Corpus de registres différents pour le développement d'un aligneur d'unités polylexicales

Claire Lemaire

Corpus de registres différents pour le développement d'un aligneur d'unités polylexicales

Claire Lemaire
Laboratoire d'Informatique de Grenoble
claire.lemaire@imag.fr

Résumé

Comment trouver des données exprimant les mêmes concepts dans des registres de langue différents ? Après un essai infructueux d'extraction terminologique à partir de corpus comparables spécialisés dans trois langues différentes, l'idée est d'ajouter pour chaque langue des sous-corpus du registre courant afin d'y détecter des relations de synonymie. Or ce type de ressource n'existe pas. Nous présentons la constitution d'un corpus de 400 K mots en allemand dans le domaine de la cancérologie, subdivisé en deux sous-corpus de même taille. À partir d'équivalents en allemand du mot clé « cancer du sein », nous avons recueilli pour un premier sous-corpus, des textes qui s'adressent à des patientes (et des patients) ou à leurs familles, et pour un second sous-corpus, des textes qui s'adressent à des médecins ou des chercheurs en médecine.

Introduction

Dans le cadre de la communication d'entreprise, il est exprimé le besoin d'obtenir des données décrivant les mêmes signifiés mais dans des registres différents. La première idée a été de collecter des textes d'experts dans un domaine spécialisé et dans différentes langues, afin d'extraire automatiquement de ces corpus des termes ayant à l'intérieur d'une même langue une relation de synonymie. L'extraction terminologique à partir des corpus de textes pour experts n'a pas fourni les résultats escomptés (Delpech, 2013). La seconde idée pour créer cet aligneur d'unités polylexicales est alors d'ajouter au corpus des textes sur un sujet d'expertise, mais cette fois rédigés en langue générale, afin de s'assurer de la présence dans la même langue de plusieurs termes ayant une relation paraphrastique entre eux. Un nouveau besoin apparaît alors en corpus de qualité, construit à partir de textes d'experts du domaine médical dans différentes langues, pour une part rédigés pour d'autres experts, et pour une autre part égale, à destination du grand public. Or il s'avère qu'une telle ressource n'existe pas, du moins dans le domaine médical.

Nous avons construit la ressource nécessaire pour l'allemand, à partir de textes contenant les mots « Brustkrebs » ou « Mammakarzinom », équivalents du terme « cancer du sein ». La première partie de cet article présente comment les textes ont été collectés, choisis et recensés. La seconde partie décrit le nettoyage des textes, leur anonymisation et le résultat final.

1 Recueil des données

Dans un premier temps, nous avons recherché des textes en allemand médical rédigés par des médecins, des journalistes médicaux, et des chercheurs en médecine. Au bout de 6 mois, nous sommes arrivés au bout de ce qui était publié en ligne dans l'espace germanophone représenté par l'Allemagne, l'Autriche et la Suisse ; nous avons ensuite enrichi le corpus au fur et à mesure des sorties de nouveaux articles. Nous avons rassemblé des textes principalement issus des sites suivants :

(1) textes scientifiques :

- « Deutsches Ärzteblatt », un magazine hebdomadaire médical publié en Allemagne par l'éditeur « Deutscher Ärzte-Verlag » à 370 000 exemplaires ;
- les articles publiés en ligne par « Universitäts-Brustzentrum Tübingen », l'institut de recherche en sénologie de l'Université de Tübingen ;
- « Ärzte Zeitung », un quotidien allemand pour les médecins publié par « Springer », qui paraît à 49 030 exemplaires ;
- les articles publiés en ligne par la « Deutsche Gesellschaft für Senologie », l'institut de recherche en sénologie de l'Allemagne ;

(2) textes généraux :

- « Netdoktor », un site scandinave d'information en matière de santé, de pharmacie et de médecine ;
- « Onkosupport », un site d'information et de soutien aux malades du cancer.

Nous avons collecté 278 textes en langue de spécialité en tout, 104 dans un registre courant et 163 autres dans un registre soutenu, soit 219 206 et 204 960 mots respectivement, soit un total de 424 166 mots avant traitement.

Notons que si l'unité choisie en linguistique de corpus est traditionnellement le mot, celle-ci est particulièrement mal adaptée pour l'allemand. En effet, l'allemand recourt très fréquemment aux mots composés, y compris dans le registre courant. C'est pourquoi, le caractère est souvent utilisé Voici un exemple de mots issus de notre corpus ainsi que leur correspondant en anglais, puisqu'il s'agit de la langue de base de l'aligneur terminologique en question (et en français pour donner un ordre d'idée au lecteur).

anglais	allemand	français
family members	Familienangehörigen	membre de la famille
quality controls	Qualitätskontrollen	contrôles de qualité
risk of disease	Erkrankungsrisiko	risque de maladie
breast cancer cases	Brustkrebserkrankungen	cas de cancer du sein
size difference	Größendifferenz	différence de taille
early detection concept	Früherkennungskonzept	concept de dépistage précoce
probability of developing the disease	Erkrankungswahrscheinlichkeiten	probabilité de développer la maladie
20 mots	**7 mots**	**27 mots**

Figure 1 : Les mots composés en allemand

2 Traitement des données

Nous avons lu chaque texte intégralement et retiré manuellement de chacun d'entre eux les références bibliographiques, les noms d'auteurs, les dates, les publicités, les résumés desdits textes, les biographies et les notes sur les intérêts et influences des auteurs. Les images ont également été retirées, ainsi que toutes les balises html. Voici un tableau récapitulatif du nombres de mots (allemands !) recueillis avant et après le traitement des fichiers. Nous obtenus en tout 406 049 mots exploitables, soit un peu plus de 3 millions de caractères.

	Langue scientifique	Langue générale	Total
Nombre de textes	104	163	278
Nombre de mots de la ressource brute	219 206	204 960	424 166
Nombre de pages de la ressource brute	151	141	292
Nombre de mots après traitement	201 119	204 930	406 049
Nombre de pages après traitement	138	141	279

Figure 2 : Le corpus avant et après le formatage

Cependant, si le corpus a été formaté, il n'a volontairement pas été normalisé. D'une part, les paragraphes ont été conservés et d'autre part, la casse n'a pas été modifiée. En effet, quelle que soit leur position dans la phrase, tous les substantifs commencent par une majuscule en allemand, il est donc important de conserver celles-ci non seulement pour cet aligneur d'unités polylexicales mais également pour toutes études terminologiques ultérieures.

Conclusion

Outre le développement de cet aligneur, une petite partie de notre ressources (6 K caractères) vient de permettre d'améliorer la couverture d'un analyseur morphologique. Nous poursuivons actuellement les travaux sur cet analyseur en y passant l'intégralité du corpus, (par section de 100 K caractères), afin d'obtenir la liste exhaustives des différents termes.

Références

Colstoun F., Delpech E., Monneret É. : *Libellex : une plateforme multiservices pour la gestion des contenus multilingues*. In Actes de la 18ème conférence sur le Traitement Automatique des Langues Naturelles (TALN), 27 juin–1er juillet, Montpellier, France, 2011.

Dejean, E. et Gaussier E. (2002). *Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables.* Lexicometrica. Alignement lexical dans les corpus multilingues. Pages 1-22.

Delpech, E. (2013). *Traduction assistée par ordinateur et corpus comparables : contributions à la traduction compositionnelle*. Thèse de doctorat. Université de Nantes.

Kilgariff, A. (2001). *Comparing Corpora*. International Journal of Corpus Linguistics. Volume 6, Issue 1. Pages 97-133.

Loock, R. (2016). *La traductologie de corpus. Editions Universitaires.* Traductologie. Septentrion. Paris. 261 pages.

Morin É., Daille B. (2010). *Compositionality and lexical alignment of multi-word terms*. Language Resources and Evaluation (LRE), Springer Netherlands, 44, 79–95.

Saldanha G., O'Brien S. (2013) *Research methodologies in translation studies.* London: Routledge, 2013.

Sinclair, J. (2005) *Corpus and Text − Basic Principles*. In: Wynne, Martin (ed.) Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books. Pages 1-16.