Construction d'un corpus biomédical pour la recherche d'informations

Valentina Dragos, Marie-Christine Jaulent
INSERM, UMRS 872, eq. 20
15, Rue de l'école de médecine
75006 Paris
valentina.dragos@spim.jussieu.fr
marie-christine.jaulent@spim.jussieu.fr

Résumé. Un corpus textuel est composé d'un ensemble de textes existants réunis selon une visée interprétative. Dans le domaine du traitement automatique de la langue ou dans le domaine de la recherche d'informations, les approches à base de corpus sont nombreuses, car utiliser un corpus offre un ancrage dans le réel qui apporte de la valeur aux résultats obtenus. Nous présentons dans ce papier un travail réalisé pour constituer un corpus scientifique du domaine biomédical. Nous présentons d'abord les motivations de notre travail, le protocole utilisé pour constituer le corpus ainsi que les résultats préliminaires issus de l'analyse de ce corpus. La phase d'analyse étudie plus particulièrement l'émergence de termes médicaux autour du thème « tabac », qui n'est pas spécifique mais uniquement incident au domaine médical. Nous terminons sur l'utilisation actuelle du corpus, en lien avec l'application des principes de la sémantique textuelle interprétative à la recherche d'informations.

1 Introduction

La construction d'un corpus électronique est contrainte par l'hypothèse de recherche qui en a motivé le besoin et par le paradigme théorique dans lequel elle s'inscrit. Dans ce travail, nous nous sommes intéressées à la construction d'un corpus scientifique du domaine biomédical, dédié, indirectement, à la recherche d'informations. Cette tâche s'inscrit dans le cadre d'un projet visant à appliquer les principes de la sémantique textuelle interprétative à la recherche d'informations. La sémantique textuelle vise à caractériser de manière générale les textes, son hypothèse de travail résidant dans le postulat de la diversité des textes, car d*e la même façon que la diversité des langues est le problème fondateur de la linguistique, la diversité des textes fonde la sémantique des textes*, cf. (Rastier, 2001).

La sémantique textuelle s'intéresse aux manifestations linguistiques en problématisant le lien entre les textes et les différentes facettes (social, historique, etc.) du contexte de leur production. Elle suppose une linguistique des genres et des discours, mais également une analyse visant à identifier les traits sémantiques (thématiques, structurels, dialectiques, etc.) des textes dont il s'agit de décrire l'interaction.

C'est pour cette raison que nous affirmons que le corpus construit est dédié *indirectement* à la recherche d'informations : il permet de délimiter un univers thématique (le domaine médical) et discursif (le genre scientifique) dont il est question d'identifier les traits sémantiques. Les traits sémantiques ainsi mis en évidence sont utilisés, a posteriori, pour la caractérisation automatique des textes permettant ainsi l'optimisation des processus de recherche

d'informations. Les traits sémantiques sont identifiés en deux étapes : tout d'abord en analysant le corpus constitué et ensuite en le contrastant avec différents corpus. Dans le cadre du projet, l'utilisation de plusieurs corpus est motivée du point de vue méthodologique, car il est nécessaire d'identifier les traits sémantiques de plusieurs types de textes afin de faciliter, ultérieurement, la recherche d'informations. Le genre scientifique a été choisi pour assurer la diversité de genres dans le cadre du projet global.

2 Constitution du corpus

L'objectif du travail réalisé a été de construire un corpus spécialisé du domaine biomédical, dont le thème est le tabac. Par corpus on entend ici un ensemble de documents réuni selon une visée interprétative (cf. Pincemin, 1999). La spécialisation est imposée par le domaine biomédical, le thème et le genre choisis. La construction a été réalisée afin d'obtenir un échantillon représentatif d'une langue de spécialité et d'un genre particulier et en ayant pour origine une hypothèse de recherche. Il s'agit de l'hypothèse de recherche propre à la sémantique textuelle qui représente le paradigme théorique adopté. Nos critères de section ont été fondés sur une configuration domaine-genre (le domaine médical, le genre scientifique).Nous avons constitué le corpus manuellement, en effectuant des recherches sur des sites web de professionnels du domaine médical Il s'agit de sites institutionnels Inserm (Institut National de la Santé et de la Recherche Médicale), Cismef (CHU de Rouen), HAS (Haute Autorité de la Santé), HON (Health on The net) et l'Affsafs (Agence française de sécurité sanitaire des produits de santé)), qui s'adressent aux professionnels de la santé mais également au grand public, en mettant à leur disposition différents types de documents : articles scientifiques, expertises collectives, brochures informatives., etc.. La recherche a été limitée aux textes produits au cours des 5 dernières années et dans notre démarche nous avons utilisé comme expert un médecin tabacologue.

Les documents ont été retrouvés en utilisant des requêtes d'interrogation. Ces requêtes ont été élaborées en utilisant les mots clés : tabac, tabagisme, nicotine. Un ensemble de syntagmes identifié sur le site HON nous a permis d'enrichir nos requêtes. Il s'agit de syntagmes nominaux mettant en évidence différents aspects du tabagisme dans le domaine médical et de la santé publique : tabagisme passif, pollution environnementale par la fumée de tabac, pollution de l'air par la fumée du tabac, consommation de tabac, habitude de fumer, utilisation du tabac, usage du tabac, sevrage tabagique et désaccoutumance au tabac.

Sur les sites sélectionnés nous avons retrouvé des articles scientifiques publiés dans 11 revues médicales et une expertise collective élaborée par des experts de l'Inserm. Le corpus construit est constitué de 120 documents et contient environ 725 000 mots (18M). Chaque document est enrichi par un ensemble de méta-données indiquant : les auteurs (noms et affiliations), l'année de publication, la revue source et les mots-clés du document.

3 Analyse du corpus

L'analyse du corpus vise à identifier les traits sémantiques du corpus constitué. Cette identification permet, par la suite, d'améliorer les processus de recherche d'informations en développant des procédures automatiques pour la reconnaissance de documents. Les traits sémantiques sont identifiés en analysant le corpus constitué, mais également en le contrastant avec différents corpus afin de mettre en évidence des similitudes et des contrastes pouvant

apparaître en changeant le type de discours ou le genre des documents. L'étude constrastive du corpus étant en cours, nous présentons dans ce papier les résultats d'une analyse réalisée d'un point de vue linguistique. Cette analyse a été réalisée semi-automatiquement. Dans un premier temps nous avons identifié des catégories homogènes au sein du corpus ; dans un deuxième temps nous avons étudié la lexicalisation des catégories, en mettant en évidence les termes spécifiques à chaque catégorie de documents. Ces deux étapes sont détaillées *infra.*

3.1 Catégorisation manuelle des données

Les documents constituants du corpus ont été structurés manuellement en six catégories homogènes. I*nfra* nous présentons ces catégories en illustrant chacune par un paragraphe extrait d'un article scientifique lui appartenant.

La catégorie « tabac facteur négatif » regroupe 36 articles qui expriment, de manière explicite, un lien causal entre le tabac et différentes pathologies. Ex : *le développement des cancers urothéliaux est favorisé par une substance chimique comme les nitrosamines mais aussi par des composés retrouvés dans la fumée du tabac.*

La catégorie « tabac facteur négatif implicite » est constituée de 18 articles scientifiques dont la thématique (portant, le plus souvent sur le sevrage, les thérapies cognitives, etc.) sous-entend une hypothèse de travail implicite liée à la nocivité du tabac. Ex : *Pour l'arrêt du tabac, le premier choix est l'utilisation de substituts nicotiniques, à la fois efficaces et sûrs.* Les textes de cette catégorie n'expriment pas, explicitement un lien causal entre le tabac et les différentes pathologies.

Dans la catégorie « tabac facteur négatif indirect» nous retrouvons 28 articles dans lesquels le tabac apparaît, explicitement comme un facteur aggravant d'une pathologie. Ex. : *Les principaux facteurs de risque d'athérosclérose, comme l'hypercholestérolémie, le diabète, l'hypertension et le tabac…*La catégorie « tabac facteur neutre» est constituée d'uniquement 3 articles dans lesquels le tabac apparaît comme un facteur neutre par rapport à une pathologie considérée. Ex. : *Le tabac, le diabète, l'hypertension artérielle et l'hypercholestérolémie n'influencent pas le risque de développer un accident thrombotique.*

La cinquième catégorie s'intitule « tabac facteur non déterminé» et est constituée de 9 articles affirmant l'impossibilité d'identifier, le rôle du tabac dans l'évolution d'un pathologie. Ex. *: Aucun étude environnementale ni épidémiologique n'a mis en évidence le rôle du tabac sur les niveaux d'endotoxines.*

La catégorie « sans objet » regroupe des textes abordant le sujet «tabac » selon un point de vue qui n'est pas propre au domaine médical. Ex. : *...ont été simultanément introduites dans le génome de cellules de tabac….*

3.2 Analyse des catégories

Pour mettre en évidence les termes spécifiques à chaque catégorie nous avons exploité à la fois les mots clés de chaque document et les résultats fournis par l'outil Lexico, (cf . Salem et all., 2003), c'est-à-dire les formes ayant une fréquence importante d'apparition au sein de chaque catégorie. Les résultats obtenus sont présentés dans le tableau 1.

Catégorie	Mots-clés des documents	Termes fréquents (Lexico)
Tabac facteur négatif (1)	Tabagisme, cancer, stress psychosocial, chimiothérapie, sevrage, dépendance, tabagisme passif	Tabac, risque, tabagisme, cancer, études, patients, nicotine, fumeurs
Tabac facteur négatif implicite (2)	Cancer, grossesse, arrêt du tabac, consommation de tabac, exposition au tabac, sevrage tabagique, dépendances	Tabac, fumeurs, nicotine, tabagisme, sevrage, dépendance
Tabac facteur négatif indirect (3)	Pathologie infantile, maladie respiratoire, cancer, hypertension artérielle	Patients, risque, cancer, traitement
Tabac facteur neutre (4)	Cancer, consommation cannabis, stratégies thérapeutiques	Patients, cannabis, mésothéliome
Tabac facteur non déterminé (5)	Cancer de la prostate, condylome externe	Nicotine, fumeurs, cancer, études, tabac, risque
Sans objet (6)	Virologie, médecine générale, analyse chimique, coût social	Nicotine, tabac, virus, récepteurs, fumeurs

Tableau 1 : Catégories de documents et lexicalisation

L'étude des résultats obtenus permet de repérer facilement deux catégories (la première et la deuxième) dont la thématique, bien que médicale, reste centrée autour du tabac. Le tabac est reflété selon trois dimensions : les pathologies sous-jacentes (cancer, chimiothérapie, patients), la santé publique (tabagisme, exposition au tabac, risque) et les troubles comportementales (consommation de tabac, sevrage tabagique, dépendances). Notons au passage que le lexème «nicotine» appartient à cette dernière dimension, en sa qualité de substance psychoactive. Les documents de ces catégories mettent en évidence l'émergence de concepts relatifs à la tabacologie. Ce processus est réalisé par deux mécanismes : le premier consiste à construire des structures linguistiques construites autour d'une racine commune, comme par exemple : *consommer du tabac*, *consommateurs de tabac*, *consommation de tabac*, *consommation chronique de tabac* et *consommation compulsive de tabac*. Le deuxième mécanisme repose sur la construction de syntagmes dont les termes centraux sont proches semantiquemment, comme c'est la cas pour : *risques du tabac* et *dangers du tabac*.

La deuxième et la troisième catégorie élargissent la thématique médicale, en évoquant les maladies respiratoires et de la petite enfance, mais nous ne trouvons plus des termes relatifs au tabac. La cinquième catégorie est caractérisée par des mots clés évoquant une thématique médicale mais ses formes fréquentes sont spécifiques au tabac. Néanmoins, ces résultats doivent être interprétés prudemment, car cette catégorie a une taille critique (3 articles). Enfin, la dernière catégorie est caractérisée par une thématique très large, relevant à la fois du domaine médical mais aussi des domaines connexes (virologie, génétique). Le lexème «tabac», bien que présent parmi les formes fréquentes, représente, le plus souvent la plante.

De manière générale, les textes du corpus sont rédigés dans un langage scientifique impersonnel, caractérisé par l'absence des pronoms personnels et l'emploi souvent de *on* et *nous* . Notons également la prévalence de formes de pluriel : 1330 occurrences de *fumeurs* contre 330 pour *fumeur* , 1437 occurrences pour *patients* contre 329 pour *patient* , la tendance étant confirmée également au niveau du genre : 81 occurrences pour *patientes* contre 13 occurrences de *patiente*.

4 Perspectives

Dans ce papier nous avons présenté la construction d'un corpus spécialisé du domaine médical dédié à la recherche d'informations ainsi que des résultats préliminaires issus de l'analyse de ce corpus. Nous nous sommes attachées à construire un corpus homogène à la fois d'un point de vue thématique et générique. Nous estimons que le corpus satisfait les critères énoncés pas (Pincemin, 1999) de signifiance (car nous avons enrichi les requêtes utilisées pour retrouver des documents), acceptabilité (construction sans contraintes externes) et un peu moins le critère d'exploitabilité car il est constitué de moins d'1 0000 0000 de mots. Cependant, le corpus est exploité dans le cadre d'un projet plus ample, dont il représente un sous-corpus, ce qui permet de pallier cet inconvénient et on pourrait le considérer représentatif (cf. Habert, 2000) du domaine médical pour la thématique «tabac».

Une première perspective de ce travail concerne l'approfondissement de l'analyse du corpus constitué qui permettrait d'identifier ses traits sémantiques. La deuxième perspective concerne son étude contrastive avec d'autres types de corpus, qui permettrait de mettre en évidence les similitudes ou les différences engendrées par le changement du type de discours ou du genre au sein de la même thématique. Plus généralement, le corpus construit pourrait être intégré dans d'autres corpus du domaine biomédical, pour mettre en évidence la diversité de la langue employée dans ce domaine, voir (Zweigenbaum et all., 2001). ou pourrait être utilisé pour la mise en oeuvre d'autres tâches, comme par exemple la fouille de textes, cf. (Tsalidis et all., 2007), (Cohen et all., 2005).

Références

Cohen B., L. Fox, P. Ogren et L. Hunter (2005) Corpus Design for biomedical natural language processing. ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Detroit.

Habert B. (2000) Des corpus représentatifs : de quoi, pour quoi, comment ? In: Linguistique sur corpus. Études et réflexions, M. Bilger (éd), Perpignan, Presses Universitaires de Perpignan, 11-58.

Pincemin Bénédicte (1999) Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative, Atelier Corpus et TAL : pour une réflexion méthodologique, 6e Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 99), Cargèse (Corse, France), 12-17 juillet 1999, Anne Condamines, Marie-Paule Péry-Woodley et Cécile Fabre (éds), pp. 26-36.

Rastier, F. (2001) Arts et sciences du texte, Paris, Presses Universitaires de France.

Salem, A., Lamaille, C., Martinez, W. et Fleury, S. (2003). *Manuel Lexico 3*, version 3.41.

Tsalidis, C., Orphanos, G. , Mantzari, E., Pantazara, M., Diolis, C. et Aristides Vagelatos (2007) Developing a Greek Biomedical Corpus towards Text Mining**,** The fourth Corpus Linguistics conference, Birmingham.

Zweigenbaum P., Jacquemart P., Grabar N., Habert B. (2001) Building a text corpus for representing the variety of medical language, Medinfo 2001: Proceedings of the 10th World Congress on Medical Informatics.