances permettant d'identifier automatiquement les maladies étudiées dans des essais enregistrés en utilisant la taxonomie de l'étude GBD 2010. Cet outil est disponible librement pour la communauté des chercheurs et peut être utilisé pour des études de santé publique à grande échelle.

9. Le *fardeau d'une maladie* représente l'impact total de la maladie sur un individu ou sur une société, en termes de coût financier, de morbidité et de mortalité. L'Organisation Mondiale de la Santé propose d'utiliser les années de vie corrigées de l'incapacité (Disability-Adjusted Life Year, DALY) comme mesure du fardeau `https://www.who.int/topics/global_burden_of_disease/fr/`.

4.2.4 Discussion

La désidentification peut presque être considérée comme un problème résolu pour ce qui est de l'utilisation de documents cliniques pour des études internes à l'hôpital, dans la mesure où les méthodes actuelles sont en mesure d'offrir des performances très élevées pour l'extraction d'entités identifiantes telles que les noms, adresses, et numéros identifiants divers. Le partage de documents désidentifiés dans un autre cadre n'est cependant pas permis par la réglementation européenne actuelle.

En pratique, nous manquons toujours d'un outil générique d'analyse des textes cliniques dans les langues autres que l'anglais comme MetaMap. Grace au travail accompli ces dernières années, un bon nombre d'éléments qui devraient constituer des briques de base de cet outil ont été étudiés, et nous disposons maintenant de ressources pour développer et évaluer des outils : découpage en phrases, découpage en sections, analyse morphosyntaxique, extraction d'entités, liaison référentielle (notamment vers l'UMLS), extraction de relations. Il faut donc progresser vers la mise en place et le partage d'un outil d'analyse générique. En parallèle, de nombreuses autres pistes restent à étudier ; par exemple la reconnaissance d'acronymes ou la désambiguïsation.

On peut également retenir qu'une analyse complexe n'est pas toujours nécessaire pour certaines applications : ainsi, l'exemple des maladies thromboemboliques montre qu'une approche de classification sac de mot donne déjà des résultats satisfaisants, même si l'extraction d'entités et de relations apporte une contribution supplémentaire. En revanche, pour la détection d'incidentalômes, l'apport de l'identification du contexte est significatif.

Chapitre 5

Conclusion et Perspectives

5.1 Conclusion

Ce manuscrit a présenté le travail que j'ai réalisé dans le domaine du traitement automatique de la langue biomédicale. Mon travail s'est attaché à modéliser l'information de santé avec plusieurs niveaux de granularité : au niveau du texte (codage, indexation, classification de documents), au niveau de l'énoncé (extraction d'entités, de relations et de leur contexte) ainsi qu'à des niveaux intermédiaires (découpage de textes en sections, extractions de relations temporelles au-delà de l'énoncé, analyse en coréférence dans le cadre de la thèse de Julien Tourille).

Mes travaux ont porté principalement sur l'anglais et le français, avec un soucis d'ouverture sur la prise en compte de multiples langues, en particulier les langues autres que l'anglais. Cette thématique a été abordée sous l'angle de la traduction automatique (production de ressources, organisation de campagnes WMT), de l'adaptation (analyse temporelle, organisation de campagnes CLEF eHealth) et de la revue de la littérature (notamment, [Névéol et al., 2018]).

Importance de la reproductibilité. Il est essentiel de disposer de ressources de qualité, illustrant la modélisation du problème considéré, et permettant des évaluations reproductibles. Ainsi, lors de la constitution du schéma d'annotation de CABeRneT et du corpus MERLoT, nous avons vérifié la cohérence des annotations à intervalle régulier pour constituer un guide d'annotation de qualité. Nous avons également validé la pertinence médicale des annotations par l'intervention d'un médecin sur un échantillon du corpus. La campagne d'annotation s'est conclue par une phase d'harmonisation des annotations afin de corriger les incohérences résiduelles. La thèse de Christopher Norman a donné lieu à plusieurs études de reproductibilité. Tout d'abord, la participation de Christopher à la campagne CLEF eHealth TAR 2018 s'est appuyée sur une méthode de relevance feedback qui offrait de bonnes performances sur l'édition 2017. La reproduction de ces résultats a permis d'élucider les détails de l'implémentation de la méthode avant de la combiner avec une méthode de classification permettant de faire avancer l'état de l'art. De même, la constitution du corpus d'informations extraites d'articles inclus dans les revues systématiques Cochrane a été validée par la reproduction de certaines analyses publiées dans les revues systématiques.

5.2 Perspectives

Cette section reprend des points importants abordés dans mes travaux, qui restent des pistes de recherche actives.

5.2.1 Partage de ressources pour le français biomédical

Une perspective de recherche importante porte sur la disponibilité de ressources pour le traitement de la langue biomédicale dans les langues autres que l'anglais. Le partage de corpus cliniques annotés constitue notamment un enjeu fondamental. En effet, la disponibilité des ces ressources est nécessaire pour permettre à la communauté de travailler ensemble sur des problèmes avec une définition partagée, d'explorer différentes représentations d'un même phénomène linguistique d'intérêt, et surtout d'effectuer des comparaisons directes et reproductibles entre méthodes. Les travaux sur l'anglais par exemple pour la détection de la négation ont montré la richesse que constituent de telles ressources [Wu et al., 2014]. La réglementation en matière de données personnelles de santé et de confidentialité en Europe est en cours d'évolution et ne permet pas à l'heure actuelle un partage de données comparable à ce qui existe sur les données en anglais. Cependant, la création de documents cliniques synthétiques ou l'utilisation de substituts comme les cas cliniques anonymes constituent autant de directions intéressantes à poursuivre. Il conviendra de montrer dans quelle mesure ce type de corpus peut être considéré comme comparable à un corpus clinique original, et quels sont les points de divergence qui devront faire l'objet d'éventuels paramétrages ou adaptations pour les modèles.

5.2.2 Vers des modèles génériques ou adaptables ?

La variabilité entre corpus issus de différentes spécialités médicales, différents hôpitaux ou différents genres fait qu'il est difficilement envisageable de disposer de corpus annotés pour l'ensemble des tâches et des types de texte concernés. Ainsi, une perspective de recherche porte sur la compréhension des différences entre ces textes, afin de s'orienter soit vers des modèles génériques exploitables sur différents corpus (par exemple, des modèles de transfert) soit vers des modèles adaptables qui soient capables, à partir d'un corpus source d'offrir une analyse performante sur un corpus cible dont les différences avec le corpus source seraient caractérisées. Ainsi, il serait intéressant de définir quelle est la meilleure stratégie en fonction des applications considérées.

Dans le cadre de la thèse de Julien Tourille, nous nous sommes intéressés à l'adaptation en domaine lors de notre participation à la campagne SemEval 2017 qui avait pour but l'adaptation de l'analyse temporelle du domaine du cancer de côlon vers le cancer du cerveau. Nous avons ensuite montré qu'une modélisation des informations réalisée dans des corpus comparables en anglais et en français permettait un transfert méthodologique entre les langues conservant les performances pour l'analyse temporelle. Les collaborations du LIMSI avec l'APHP ont également permis d'étudier le transfert de modèles pour la tâche de désidentification de textes cliniques. Cependant, l'applicabilité pratique de ces travaux se heurte à la problématique de la confidentialité des modèles statistiques, et en particulier neuronaux. On peut également replacer ce problème dans le contexte du partage des ressources évoqué ci-dessus.

5.2.3 Intégration de l'expertise biomédicale

Un défi du TAL biomédical reste l'interaction avec les utilisateurs des méthodes d'analyse de textes développées : praticiens hospitaliers, auteurs de revues systématiques, in-

dexeurs ou codeurs... D'une part, la recherche en TAL doit continuer de s'intéresser à des questions fondamentales de compréhension et de modélisation des phénomènes linguistiques sans devenir de l'ingénierie médicale. D'autre part, les problématiques de recherche doivent néanmoins être guidées par les applications qui intéressent les utilisateurs.

En termes de modélisation des problèmes à résoudre, il est crucial de communiquer aux utilisateurs l'importance de disposer de données de travail de quantité et de nature appropriée aux méthodes que l'on souhaite mettre en œuvre. Les données sont un point d'accès à l'expertise de ces utilisateurs, qui peuvent déjà disposer de ressources directement utilisables. Par exemple, dans le cadre de la campagne CLEF eHealth, l'historique du codage des certificats de décès électroniques en anglais, français, hongrois et italien a pu donner lieu à l'évaluation de nombreuses méthodes, dont certaines sont en cours d'intégration dans le workflow du CépiDC.

En termes d'évaluation, il est important de proposer des métriques ou des évaluations orientées vers l'application. Par exemple, pour la désidentification, les métriques d'extraction d'information que sont la précision, le rappel et la F-mesure ne permettent qu'imparfaitement de rendre compte des performances des systèmes du point de vue de l'estimation des risques de ré-identification. Je me suis intéressée à ce problème en collaboration avec Cyril Grouin et Nicolas Griffon, dans une étude qui a montré qu'un système de désidentification entièrement automatique offrant un rappel de 0,90 permettait de masquer les informations identifiantes des patients afin de limiter les possibilités de réidentification à des utilisateurs bénéficiant d'un accès privilégié au système d'information hospitalier de l'établissement d'origine des documents.

Cette dernière section a dressé les perspectives de ma recherche pour les prochaines années. Il s'agira notamment de continuer à travailler en collaboration proche avec les collègues hospitalo-universitaires afin de progresser sur l'intégration des méthodes de TAL et des applications cliniques.

Bibliographie

[Abeillé et al., 2003] Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In Abeillé, A., editor, *Treebanks*. Kluwer, Dordrecht.

[Altman, 2011] Altman, R. (2011). Pharmacogenomics :"noninferiority" is sufficient for initial implementation. *Clinical Pharmacology & Therapeutics*, 89(3) :348–350.

[Aronson and Lang, 2010] Aronson, A. and Lang, F. (2010). An overview of MetaMap : historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3) :229–36.

[Aronson et al., 2007] Aronson, A. R., Bodenreider, O., Demner-Fushman, D., Fung, K. W., Lee, V. K., Mork, J. G., Névéol, A., Peters, L., and Rogers, W. J. (2007). From indexing the biomedical literature to coding clinical text : experience with mti and machine learning approaches. In *Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, pages 105–112. Association for Computational Linguistics.

[Aronson et al., 2004] Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., Rogers, W. J., et al. (2004). The nlm indexing initiative's medical text indexer. *Medinfo*, 89.

[Atal, 2017] Atal, I. (2017). *Cartographie globale des essais cliniques*. PhD thesis. Thèse de doctorat dirigée par Porcher, Raphaël Santé publique Sorbonne Paris Cité 2017.

[Atal et al., 2016] Atal, I., Zeitoun, J.-D., Névéol, A., Ravaud, P., Porcher, R., and Trinquart, L. (2016). Automatic classification of registered clinical trials towards the global burden of diseases taxonomy of diseases and injuries. *BMC Bioinformatics*, 17(1) :392.

[Aurnague et al., 2000] Aurnague, M., Boulanouar, K., Nespoulous, J.-L., Borillo, A., and Borillo, M. (2000). Spatial semantics : the processing of Internal Localization Nouns. *Cahiers de Psychologie Cognitive - Current Psychology of Cognition*, 19(1) :69–110.

[Biber, 1988] Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, New York, NY.

[Boyer and Névéol, 2018] Boyer, A. and Névéol, A. (2018). Détection automatique de phrases en domaine de spécialité en français. In *Conférence sur le Traitement Automatique des Langues Naturelles*, pages 205–2013.

[Campillos et al., 2018] Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A.-L., and Névéol, A. (2018). A french clinical corpus with comprehensive semantic annotations : development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2) :571–601.

[Candito and Seddah, 2012] Candito, M.-H. and Seddah, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN 2012*, pages 321–334.

[Chapman et al., 2001] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5) :301–310.

[Chapman and Cohen, 2009] Chapman, W. W. and Cohen, K. B. (2009). Guest editorial : Current issues in biomedical text mining and natural language processing. *Journal of biomedical informatics*, 42(5) :757–759.

[Chapman et al., 2011] Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text : the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5) :540–543.

[Charniak and Johnson, 2005] Charniak, E. and Johnson, M. (2005). Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan. Association for Computational Linguistics.

[Charnock, 1999] Charnock, R. (1999). Les langues de spécialité et le langage technique : considérations didactiques. *ASp*, 23-26 :281–302.

[Cohen et al., 2017] Cohen, K., Névéol, A., Xia, J., Hailu, N., Hunter, L., and Zweigenbaum, P. (2017). Reproducibility in biomedical natural language processing. In *AMIA annual symposium proceedings*. American Medical Informatics Association.

[Cohen et al., 2018] Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., Ide, N., Névéol, A., Grouin, C., and Hunter, L. E. (2018). Three dimensions of reproducibility in natural language processing. In *Proceeding of LREC 2018. International Conference on Language Resources & Evaluation*, volume 2018, pages 156–165. NIH Public Access.

[Cohen et al., 2013] Cohen, R., Elhadad, M., and Elhadad, N. (2013). Redundancy in electronic health record corpora : analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, 14 :10.

[Collier et al., 2006] Collier, N., Nazarenko, A., Baud, R., and Ruch, P. (2006). Recent advances in natural language processing for biomedical applications. *International Journal of Medical Informatics*, 75(6) :413 – 417. Recent Advances in Natural Language Processing for Biomedical Applications Special Issue.

[Darmoni et al., 2008] Darmoni, S., Pereira, S., Névéol, A., Massari, P., Dahamna, B., Letord, C., Kerdelhué, G., Piot, J., Derville, A., and Thirion, B. (2008). French infobutton : an academic and business perspective. In *AMIA... Annual Symposium proceedings. AMIA Symposium*, pages 920–920.

[Darmoni et al., 2000] Darmoni, S. J., Leroy, J.-P., Baudic, F., Douyère, M., Piot, J., and Thirion, B. (2000). Cismef : a structured health resource guide. *Methods of information in medicine*, 39(01) :30–35.

[Darmoni et al., 2012] Darmoni, S. J., Soualmia, L. F., Letord, C., Jaulent, M.-C., Griffon, N., Thirion, B., and Névéol, A. (2012). Improving information retrieval using mesh concepts : a test case on rare and chronic diseases. *J Med Libr Assoc*, (3) :176–83.

[Deléger et al., 2017] Deléger, L., Campillos, L., Ligozat, A.-L., and Névéol, A. (2017). Design of an extensive information representation scheme for clinical narratives. *Journal of Biomedical Semantics*, 8(1) :37.

[Deléger and Névéol, 2014] Deléger, L. and Névéol, A. (2014). Automatic identification of document sections for designing a french clinical corpus (identification automatique de zones dans des documents pour la constitution d'un corpus médical en français)[in french]. *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, 2 :568–573.

[Delozier and Lingle, 1992] Delozier, E. P. and Lingle, V. A. (1992). Medline and mesh : challenges for end users. *Medical reference services quarterly*, 11(3) :29–46.

[Deléger et al., 2014] Deléger, L., Grouin, C., Ligozat, A.-L., Zweigenbaum, P., and Névéol, A. (2014). Annotation of specialized corpora using a comprehensive entity and relation scheme. In *Proc of LREC*, pages 1267–1274.

[Demner-Fushman et al., 2010] Demner-Fushman, D., Apostolova, E., Islamaj Dogan, R., Lang, F.-M., Mork, J. G., Névéol, A., Shooshan, S. E., Simpson, M., and Aronson, A. R. (2010). Nlm's system description for the fourth i2b2/va challenge. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA : i2b2*.

[Demner-Fushman et al., 2009] Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support ? *J Biomed Inform*, 42 :760–772.

[Demner-Fushman et al., 2017] Demner-Fushman, D., Rogers, W. J., and Aronson, A. R. (2017). Metamap lite : an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4) :841–844.

[Dernoncourt et al., 2017] Dernoncourt, F., Lee, J. Y., and Szolovits, P. (2017). Neuro-NER : An Easy-to-Use Program for Named-Entity Recognition Based on Neural Networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, volume System Demonstrations, pages 97–102. Association for Computational Linguistics.

[D'hondt et al., 2014] D'hondt, E., Grau, B., Darmoni, S., Névéol, A., Schuers, M., and Zweigenbaum, P. (2014). Limsi@ 2014 clinical decision support track. Technical report, NATIONAL CENTER FOR SCIENTIFIC RESEARCH ORSAY (FRANCE) COMPUTER SCIENCES LAB FOR MECHANICS AND ENGINEERING SCIENCES.

[D'hondt et al., 2016] D'hondt, E., Grouin, C., Névéol, A., Stamatatos, E., and Zweigenbaum, P. (2016). Detection of text reuse in french medical corpora. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 108–114.

[D'hondt et al., 2015] D'hondt, E., Tannier, X., and Névéol, A. (2015). Redundancy in french electronic health records : A preliminary study. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 21–30.

[Doğan et al., 2014] Doğan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus : A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47 :1 – 10.

[Elhadad et al., 2015] Elhadad, N., Pradhan, S., Gorman, S., Manandhar, S., Chapman, W., and Savova, G. (2015). Semeval-2015 task 14 : Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

[Escudié et al., 2017] Escudié, J., Rance, B., Malamut, G., Khater, S., Burgun, A., Cellier, C., and Jannot, A. (2017). A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease : a case study on autoimmune comorbidities in patients with celiac disease. *BMC Med. Inf. & Decision Making*, 17(1) :140 :1–140 :10.

[Ferraro et al., 2013] Ferraro, J., Daumé, H., DuVall, S., Chapman, W., Harkema, H., and Haug, P. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *J Am Med Inform Assoc*, 20(5) :931–939.

[Friedman et al., 1995] Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S. B., and Clayton, P. D. (1995). Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1) :83–108.

[Friedman et al., 2002] Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages : a description based on the theories of zellig harris. *Journal of Biomedical Informatics*, 35(4) :222 – 235. Sublanguage - Zellig Harris Memorial.

[Funk and Reid, 1983] Funk, M. E. and Reid, C. A. (1983). Indexing consistency in medline. *Bulletin of the Medical Library Association*, 71(2) :176.

[Garcelon et al., 2017] Garcelon, N., Neuraz, A., Benoit, V., Salomon, R., and Burgun, A. (2017). Improving a full-text search engine : the importance of negation detection and family history context to identify cases in a biomedical data warehouse. *Journal of the American Medical Informatics Association*, 24(3) :607–613.

[Gault et al., 2002] Gault, L. V., Shultz, M., and Davies, K. J. (2002). Variations in medical subject headings (mesh) mapping : from the natural language of patron terms to the controlled vocabulary of mapped lists. *Journal of the Medical Library Association*, 90(2) :173.

[Gonzalez-Hernandez et al., 2017] Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., and Savova, G. (2017). Capturing the patient's perspective : a review of advances in natural language processing of health-related text. *Yearb Med Inform*, 26(1) :214–227.

[Grouin, 2013] Grouin, C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. PhD thesis, Université Pierre et Marie Curie, Paris, France.

[Grouin et al., 2014a] Grouin, C., Deléger, L., Escudié, J.-B., Groisy, G., Jannot, A.-S., Rance, B., Tannier, X., and Névéol, A. (2014a). How to de-identify a large clinical corpus in 10 days. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

[Grouin et al., 2015] Grouin, C., Griffon, N., and Névéol, A. (2015). Is it possible to recover personal health information from an automatically de-identified corpus of french ehrs ? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2015, Lisbon, Portugal, September 17, 2015*, pages 31–39.

[Grouin et al., 2014b] Grouin, C., Lavergne, T., and Neveol, A. (2014b). Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 54–58. Association for Computational Linguistics and Dublin City University.

[Grouin and Névéol, 2014] Grouin, C. and Névéol, A. (2014). De-identification of clinical notes in French : towards a protocol for reference corpus developpement. *J Biomed Inform*, 46(3) :506–515.

[Hahn et al., 2007] Hahn, U., Wermter, J., Blasczyk, R., and Horn, P. A. (2007). Text mining : powering the database revolution. *Nature*, 448(7150) :130.

[Harris, 1991] Harris, Z. (1991). *A Theory of Language and Information : A Mathematical Approach*. Clarendon Press, Oxford.

[Herskovic et al., 2007] Herskovic, J. R., Tanaka, L. Y., Hersh, W., and Bernstam, E. V. (2007). A day in the life of pubmed : analysis of a typical day's query log. *Journal of the American Medical Informatics Association*, 14(2) :212–220.

[Hébert, 2010] Hébert, L. (2010). Typologie des structures du signe : le signe selon le groupe µ. *Actes Sémiotiques*, 113.

[Intxaurrondo et al., 2017] Intxaurrondo, A., Pérez-Pérez, M., Pérez-Rodríguez, G., López-Martín, J. A., Santamaria, J., de la Pena, S., Villegas, M., Akhondi, S. A., Valencia, A., Lourenço, A., et al. (2017). The biomedical abbreviation recognition and resolution (barr) track : benchmarking, evaluation and importance of abbreviation recognition systems applied to spanish biomedical abstracts.

[Islamaj Dogan et al., 2009] Islamaj Dogan, R., Murray, G. C., Névéol, A., and Lu, Z. (2009). Understanding pubmed® user search behavior through log analysis. *Database*, 2009 :bap018.

[Islamaj Doğan et al., 2011] Islamaj Doğan, R., Névéol, A., and Lu, Z. (2011). A context-blocks model for identifying clinical relationships in patient records. *BMC bioinformatics*, 12(3) :S3.

[Jimeno Yepes and Névéol, 2013] Jimeno Yepes, A. and Névéol, A. (2013). Effect of additional in-domain parallel corpora in biomedical statistical machine translation. In *Proceedings of the 4th International Workshop on Health Document Text Mining and Information Analysis with the Focus of Cross-Language Evaluation (Louhi 2013)*.

[Jimeno Yepes et al., 2013] Jimeno Yepes, A., Prieur-Gaston, E., and Névéol, A. (2013). Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14 :146.

[Jonquet et al., 2009] Jonquet, C., Shah, N. H., and Musen, M. A. (2009). The open biomedical annotator. *Summit on translational bioinformatics*, 2009 :56.

[Kanoulas et al., 2017] Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2017). Overview of the CLEF technologically assisted reviews in empirical medicine. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*, CEUR Workshop Proceedings.

[Kiss and Strunk, 2006] Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4) :485–525.

[Kors et al., 2015] Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M., and Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition : the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5) :948–956.

[Lavergne et al., 2010] Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

[Lavergne et al., 2016] Lavergne, T., Névéol, A., Robert, A., Grouin, C., Rey, G., and Zweigenbaum, P. (2016). A dataset for icd-10 coding of death certificates : Creation and usage. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 60–69, Osaka, Japan. The COLING 2016 Organizing Committee.

[Legrand et al., 2017] Legrand, J., Toussaint, Y., Raïssi, C., and Coulet, A. (2017). Tree-LSTM and Cross-Corpus Training for Extracting Biomedical Relationships from Text. In *DLPM2017 Workshop - 2nd International Workshop on Deep Learning for Precision Medicine*, Skopje, Macedonia. Held in conjunction with ECML-PKDD 2017.

[Ligozat, 2010] Ligozat, G. (2010). *Raisonnement qualitatif sur le temps et l'espace*. Hermès Lavoisier, Paris.

[Lindberg et al., 1993] Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The unified medical language system. *Methods of information in medicine*, 32(04) :281–291.

[Lippincott et al., 2011] Lippincott, T., Séaghdha, D. Ó., and Korhonen, A. (2011). Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, 12(1) :212.

[Lipscomb, 2000] Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3) :265.

[Liu et al., 2007] Liu, K., Chapman, W., Hwa, R., and Crowley, R. (2007). Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *J Am Med Inform Assoc*, 14(5) :641–650.

[Lok, 2010] Lok, C. (2010). Literature mining : Speed reading. *Nature News*, 463(7280) :416–418.

[Lomax and McCray, 2004] Lomax, J. and McCray, A. T. (2004). Mapping the gene ontology into the unified medical language system. *International Journal of Genomics*, 5(4) :354–361.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). NLTK : The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania. Association for Computational Linguistics.

[Manchikanti, 2008] Manchikanti, L. (2008). Evidence-based medicine, systematic reviews, and guidelines in interventional pain management, part i : introduction and general considerations. *Pain Physician*, 11(2) :161–86.

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

[Max et al., 2013] Max, A., Névéol, A., Yvon, F., Zweigenbaum, P., and Ravaud, P. (2013). Traduction automatique en français des revues cochrane. In *Actes du 1er SIG IMIA Francophone*. International Medical Informatics Association (IMIA).

[McClosky, 2010] McClosky, D. (2010). *Any Domain Parsing : Automatic Domain Adaptation for Natural Language Parsing*. PhD thesis, Department of Computer Science, Brown University.

[McCray et al., 2001] McCray, A. T., Burgun, A., and Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. In *Proc of MedInfo*, volume 10, pages 216–20.

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Mork et al., 2017] Mork, J., Aronson, A., and Demner-Fushman, D. (2017). 12 years on – is the nlm medical text indexer still useful and relevant ? *Journal of Biomedical Semantics*, 8(1) :8.

[Mork et al., 2010] Mork, J. G., Bodenreider, O., Demner-Fushman, D., Doğan, R. I., Lang, F.-M., Lu, Z., Névéol, A., Peters, L., Shooshan, S. E., and Aronson, A. R. (2010). Extracting rx information from clinical narrative. *Journal of the American Medical Informatics Association*, 17(5) :536–539.

[Murray et al., 2012] Murray, C. J., Ezzati, M., Flaxman, A. D., Lim, S., Lozano, R., Michaud, C., Naghavi, M., Salomon, J. A., Shibuya, K., Vos, T., Wikler, D., and Lopez, A. D. (2012). Gbd 2010 : design, definitions, and metrics. *The Lancet*, 380(9859) :2063 – 2066.

[Névéol et al., 2016a] Névéol, A., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., and Zweigenbaum, P. (2016a). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In Balog, K., Cappellato, L., Ferro, N., and Macdonald, C., editors, *CLEF 2016 Working Notes*. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, `http://ceur-ws.org/Vol-1609/`.

[Névéol et al., 2016b] Névéol, A., Cohen, K., Grouin, C., and Robert, A. (2016b). Replicability of research in biomedical natural language processing : a pilot evaluation for a coding task. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 78–84.

[Névéol et al., 2018] Névéol, A., Dalianis, H., Velupillai, S., Savova, G., and Zweigenbaum, P. (2018). Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of Biomedical Semantics*, 9(1) :12.

[Névéol et al., 2009a] Névéol, A., Deserno, T. M., Darmoni, S. J., Güld, M. O., and Aronson, A. R. (2009a). Natural language processing versus content-based image analysis for medical document retrieval. *Journal of the American Society for Information Science and Technology*, 60(1) :123–134.

[Névéol et al., 2015] Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., and Zweigenbaum, P. (2015). CLEF eHealth evaluation lab 2015 task 1b : clinical named entity recognition. In *CLEF 2015 Online Working Notes*. CEUR-WS.

[Névéol et al., 2010] Névéol, A., Islamaj Doğan, R., and Lu, Z. (2010). Author keywords in biomedical journal articles. In *AMIA Annual Symposium Proceedings*, page 537–541. American Medical Informatics Association.

[Névéol et al., 2009b] Névéol, A., Kim, W., Wilbur, W. J., and Lu, Z. (2009b). Exploring two biomedical text genres for disease recognition. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 144–152. Association for Computational Linguistics.

[Névéol et al., 2012] Névéol, A., Li, J., and Lu, Z. (2012). Linking multiple disease-related resources through umls. In *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, pages 767–772. ACM.

[Névéol and Lu, 2010] Névéol, A. and Lu, Z. (2010). Automatic integration of drug indications from multiple health resources. In *Proceedings of the 1st ACM international health informatics symposium*, pages 666–673. ACM.

[Névéol et al., 2006] Névéol, A., Pereira, S., Soualmia, L. F., Thirion, B., and Darmoni, S. J. (2006). A method of cross-lingual consumer health information retrieval. *Studies in health technology and informatics*, 124 :601–608.

[Névéol et al., 2008] Névéol, A., Shooshan, S. E., and Claveau, V. (2008). Automatic inference of indexing rules for medline. *BMC bioinformatics*, 9(11) :S11.

[Névéol et al., 2007a] Névéol, A., Shooshan, S. E., Humphrey, S. M., RINDFLESH, T. C., and Aronson, A. R. (2007a). Multiple approaches to fine-grained indexing of the biomedical literature. In *Proceedings of the Biocomputing Symposium*, pages 292–303. World Scientific.

[Névéol et al., 2007b] Névéol, A., Shooshan, S. E., Mork, J. G., and Aronson, A. R. (2007b). Fine-grained indexing of the biomedical literature : Mesh subheading attachment for a medline indexing tool. In *AMIA Annual Symposium Proceedings*, pages 553–7. American Medical Informatics Association.

[Névéol et al., 2011] Névéol, A., Wilbur, W. J., and Lu, Z. (2011). Extraction of data deposition statements from the literature : a method for automatically tracking research results. *Bioinformatics*, 27(23) :3306–3312.

[Norman et al., 2017] Norman, C., Leeflang, M., and Névéol, A. (2017). LIMSI@CLEF eHealth 2017 task 2 : Logistic regression for automatic article ranking.

[Norman et al., 2018a] Norman, C., Leeflang, M., and Névéol, A. (2018a). Data extraction and synthesis in systematic reviews of diagnostic test accuracy : A corpus for automating and evaluating the process. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

[Norman et al., 2018b] Norman, C., Leeflang, M., and Névéol, A. (2018b). Limsi@clef ehealth 2018 task 2 : Technology assisted reviews by stacking active and static learning.

[Norman et al., 2018c] Norman, C., Leeflang, M., Zweigenbaum, P., and Névéol, A. (2018c). Automating Document Discovery in the Systematic Review Process : How to Use Chaff to Extract Wheat. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[Nzali et al., 2015] Nzali, M. D. T., Névéol, A., and Tannier, X. (2015). Analyse d'expressions temporelles dans les dossiers électroniques patients. In *TALN : Traitement Automatique des Langues Naturelles*.

[Névéol et al., 2014a] Névéol, A., Dalianis, H., Savova, G., and Zweigenbaum, P. (2014a). Didactic panel : Clinical natural language processing in languages other than English. In *Proc AMIA Annu Symp*.

[Névéol et al., 2011] Névéol, A., Doğan, R. I., and Lu, Z. (2011). Semi-automatic semantic annotation of pubmed queries : A study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics*, 44(2) :310 – 318.

[Névéol et al., 2017] Névéol, A., Elhadad, N., Velupillai, S., Xu, H., and Savova, G. (2017). Didactic panel : Clinical natural language processing in languages other than English. In *Proc AMIA Annu Symp*.

[Névéol et al., 2014b] Névéol, A., Grosjean, J., Darmoni, S., and Zweigenbaum, P. (2014b). Language resources for French in the biomedical domain. In *Proc Language and Resource Evaluation Conference, LREC 2014*, pages 2146–2151.

[Névéol et al., 2014c] Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014c). The QUAERO French medical corpus : A resource for medical entity recognition and normalization. In *Proc of BioTextMining Workshop, LREC 2014*, BioTxtM 2014, pages 24–30, Reykjavik, Iceland.

[Névéol et al., 2006] Névéol, A., Rogozan, A., and Darmoni, S. (2006). Automatic indexing of online health resources for a french quality controlled gateway. *Information Processing & Management*, 42(3) :695 – 709.

[Névéol et al., 2009] Névéol, A., Shooshan, S. E., Humphrey, S. M., Mork, J. G., and Aronson, A. R. (2009). A recent advance in the automatic indexing of the biomedical literature. *Journal of Biomedical Informatics*, 42(5) :814 – 823. Biomedical Natural Language Processing.

[Névéol et al., 2004] Névéol, A., Soualmia, L. F., Douyère, M., Rogozan, A., Thirion, B., and Darmoni, S. J. (2004). Using cismef mesh "encapsulated" terminology and a cate-

gorization algorithm for health resources. *International Journal of Medical Informatics*, 73(1) :57 – 64.

[Névéol et al., 2018] Névéol, A., Yepes, A. J., Neves, M., and Verspoor, K. (2018). Parallel Corpora for the Biomedical Domain. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[Névéol and Zweigenbaum, 2015] Névéol, A. and Zweigenbaum, P. (2015). Clinical natural language processing in 2014 : foundational methods supporting efficient healthcare. *Yearb Med Inform*, 10(1) :194–198.

[Névéol and Zweigenbaum, 2016] Névéol, A. and Zweigenbaum, P. (2016). Clinical natural language processing in 2015 : Leveraging the variety of texts of clinical interest. *Yearb Med Inform*, 10(1) :234–239.

[Névéol and Zweigenbaum, 2017] Névéol, A. and Zweigenbaum, P. (2017). Making sense of big textual data for health care : Findings from the section on clinical natural language processing. *Yearb Med Inform*, 26(1) :228–234.

[Névéol and Zweigenbaum, 2018] Névéol, A. and Zweigenbaum, P. (2018). Expanding the diversity of texts and applications : Findings from the section on clinical natural language processing of the international medical informatics association yearbook. *Yearb Med Inform*, pages 188–193.

[Osborne et al., 2018] Osborne, J. D., Neu, M. B., Danila, M. I., Solorio, T., and Bethard, S. J. (2018). Cuiless2016 : a clinical corpus applying compositional normalization of text mentions. *Journal of Biomedical Semantics*, 9(1) :2.

[Pakhomov et al., 2006] Pakhomov, S., Coden, A., and Chute, C. (2006). Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, 75(6) :418–429.

[Pécheux et al., 2014] Pécheux, N., Gong, L., Do, Q. K., Marie, B., Ivanishcheva, Y., Allauzen, A., Lavergne, T., Niehues, J., Max, A., and Yvon, F. (2014). Limsi @ wmt'14 medical translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 246–253, Baltimore, Maryland, USA. Association for Computational Linguistics.

[Pereira et al., 2006] Pereira, S., Névéol, A., Massari, P., Joubert, M., and Darmoni, S. (2006). Construction of a semi-automated icd-10 coding help system to optimize medical and economic coding. In *MIE*, pages 845–850.

[Pham et al., 2014] Pham, A.-D., Névéol, A., Lavergne, T., Yasunaga, D., Clément, O., Meyer, G., Morello, R., and Burgun, A. (2014). Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC Bioinformatics*, 15(1) :266.

[Poibeau, 2005] Poibeau, T. (2005). Parcours interprétatifs et terminologie. page 12. Université de Rouen. http ://www.loria.fr/ yannick/TIA2005/doc/poibeau.pdf.

[Pradhan et al., 2015] Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W., and Savova, G. (2015). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1) :143–154.

[Pustejovsky et al., 2003] Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). TimeML : Robust specification of event and temporal expressions in text. *New directions in question answering*, 3 :28–34.

[Pustejovsky and Stubbs, 2011] Pustejovsky, J. and Stubbs, A. (2011). Increasing informativeness in temporal annotation. In *Proceedings of LAW V - The 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

[Rabary et al., 2015] Rabary, C. T., Lavergne, T., and Névéol, A. (2015). Etiquetage morpho-syntaxique en domaine de spécialité : le domaine médical. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles, Caen, France.*

[Rindflesch and Fiszman, 2003] Rindflesch, T. C. and Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing : interpreting hypernymic propositions in biomedical text. *Journal of biomedical informatics*, 36(6) :462–477.

[Roland Roller and Leser, 2018] Roland Roller, Madeleine Kittner, D. W. and Leser, U. (2018). Cross-lingual candidate search for biomedical concept normalization. In Melero, M., Krallinger, M., and Gonzalez-Agirre, A., editors, *Proceedings of the Multilingual-BIO Workshop : Multilingual Biomedical Text Processing*, pages 16–21, Paris, France. European Language Resources Association (ELRA).

[Ruiz and Névéol, 2007] Ruiz, M. E. and Névéol, A. (2007). Evaluation of automatically assigned mesh terms for retrieval of medical images. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 641–648. Springer.

[Sakji et al., 2010] Sakji, S., Gicquel, Q., Pereira, S., Kergourlay, I., Proux, D., Darmoni, S. J., and Metzger, M. H. (2010). Evaluation of a french medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. In *MedInfo*, pages 252–256.

[Savova et al., 2010] Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., and Chute, C. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) : architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5) :507–13.

[Savova et al., 2012] Savova, G., Styler, W., Albright, D., Palmer, M., Harris, D., Zaramba, G., Haug, P., Clark, C., Wu, S., and Ihrke, D. (2012). SHARP template annotations : Guidelines. Technical report, Mayo Clinic.

[Schulz and Hahn, 2005] Schulz, S. and Hahn, U. (2005). Part-whole representation and reasoning in formal biomedical ontologies. *Artificial Intelligence in Medicine*, 34(3) :179 – 200.

[Smith et al., 2006] Smith, L. H., Rindflesch, T. C., and Wilbur, W. J. (2006). The importance of the lexicon in tagging biological text. *Natural Language Engineering*, 12(4) :335–351.

[Soysal et al., 2017] Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., and Xu, H. (2017). Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3) :331–336.

[Stenetorp et al., 2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT : a web-based tool for NLP-assisted text annotation. In *Proc of EACL Demonstrations*, pages 102–107, Avignon, France. ACL.

[Sun et al., 2013] Sun, W., Rumshisky, A., and Uzuner, O. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5) :806–813.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Tapi Nzali et al., 2015] Tapi Nzali, M. D., Tannier, X., and Neveol, A. (2015). Automatic extraction of time expressions accross domains in french narratives. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 492–498, Lisbon, Portugal. Association for Computational Linguistics.

[Tchechmedjiev et al., 2018] Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., and Jonquet, C. (2018). Enhanced functionalities for annotating and indexing clinical text with the ncbo annotator+. *Bioinformatics*, 34(11) :1962–1965.

[Tiedemann, 2009] Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

[Tourille et al., 2018] Tourille, J., Doutreligne, M., Ferret, O., Paris, N., Névéol, A., and Tannier, X. (2018). Evaluation of a sequence tagging tool for biomedical texts. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*.

[Tourille et al., 2016] Tourille, J., Ferret, O., Névéol, A., and Tannier, X. (2016). Limsi-cot at semeval-2016 task 12 : Temporal relation identification using a pipeline of classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1136–1142, San Diego, California. Association for Computational Linguistics.

[Tourille et al., 2017a] Tourille, J., Ferret, O., Neveol, A., and Tannier, X. (2017a). Neural architecture for temporal relation extraction : A bi-lstm approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 224–230. Association for Computational Linguistics.

[Tourille et al., 2017b] Tourille, J., Ferret, O., Tannier, X., and Névéol, A. (2017b). Limsi-cot at semeval-2017 task 12 : Neural architecture for temporal information extraction from clinical narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 597–602, Vancouver, Canada. Association for Computational Linguistics.

[Tourille et al., 2017c] Tourille, J., Ferret, O., Tannier, X., and Névéol, A. (2017c). Temporal information extraction from clinical text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, pages 739–745, Valencia, Spain. Association for Computational Linguistics.

[Tsafnat et al., 2014] Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., and Coiera, E. (2014). Systematic review automation technologies. *Systematic reviews*, 3(1) :74.

[Tutubalina et al., 2018] Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., and Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84 :93 – 102.

[Uzuner, 2007] Uzuner, . (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5) :550 – 563.

[Verspoor et al., 2013] Verspoor, K., Jimeno Yepes, A., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., and Plazzer, J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, 2013.

[Wu et al., 2014] Wu, S., Miller, T., Masanz, J., Coarr, M., Halgrim, S., Carrell, D., and Clark, C. (2014). Negation's not solved : generalizability versus optimizability in clinical natural language processing. *PLoS One*, 9(11) :e112774.

[Wüster, 1981] Wüster, E. (1981). *L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. Textes choisis de terminologie 1, Fondements théoriques de la terminologie*. Presses de l'université de Laval.

[Zweigenbaum et al., 2007] Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining : current progress. *Briefings in bioinformatics*, 8(5) :358–375.

[Zweigenbaum et al., 2016] Zweigenbaum, P., Grouin, C., and Lavergne, T. (2016). Supervised classification of end-of-lines in clinical text with no manual annotation. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 80–88, Osaka, Japan. The COLING 2016 Organizing Committee.

Annexes

Annexe 1

1 Résultats des participants aux campagnes CLEF eHealth (2015-2018)

Le tableau 5.1 présente les performances des systèmes sur la tâche de reconnaissance d'entités nommées ; le tableau 5.2 présente les performances des systèmes sur la tâche de reconnaissance d'entités normalisées.

Le tableau 5.3 présente les performances des systèmes sur la tâche de normalisation d'entités.

Le Tableau 5.4 présente les performances des systèmes sur les textes non alignés pour l'anglais (2017), le français, le hongrois et l'italien (2018).

2 Résultats des participants à la tâche de traduction biomédicale de WMT (2016-2018).

Le Tableau 5.5 présente les performances des systèmes pour les paires de langues EN/ES, EN/FR et EN/PT sur les corpus Scielo et EDP lors des campagnes WMT de 2016 à 2018.

	EMEA						MEDLINE					
	2015			2016			2015			2016		
Team	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Erasmus-run3.unofficial	-	-	-	*0.72*	*0.79*	***0.75***	-	-	-	***0.68***	***0.72***	***0.70***
Erasmus-run1	0.75	0.76	**0.76**	0.62	**0.80**	0.70	0.71	0.63	**0.67**	0.62	0.69	0.65
Erasmus-run2	0.71	**0.78**	0.74	0.63	0.79	0.70	0.68	**0.64**	0.66	0.62	0.68	0.65
IHS-RD-run1-fix	*0.86*	*0.60*	*0.70*	-	-	-	*0.76*	*0.40*	*0.53*	-	-	-
Watchdogs-run1	**0.86**	0.55	0.67	-	-	-	0.71	0.41	0.52	-	-	-
IHS-RD-run2-fix	*0.80*	*0.57*	*0.67*	-	-	-	-	-	-	-	-	-
HIT-WI-run1-fix	*0.81*	*0.43*	*0.56*	-	-	-	-	-	-	-	-	-
LITL-run1	-	-	-	**0.78**	0.40	0.53	-	-	-	0.64	0.32	0.43
LITL-run2	-	-	-	0.77	0.39	0.52	-	-	-	0.64	0.32	0.43
LIMSI-run1	0.60	0.42	0.49	-	-	-	0.57	0.38	0.46	-	-	-
Watchdogs-run2	0.36	0.58	0.44	-	-	-	0.40	0.46	0.43	-	-	-
SIBM-run1	0.00	0.00	0.00	0.54	0.38	0.44	0.13	0.23	0.17	0.54	0.48	0.51
SIBM-run2	-	-	-	0.60	0.33	0.43	-	-	-	0.64	0.44	0.52
BITEM-run1	-	-	-	0.52	0,18	0.27	-	-	-	0.57	0.44	0.50
UPF-run1	0.00	0.00	0.00	0.13	0.22	0.16	-	-	-	0.13	0.24	0.17
UPF-run1-fix	*0,05*	*0,14*	*0,07*	-	-	-	-	-	-	-	-	-
UPF-run2.unofficial	-	-	-	*0.10*	*0.19*	*0.13*	-	-	-	0.13	0.24	0.17
HIT-WI Lab-run1	0.01	0.01	0.01	-	-	-	0.61	0.36	0.45	-	-	-
IHS-RD-run1	0.00	0.00	0.00	-	-	-	0.31	0.03	0.05	-	-	-
IHS-RD-run2	0.00	0.00	0.00	-	-	-	**0.76**	0.40	0.52	-	-	-
average	0.33	0.31	0.31	0.58	0.44	0.47	0.50	0.36	0.40	0.50	0.43	0.45
median	0.18	0.21	0.22	0.61	0.39	0.48	0.65	0.40	0.49	0.62	0.44	0.50

TABLE 5.1 – System performance for plain entity recognition on the development (2015) and test (2016) corpus. Data shown in *italic font* presents runs that were submitted after the official deadline. The median and average are computed solely using the official runs.

	EMEA						MEDLINE					
	2015			2016			2015			2016		
Team	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Erasmus-run3.unofficial	-	-	-	***0.70***	***0.64***	***0.67***	-	-	-	***0.60***	***0.61***	***0.61***
Erasmus-run1	**0.71**	**0.71**	**0.71**	0.49	0.58	0.53	**0.547**	**0.634**	**0.587**	0.410	0.562	0.474
Erasmus-run2	0.71	0.65	0.68	0.48	0.59	0.53	0.55	0.60	0.58	0.40	0.57	0.47
SIBM-run1	0.00	0.00	0.00	0.27	0.38	0.32	0.30	0.19	0.23	0.36	0.40	0.38
SIBM-run2	-	-	-	0.21	0.39	0.27	-	-	-	0.33	0.48	0.39
BITEM-run1	-	-	-	0.16	0.45	0.23	-	-	-	0.38	0.49	0.43
HIT-WI-run1-fix	*0.19*	*0.37*	*0.25*	-	-	-	*0.05*	*0.59*	*0.09*	-	-	-
IHS-RD-run1-fix	*0.051*	*0.57*	*0.09*	-	-	-	0.04	0.40	0.07	-	-	-
HIT-WI-run1	0.00	0.01	0.01	-	-	-	0.17	0.30	0.22	-	-	-
average	0.29	0.27	0.28	0.32	0.48	0.38	0.32	0.42	0.37	0.38	0.50	0.43
median	0.00	0.01	0.01	0.27	0.48	0.32	0.30	0.40	0.23	0.38	0.49	0.43

TABLE 5.2 – System performance for normalized entity recognition on the development (2015) and test (2016) corpus. Data shown in *italic font* presents runs that were submitted after the official deadline. The median and average are computed solely using the official runs.

	EMEA						MEDLINE					
	2015			2016			2015			2016		
Team	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Erasmus-run2	**0.77**	**1.00**	**0.87**	-	-	-	**0.58**	0.81	0.67	-	-	-
Erasmus-run1	0.77	**1.00**	0.87	-	-	-	0.57	**0.82**	**0.67**	-	-	-
SIBM-run2	-	-	-	**0.60**	0.46	**0.52**	-	-	-	**0.60**	0.47	0.52
SIBM-run1	-	-	-	0.57	**0.48**	0.52	-	-	-	0.59	**0.52**	**0.55**
HIT-WI-run1	0.56	0.55	0.56	-	-	-	0.47	0.47	0.47	-	-	-
IHS-RD-run1	0.06	0.69	0.10	-	-	-	0.04	0.58	0.08	-	-	-
UPF-run1	-	-	-	0.48	0.48	0.48	-	-	-	0.48	0.47	0.47
average	0.53	0.90	0.62	0.55	0.47	0.51	0.40	0.73	0.48	0.56	0.49	0.57
median	0.77	1.00	0.87	0.57	0.48	0.52	0.57	0.81	0.67	0.59	0.47	0.53

TABLE 5.3 – System performance for entity normalization on the development (2015) and test (2016) corpus.

	2017			2018								
	American			**French**			**Hungarian**			**Italian**		
Team	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
IxaMed run2	-	-	-	.877	.588	.704	**.970**	**.955**	**.963**	.945	.922	.934
IxaMed run1	-	-	-	.872	**.597**	**.709**	.968	.954	.961	**.960**	**.945**	**.952**
LSI UNED-run2	-	-	-	.879	.540	.669	.946	.911	.928	.931	.861	.895
LSI UNED-run1	-	-	-	.842	.556	.670	.932	.922	.927	.917	.875	.895
IAM-run2	-	-	-	.820	.560	.666	-	-	-	-	-	-
IAM-run1	-	-	-	.807	.555	.657	-	-	-	-	-	-
TorontoCL-run2	-	-	-	.922	.897	.910	.900	.829	.863			
TorontoCL-run1	-	-	-	.901	.887	.894	.908	.824	.864			
WebIntelligentLab	-	-	-	.702	.495	.580	-	-	-	-	-	-
ECNUica-run1	-	-	-	.790	.456	.578	-	-	-	-	-	-
KFU-run1	.893	.811	**.850**	-	-	-	-	-	-	-	-	-
KFU-run2	.891	**.812**	.850	-	-	-	-	-	-	-	-	-
LIMSI-run1	.909	.765	.831	-	-	-	-	-	-	-	-	-
TUC-MI-run1	**.940**	.725	.819	-	-	-	-	-	-	-	-	-
SIBM-run1	.839	.783	.810	-	-	-	-	-	-	-	-	-
TUC-MI-run2	.929	.717	.809	-	-	-	-	-	-	-	-	-
LIRMM-run1	.691	.514	.589	-	-	-	-	-	-	-	-	-
LIRMM-run2	.646	.527	.580	-	-	-	-	-	-	-	-	-
ims unipd-run1	.496	.442	.468	.653	.396	.493	.761	.748	.755	.535	.484	.509
Mondeca-run1	.691	.309	.427	-	-	-	-	-	-			
ims unipd-run2	.382	.341	.360	-	-	-	-	-	-	-	-	-
techno	-	-	-	.569	.286	.380	-	-	-	-	-	-
WBI-run2	.616	.606	.611	.512	.253	.339	.522	.388	.445	.862	.689	.766
WBI-run1	.616	.606	.611	.494	.246	.329	.518	.384	.441	.857	.685	.761
UNSW-run1	.401	.352	.375	-	-	-	-	-	-	-	-	-
UNSW-run2	.371	.328	.348	-	-	-	-	-	-	-	-	-
KCL-Health-NLP-run1	-	-	-	.738	.405	.523	-	-	-	.746	.636	.687
KCL-Health-NLP-run2	-	-	-	.724	.394	.510	-	-	-	.725	.616	.666
APHP-run1	-	-	-	.668	.601	.633	-	-	-	-	-	-
APHP-run2	-	-	-	.816	.607	.696	-	-	-	-	-	-
KR-ISPED-corrected	-	-	-	.676	.323	.437	-	-	-	-	-	-
average	.670	.582	.622	.723	.410	.507	.844	.761	.799	.827	.783	.803
median	.646	.606	.611	.798	.475	.579	.900	.824	.863	.922	.897	.910

TABLE 5.4 – System performance for ICD10 coding on the **American**, **French**, **Hungarian** and **Italian** test corpus in terms of Precision (P), recall (R) and F-measure (F).

		Scielo				EDP	
		2016		2017	2018	2017	2018
Language	**Team**	BLEU		BLEU	BLEU	BLEU	BLEU
EN-PT	IstrionBox	**17.55**	**19.01**	-	-	NA	NA
	UHH	-	-	**39.38**	34.92	NA	NA
	UFRGS	-	-	-	**39.43**	NA	NA
	baseline	15.38	17.22	30.52	-	NA	NA
PT-EN	IstrionBox	**20.88**	21.50	-	-	NA	NA
	UHH	-	-	**43.93**	41.84	NA	NA
	TALP	-	-	-	39.49	NA	NA
	UFRGS	-	-	-	**42.58**	NA	NA
	baseline	17.59	18.48	36.35	-	NA	NA
EN-ES	IXA	31.57	28.13	-	-	NA	NA
	TALP	**33.22**	**29.47**	-	-	NA	NA
	UHH	-	-	**36.23**	31.33	NA	NA
	UFRGS	-	-	-	**39.77**	NA	NA
	baseline	17.82	16.88	27.31	-	NA	NA
ES-EN	IXA	30.66	28.12	-	-	NA	NA
	TALP	29.83	27.42	40.49	-	NA	NA
	uedin	**31.49**	**29.02**	-	-	NA	NA
	UHH	-	-	**37.49**	36.16	NA	NA
	UFRGS	-	-	-	**43.41**	NA	NA
	baseline	18.78	16.92	27.31	-	NA	NA
EN-FR	LIMSI	-	**22.75**	NA	NA	-	-
	Hunter	-	-	NA	NA	17.50	**23.24**
	kyoto	-	-	NA	NA	**27.04**	-
	UHH	-	-	NA	NA	22.79	-
	baseline	-	9.24	NA	NA	12.32	-
FR-EN	Hunter	-	-	NA	NA	15.18	-
	kyoto	-	-	NA	NA	**25.21**	-
	UHH	-	-	NA	NA	23.41	-
	TALP	-	-	NA	NA	-	**25.78**
	baseline	-	-	NA	NA	17.47	-

TABLE 5.5 – Translation performance for biomedical text in ES/EN, FR/EN and PT/EN language pairs in WMT biomedical campaigns (Best BLEU scores reported for each team).