Constitution et traitement d'un corpus bilingue d'articles scientifiques : exemple de mise en oeuvre automatique avec une architecture légère en Perl

Olivier Kraif
Université Grenoble Alpes

Résumé
Nous présentons dans cet article un outil dédié à la constitution de corpus parallèles alignés constitués à partir de l'aspiration de sources sur le Web. Nous montrons comment cet outil a permis de constituer un corpus aligné anglais-français pour un type de texte difficile à trouver en version bilingue, à savoir les articles scientifiques. Moyennant l'élaboration de chaines de traitement basées sur l'écriture d'expressions régulières (dédiées à la définition des urls à aligner et à l'extraction des contenus) nous avons pu constituer, lors d'une première campagne, un corpus parallèle d'environ 4 millions de mots dans chaque langue, formaté en XML-TEI et TMX.
Mots-clés : alignement multilingue, corpus parallèle, corpus d'articles scientifiques
Keywords : multilingual aligning, parallel corpora, scientific paper corpora

Introduction

La constitution de corpus bilingues (ou multilingue) parallèles, c'est-à-dire réunissant des textes et leurs traductions dans une ou plusieurs langues, peut se révéler une tâche assez ardue. Pour un couple de langues tel que le français et l'anglais, sans doute un des mieux lotis dans ce domaine, on peut trouver de grandes quantités de textes traduits et disponibles au téléchargement, mais concernant un nombre limité de domaines et de types de textes : textes réglementaires et juridiques, retranscription de débats parlementaires, rapports institutionnels, documentations de logiciels libres, conférences ou sous-titre de films traduits de façon collaborative. Le site OPUS (Tiedmann, 2012), qui héberge sans doute la plus riche collection de textes alignés en ligne, fournit un bon aperçu de la très grande quantité de données traduites disponibles sur Internet, mais il montre aussi qu'en dehors des documents officiels produits par les institutions ou organisation nationales et internationales, et des productions collaboratives (logiciels libres, wiki, etc.), on trouve assez peu de choses, et de nombreux genres sont sous-représentés.
C'est notamment le cas de la production scientifique. S'il existe des corpus académiques multilingues téléchargeables ou interrogeables en ligne, tels que Scientext (Tutin et Grossmann, 2012) ou Kiap (Fløttum et al. ,2013), il n'existe pas, à notre connaissance, de tels corpus en version parallèle. Il existe plusieurs raisons à cela : d'une part, c'est un domaine ou l'on traduit assez peu, de nombreux chercheurs rédigeant directement leurs articles en anglais, ou adaptant eux-mêmes, en mélangeant traduction et réécriture, des textes déjà publiés dans leur langue d'origine. D'autre part, les traductions effectuées par des professionnels sont souvent commanditées par des éditeurs d'ouvrages ou de revues dont l'accès est payant, en version électronique ou papier.
Pourtant nous avons noté quelques exceptions, assez notables, dans le domaine des SHS. De nombreuses revues se proposent soient de fournir des éditions bilingues (ou multilingues) des articles qu'elles publient, et soit d'intégrer dans leurs contenu éditorial une sélection d'articles traduits d'autres revues : c'est le cas, par exemple, de *L'Année psychanalytique internationale* (en partenariat avec *The International Journal of Psychoanalysis*), de *ReS Futurae* (en partenariat avec *Science Fiction Studies*), de *Women, Gender, History* (en partenariat avec la revue française *Clio*). Sur le site http://journals.openedition.org, qui publie en texte intégral des revues dans les domaines des SHS, on trouve en version bilingue français-anglais de nombreuses revues. Pour n'en citer que quelques unes : la *Revue de géographie alpine*, *Champs pénal*, *Perspectives chinoises*, la *Revue internationale de politique de développement*, *Anthropologie et développement*, etc. On trouve même des revues traduites en de nombreuses langues, telle que *Via tourism* (avec des versions en français, anglais, espagnol, portugais, catalan, allemand).
Ces publications multilingues constituent un vrai trésor pour qui s'intéresse aux écrits académiques, du point de vue traductologique, terminologique ou encore pour étudier le lexique scientifique transdisciplinaire d'un point de vue contrastif (Drouin, 2010 ; Hatier 2015 ; Gilles 2017).

Nous présentons dans cette communication un outil permettant d'automatiser la collecte, l'alignement et les traitements de ces textes. Nous donnons ensuite les résultats d'une première campagne de collecte effectuée sur les revues du site http://journals.openedition.org.

Chaine de traitement

Il existe de nombreux outils, payants ou gratuits, dédiés à l'aspiration de site Web (tels que WebCopier, HTTrack) et à l'alignement de corpus parallèles (par exemple YouAlign, WinAlign, LF Aligner, Yasa ou encore Alinea). Mais de rares outils permettent d'intégrer les deux fonctionnalités, tels que le logiciel commercial AlignFactory de Terminotix. Pour pallier ce manque, nous avons choisi de développer une solution libre et légère (pour le moment sans interface graphique), écrite en Perl, permettant d'intégrer un *crawler* (pour sauvegarder des contenus extraits du web) à des chaines de traitements dédiées à la constitution de corpus, à la manière de Gromoteur (Gerdes 2014). Cet outil, baptisé PCP (pour Perl Corpus Processor), intègre des fonctionnalités standards des chaines de traitement de corpus :
- réencodage des caractères
- reformatage et extraction de texte brut
- application de recherche/remplacement d'expressions régulières en cascade (cascades de transducteurs)
- segmentation en paragraphes, phrases et tokens
- balisage structurel et re-formatage en XML-TEI
- intégration de Treetagger et d'analyseurs syntaxiques tels que XIP (à installer séparément)
- extraction de concordances, d'index hiérarchiques, de tableau de cooccurrences, de segments répétés

A ces fonctionnalités, il faut ajouter deux étapes utiles pour la collecte de textes alignés :
- le *crawler* qui parcourt le Web pour extraire des pages ou des portions de page web ;
- l'alignement, qui intègre plusieurs aligneurs libres : Yasa (Lamraoui & Langlais, 2013), LF Aligner basé sur Hunalign (Varga et al., 2005), JAM (Kraif, 2015) et Alinea (Kraif, 2001). Ces aligneurs et multialigneurs (qui peuvent aligner plus de deux langues) sont accompagnés de fonctionnalités de reformatage vers différents formats d'alignement (TXT, CesAlign, TMX, HTML, etc.).

L'originalité de PCP tient dans le parti-pris suivant : tous les paramétrages de l'outil sont basés sur les expressions régulières de Perl. Ces expressions permettent notamment de déterminer les urls des pages à parcourir, les portions de page à extraire, les méta-données à enregistrer, les expressions de nettoyage des pages, les correspondances entre pages alignées, mais aussi le nommage des fichiers et des répertoires, etc. Par exemple, pour enregistrer les articles issus de la revue *Via Tourism*, on définit les paramètres suivants :

```
url='http://journals.openedition.org/viatourism/'
urlPattern=qr/http:..journals.openedition.org.viatourism.\d+$/
contentPattern=qr/<meta   http-equiv="Content-language"   content="(?:fr|en)"   \/>.*<!--   #widgets
-->(.*)<\/div><!-- .text wResizable -->/s
namePattern=[qr/<title>(.*?)<\/title>/]
nameBase='content'
metadataPatterns=[
{label=>'authors',base=>'content',search=>qr/<meta name="citation_authors" content="(.*?)"-->/s},
{label=>'publicationDate',base=>'content',search=>qr/<meta name=".*?_pub.*?date" content="(.*?)"/s},
{label=>'date',base=>'content',search=>qr/<meta name="citation_online_date" content="(.*?)"/s},
{label=>'publisher',base=>'content',search=>qr/<meta name="citation_publisher" content="(.*?)"/s},
{label=>'volume',base=>'content',search=>qr/<meta name="citation_issue" content="(.*?)"/s},
{label=>'language',base=>'content',search=>qr/<meta name="citation_language" content="(.*?)"/s},
{label=>'languageId',base=>'content',search=>qr/<meta name="citation_language" content="(.*?)"/s},
{label=>'bibl',base=>'content',search=>qr/<div id="quotation" class="section">.*?<p>(.*?)<\/p>/s},
]
metadataFilePattern=[qr/$/,'.meta']
alignedUrlWithContextPatterns={en=>qr/<link     title=".*?"     type="text\/html"     rel="alternate"
hreflang="en" href="(.*?)"/}
```

L'outil s'adresse donc à des usagers (linguistes, terminologues, traducteurs) qui, sans avoir besoin de connaître un langage de programmation, doivent néanmoins avoir été initiés au langage des expressions régulières, et à quelques spécificités syntaxiques (utilisation des slashes, accolades, crochets, guillemets). La relative complexité du formalisme est selon nous largement compensée par sa souplesse et sa très grande puissance d'expressivité. Nos collaborations dans le domaine de la linguistique de corpus nous ont montré qu'une communauté potentielle d'utilisateurs existe, et pourrait encore s'accroitre moyennant le développement de formations ciblées.

Résultats de la première campagne

Un stagiaire de deuxième année de licence de Sciences du langage, Dorian Bellanger, a mis en place des chaines de traitement PCP sur 10 revues de OpenEdition.org, et a récolté 558 articles dans chaque langue pour un total d'environ 2 812 000 mots en anglais et 3 041 000 en français. A cela s'ajoutent 144 articles de la *Revue de recherche en psychanalyse* et 180 articles de vulgarisation issus du *Journal du CNRS* pour un total d'environ 4 millions de mots dans chaque langue. Les corpus ont été alignés en TMX et convertis en format XML-TEI (P5).
Dans de prochaine campagnes de téléchargement, nous prévoyons d'augmenter ce corpus, par l'ajout de nouvelles revues et l'élargissement à d'autres genres de texte, notamment les rapport scientifiques publiés par le service Communautaire d'Information sur la Recherche et le Développement (CORDIS) de la Commission Européenne.
PCP est librement téléchargeable depuis le site de l'auteur (http://olivier.kraif.u-grenoble3.fr/). Les chaines de traitement développées seront également prochainement mise en ligne sur le même site.

Références

Drouin, P. (2010). Extracting a bilingual transdisciplinary scientific lexicon. *eLexicography in the 21st century: new challenges, new applications*. Louvain-la-Neuve: Presses Universitaires de Louvain/Cahiers du CENTAL, 43-53.
Gilles, F. (2017) *Valorisation des analogies lexicales entre l'anglais et les langues romanes : étude prospective pour un dispositif plurilingue d'apprentissage du FLE dans le domaine de la santé*, Thèse de doctorat, sous la dir. de C. Degache et O. Kraif, Université Grenoble Alpes
Hatier, S. (2016) *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS*, Thèse de doctorat en Sciences du langage Spécialité Informatique et sciences du langage, sous la direction d'Agnès Tutin, Université Stendhal Grenoble 3.
Fløttum Kjersti, Dahl Trine, Alvsåker Didriksen Anders, Müller Gjesdal Anje (2013) KIAP – reflections on a complex corpus, In *Bergen Language and Linguistics Studies*, Vol 3, No 1, p. 137-150.
Kraif O. (2001) Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation, *TAL* 42 :3, ATALA, Paris, pp. 833-867.
Kraif O. (2015) Multi-alignement vs bi-alignement : à plusieurs, c'est mieux ! , *Actes de TALN 2015, 22ème Conférence sur le Traitement Automatique des Langues Naturelles*, Caen, 22-25 juin 2015, pp. 255-266.
Lamraoui, F., Langlais, P. (2013) "Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment?", *XIV Machine Translation Summit*, Nice, France.
Tiedemann, J. (2012) Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012).*
Gerdes, K. (2014) "Corpus collection and analysis for the linguistic layman: The Gromoteur", *Proceedings of the JADT 2014*, Paris.
Tutin, A., Grossmann, F. (2012) (Eds). *Autour du corpus Scientext*, Presses Universitaires de Rennes.
Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590-596.