# Constitution et traitement d'un corpus bilingue d'articles scientifiques

## Exemple de mise en oeuvre automatique avec une architecture légère en Perl

Olivier Kraif

LIDILEM, Univ. Grenoble Alpes, F-38040, Grenoble ; olivier.kraif@univ-grenoble-alpes.fr

**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 chaînes 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

## 1 Introduction

La constitution de corpus bilingues (ou multilingues) 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 organisations 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 où 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 commandées 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 proposent soit de fournir des éditions bilingues (ou multilingues) des articles qu'elles publient, soit d'intégrer dans leurs contenus éditoriaux 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*, *Champ 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 cet article 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, ainsi que sur d'autres périodiques tels que le Journal du CNRS.

## 2 Outils existants

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.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Nous nous pencherons ici uniquement sur les outils non commerciaux. Pour l'alignement de textes, Lamraoui et Langlais (2013) reportent des résultats au niveau de l'état de l'art [1] avec le simple et très rapide Yasa, implémenté en C++. Le code source est téléchargeable sur le site des auteurs, mais nécessite d'être compilé. Doté d'une interface graphique et prenant en charge de nombreux formats d'entrées / sorties, nous recommendons également LF Aligner de Andreas Farkas, basé sur l'aligneur Hunalign (Varga *et al.* 2005). L'utilisation de lexiques bilingues pour de nombreuses langues lui confère une certaine robustesse et il obtient de bons résultats avec de nombreuses paires de langues. Notons enfin le multi-aligneur JAM (Kraif, 2015), conçu pour aligner plus de deux langues en même temps. Ces différents aligneurs peuvent être testés en ligne grâce à une interface Web nommée Web Align Toolkit, conçue par nos soins (accessible ici : http ://phraseotext.univ-grenoble-alpes.fr/WebAlignToolkit/). Cette interface utilise en fait la boîte à outils PCP Toolkit que nous décrirons par la suite.

Pour l'aspiration de pages Web, nous citerons Gromoteur (Gerdes, 2014), un outil tout spécialement conçu pour l'aspiration de pages Web en vue de constituer un corpus linguistique. Comme la plupart des outils d'aspiration de page Web, Gromoteur permet de gérer des sessions d'aspiration (« crawling ») s'appuyant sur la définition de quelques paramètres de base :

— url(s) de départ : à partir de quelle adresse(s) Web le crawling doit-il commencer ?
— pattern des url(s) à suivre et à télécharger : on définit une expression régulière indiquant la forme des urls qu'il faut parcourir et enregistrer ;
— forme du contenu : on peut également contraindre le crawling en fonction du contenu des pages et non seulement de leur adresse ;
— pattern des urls à ne pas parcourir ;
— langue des pages à parcourir.

La gestion des sessions de crawling permet de définir s'il faut écraser le contenu des urls déjà téléchargées, le nombre maximal de pages à traiter, la taille maximale des données à télécharger, etc.

Gromoteur présente l'intérêt d'intégrer une chaîne de traitement dédiée à la linguistique de corpus : outre le nettoyage des contenus téléchargés, il permet d'effectuer la tokenisation en mots, l'étiquetage et la lemmatisation, ainsi que l'extraction de concordances et de données lexicométriques.

Mais Gromoteur n'est pas conçu, dans sa version actuelle, pour travailler sur des corpus parallèles. Pour l'aspiration de pages Web en vue de l'alignement multilingue, nous ne connaissons pas de solution non-commerciale. La boîte à outil que nous présentons dans la prochaine section vise à combler cette lacune.

1. Ils ont effectué une évaluation comparée de Yasa, BMA (Moore, 2002) et HUNALIGN (Varga *et al.*, 2005) sur différentes parties du corpus BAF. En calculant la précision et le rappel au niveau des phrases, leur système obtient de meilleurs résultats sur toutes les parties du corpus : 96,4 % de F-mesure pour les textes institutionnels, 91,7 % pour les articles scientifiques, 86,9 % pour une traduction contractée de Jules Verne et 10,9 % pour un glossaire non parallèle.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

## 3 La boîte à outil PCP Toolkit

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 chaînes de traitements dédiées à la constitution de corpus, à la manière de Gromoteur. Cet outil, baptisé PCP (pour Perl Corpus Processor), intègre des fonctionnalités standards des chaînes de traitement de corpus :

— réencodage des caractères ;
— reformatage et extraction de texte brut ;
— application de recherche/remplacement d'expressions régulières en cascade ;
— segmentation en paragraphes, phrases et tokens ;
— balisage structurel et re-formatage en XML-TEI ;
— intégration de Treetagger et d'analyseurs syntaxiques tels que Stanford CoreNLP ou XIP (à installer séparément).
— extraction de concordances, d'index hiérarchiques, de tableaux de cooccurrences, de segments répétés ;
— recherche de patterns complexes.

A ces fonctionnalités, il faut ajouter deux étapes spécifiques à 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 Toolkit tient dans le parti-pris suivant : tous les paramétrages de l'outil sont basés sur les expressions régulières (suivant la syntaxe 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. De bout en bout, il est possible de réaliser une chaîne de traitements complexe pouvant mettre en jeu des milliers de fichiers sans pour autant connaître un langage de programmation. L'outil s'adresse donc à des « spécialistes de la donnée », plutôt qu'à des spécialistes de l'algorithmique.

En tant que tels, les utilisateurs doivent être familiers des notions de noms de fichiers et de répertoires, d'encodage des caractères et de format de fichier, et connaître des formats utiles en TAL tels que TXT, CSV ou TSV, HTML, XML, XML-TEI, ou encore TMX et SRX pour les mémoires de traduction.

L'aspect sériel d'une chaîne de traitement est déterminé non pas par des structures algorithmiques (tels qu'une boucle ou des structures conditionnelles) mais 1/ par des expressions régulières qui indiquent l'ensemble des données à traiter, et 2/ par des

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

étapes de traitement séquentielles qui indiquent, pas à pas, quelles sont les opérations effectuées sur ces données.

### 3.1 Un exemple de chaîne de traitements

Pour illustrer concrètement le fonctionnement de PCP Toolkit, décomposons une des chaînes de traitements que nous avons mise en oeuvre, pour l'aspiration et l'alignement de la revue multilingue *Via Tourism*.

Les étapes de la chaîne de traitements (ou *pipeline*) sont les suivantes :

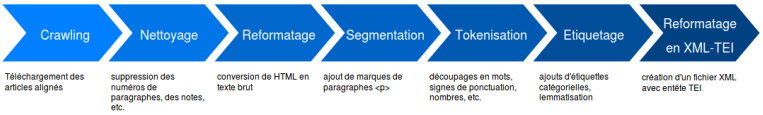



FIGURE 1: Etapes d'une chaîne de traitements

Pour l'utilisateur, un *pipeline* est simplement défini par un fichier portant une extension.pl, où il explicite la suite des paramétrages et des étapes (nommées *fonctions*) qu'il veut mettre en oeuvre. L'exécution d'un pipeline nommé myPipe.pl est effectué en ligne de commande, en écrivant :

perl runPipeline.pl myPipe.pl

### 3.2 Crawling

Pour enregistrer les articles issus de la revue *Via Tourism*, nous avons défini les paramètres suivants :

— l'url de départ :

url='http ://journals.openedition.org/viatourism/'

— la langue des pages sources :

sourceLanguage='fr'

— le pattern des url à parcourir :

urlPattern=qr/http :.journals.openedition.org.viatourism.\d+$/

— le pattern des contenus à enregistrer (tout le texte reconnu par l'expression entre parenthèses) :

contentPattern=qr/(.*?)<\/title>/]

— la liste des patterns indiquant où trouver les métadonnées du fichier ainsi enregistré. Pour chaque méta-donnée, on définit d'abord une étiquette ou label. Tout comme le nom du fichier, les méta-données peuvent être extraites de différentes bases : l'url, le contenu de la page, le lien ayant conduit vers la page. Une fois la base définie, on définit le pattern permettant d'extraire l'information correspondante (définie par l'expression entre parenthèses).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

```
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},
]
```

— le schéma de recherche/remplacement qui indique comment le nom du fichier de métadonnées doit être formé, par rapport au nom du fichier de données. Ici, on se contente d'ajouter.meta à la fin du nom de fichier.

metadataFilePattern=[qr/$/,'.meta']

— le pattern indiquant le contexte de l'url pointant vers une page traduite, ainsi que la langue de la page traduire (ici'en').

alignedUrlWithContextPatterns={en=>qr/[qr/[.]htm/,'a.htm']}

Pour la définition des pages alignées, il existe en fait plusieurs stratégies, qui correspondent à divers cas de figure que nous avons rencontrés.

### 3.3 Stratégies d'appariement des pages

— Article contenant l'hyperlien

Le premier cas de figure correspond au cas le plus simple et le plus fréquent : la page contenant l'article en langue source contient également un hyperlien pointant vers la traduction. C'est le cas correspondant à la revue Via Tourism, dont chaque article inclus des liens vers les diverses traductions disponibles.

— Transformation d'URL

Le deuxième cas de figure consiste à opérer une transformation simple de l'URL pour obtenir l'URL de l'article traduit. On peut donner l'exemple de la Revue de recherche en psychanalyse. Pour obtenir la traduction anglaise des articles, il suffit d'intercaler un 'a' dans l'URL, comme ci-dessous :

version fr : https : //www.cairn.info/revue-recherches-en-psychanalyse-2015-2-page-98.htm

version en : https : //www.cairn.info/revue-recherches-en-psychanalyse-2015-2-page-98a.htm

Pour traiter un tel cas de figure, on utilise le paramétrage suivant :

alignedUrlTransformationScheme={'en'=>[qr/[.]htm/,'a.htm']}

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

De nombreux sites se basent sur ce modèle. Par exemple, pour les rapports publiés sur le site du parlement européen, on a des fichiers au formats DOCX donc les adresses sont les suivantes :

version fr : http ://www.europarl.europa.eu/doceo/document/A-8-2019-0069*FR.docx

version en : http ://www.europarl.europa.eu/doceo/document/A-8-2019-0069*EN.docx

Pour traiter ce cas on aurait :

alignedUrlTransformationScheme={'en'=>[qr/FR[.]htm/,'EN.htm']}

— Page d'index

Le troisième cas de figure est plus rare : il s'agit de liens vers des traductions apparaissant à côté du lien vers l'article source, dans un index. Dans certains cas, en effet, la page contenant l'article source ne contient pas de lien direct vers sa traduction. C'est l'index des articles du numéro qui permet d'effectuer les appariements (le cas de figure avait été rencontré pour la revue *Clio*).

D'autres paramétrages que nous ne détaillerons pas sont proposés (pour gérer les limites de la session de crawling, le nombre de pages enregistrés, l'encodage des fichiers, etc.). La totalité de la chaîne de traitement est donnée en annexe.

Une fois tous ces paramétrages effectués, l'étape proprement dite est lancée *via* la commande :

->runCrawler()

Chaque étape est définie par une telle commande préfixée par le symbole ->

### 3.4 Suite des traitements

Toutes les étapes qui suivent sont définies par une série de paramètres standards, de la manière suivante :

— définition du répertoire contenant les fichiers en entrées : paramètre inputDir
— définition du répertoire contenant les fichiers en sorties : paramètre outputDir
— définition des fichiers à traiter dans le répertoire d'entrées : paramètre filePattern
— définition du nom des fichiers en sorties, formés par transformation du nom des fichiers en entrées : outputFileName
— définition d'une série de paramètres spécifiques à l'étape
— lancement de l'étape : ->fonction()

Par exemple, pour le nettoyage du HTML des pages téléchargées de *via* Tourism, on définit l'étape suivante :

inputDir='./data/viaTourism/html'
outputDir='./data/viaTourism/html2'
filePattern=qr/html$/
outputFileName=[qr/[.]html/,'.html']

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

```
searchReplacePatterns=[
[qr/(\d*?<\/span>)/,'$1'],
[qr/.?<|/p>|s<\/div>/s,''],
[qr/.?<|/a><|/div>|s<\/div>/s,''],
[qr/.*?<\/span>/,''],
[qr/.*?<\/ul>/s,''],
[qr//,'']
]
->findAndReplace()
```

Dans sa version actuelle, sans interface graphique, les utilisateurs de PCP Toolkit doivent donc être initiés non seulement aux expressions régulières, mais aussi à quelques spécificités syntaxiques liées à l'expression de ces paramètres : chaînes de caractères entre'guillemets simples', expressions régulières entre qr/ /, listes entre crochets, accolades, balises HTML, etc. Ce type d'outils s'adresse donc à des utilisateurs avertis, qui sans avoir des compétences de développeurs, doivent avoir une certaine maîtrise de l'expression des données : l'outil peut intéresser des linguistes, terminologues, traducteurs à l'aise avec les différents types de structurations des données langagières (XML, CSV, JSON, etc.).

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'accroître moyennant le développement de formations ciblées.

## 4 Résultats d'une campagne de téléchargement

Un stagiaire de deuxième année de licence de Sciences du langage, Dorian Bellanger, a mis en place des chaînes 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.

Les revues concernées couvrent des domaines variés :

- Architecture : *Architecture Beyond Europe*
- Sciences sociales et interdisciplinarité : *Champ pénal, Environnement urbains, Brussel studies, Perspectives interdisciplinaires sur le travail et la santé (PISTES)*
- Sciences de l'information et de la communication : *Communiquer*
- Géographie : *Revue de géographie alpine*
- Sémiotique : *Signata / annales des sémiotiques*
- Politique de développement : *International Development Policy*
- Tourisme : *Via Tourism*
- Histoire : *Témoigner*

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

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

dans chaque langue. Les corpus ont été alignés en TMX et convertis en format XML-TEI (P5). Le corpus est réparti en trois parties, en fonction des sources éditoriales, de la manière suivante :

Tableau 1 : Résultat de la première campagne de téléchargement

| | OpenEdions.org | | Le Journal du CNRS | | Revue de Rech. Psych. | | TOTAL | |
|---|---|---|---|---|---|---|---|---|
| | en | fr | en | fr | en | fr | en | fr |
| Nombre d'articles | 558 | 558 | 180 | 180 | 144 | 144 | 882 | 882 |
| Nombres de mots | 2 812 363 | 3 041 326 | 234 286 | 294 368 | 759 086 | 750 883 | 3 805 735 | 4 086 577 |

Dans de prochaines 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 rapports scientifiques publiés par le service Communautaire d'Information sur la Recherche et le Développement (CORDIS) de la Commission Européenne.

Les corpus ne sont pas diffusables par nous, du fait de la protection du droit d'auteur, mais les chercheurs peuvent se procurer PCP et les pipelines utilisés sur la page Github du projet : https ://github.com/kraifo/PCP

## 5 Conclusion et perspectives

L'étudiant stagiaire à qui nous avons demandé de prendre en main PCP Toolkit a réussi à mettre en œuvre rapidement les chaînes de traitement requises pour la constitution d'un corpus bilingue aligné d'articles scientifiques d'environ 4 millions de mots par langue, en anglais et français. Nous pensons que des linguistes familiarisés aux outils du TAL, aux formats de fichier usuels, ainsi qu'au formalisme des expressions régulières, pourront également tirer profit d'un tel outil, dont la licence est libre.

Il nous reste à réaliser un effort important de documentation, à travers la confection de tutoriels pour guider les usagers dans la réalisation de chaînes de traitement simples et le choix des paramètres appropriés, notamment à partir de textes téléchargés depuis Internet. Nous prévoyons par ailleurs de réaliser prochainement une interface Web pour générer les fichiers *pipelines* à partir d'un formulaire en ligne, avec un assistant permettant de guider l'utilisateur pas à pas : de la sorte, les aspects syntaxiques liés au langage Perl (chaîne de caractères entre guillemets, expressions régulières entre slashes, tableaux associatifs avec accolades, etc.) pourront être ignorés dans un premier temps. Une fois ces fichiers générés, les utilisateurs pourront les exécuter sur un petit échantillon de données, ou les télécharger avant de les faire tourner sur leurs propres machines avec des volumes de données plus importants.

**Remerciements**

Nous remercions chaleureusement Dorian Bellanger, étudiant en licence de l'Université Grenoble Alpes, et Brooke Stephenson, étudiante en master de l'Université de Franche-Comté, qui ont travaillé lors de leurs stages sur les chaînes de traitements pour la constitution de ce corpus.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

# Références

Drouin, Patrick (2010) "Extracting a bilingual transdisciplinary scientific lexicon", *Lexicography in the 21st century : new challenges, new applications*. Louvain-la-Neuve : Presses Universitaires de Louvain/Cahiers du CENTAL, p. 43-53.

Fløttum, Kjersti ; Dahl, Trine ; Alvsåker Didriksen, Anders ; Müller Gjesdal, Anje (2013) "KIAP – reflections on a complex corpus", *Bergen Language and Linguistics Studies*, vol. 3, No 1, p. 137-150.

Gerdes, Kim (2014) "Corpus collection and analysis for the linguistic layman : The Gromoteur", *Proceedings of the JADT 2014*, Paris, p. 261-269.

Gilles, Fabrice (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, Sylvain (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 et Marie-Paule Jacques, Université Stendhal Grenoble 3.

Kraif, Olivier (2001) « Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation » , *TAL* 42 :3, ATALA, Paris, p. 833-867.

Kraif, Olivier (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, p. 255-266.

Lamraoui, Fethi ; Langlais, Philippe (2013) "Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment ?", *XIV Machine Translation Summit*, Nice, France, p. 77-84.

Moore, Robert C. (2002) "Fast and accurate sentencealignment of bilingual corpora", *AMTA'02 Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation : From Research to Real Users*, p. 135-144.

Tiedemann, Jörg (2012) "Parallel Data, Tools and Interfaces in OPUS", *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, MAY 21-27, Istanbul, Turkey_,_ p. 2214-2218.

Tutin, Agnès ; Grossmann, Francis (dir.) (2012) *Autour du corpus Scientext*, Presses Universitaires de Rennes.

Varga, Dániel ; Németh, László ; Halácsy, Péter ; Kornai, András ; Trón, Viktor ; Nagy, Viktor (2005) "Parallel corpora for medium density langages", *Proceedings of the RANLP 2005*, p. 590-596.

# Annexe

## Chaîne de traitement utilisée pour la revue *via* Tourism

```
windows=0

verbose=1

#****************************** Step 1 : crawling

sourceId='viaTourism'

targetDir="./data/viaTourism/html"

url='http ://journals.openedition.org/viatourism/'

urlPattern=qr/http :.journals.openedition.org.viatourism.\d+$/

resetCrawling=1

limitDepth=3

downloadIfAligned=1

alignedUrlWithContextPatterns={en=>qr/(.*?)<\/title>/]

nameBase='content'
```

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

```
inputEncoding='utf8'
outputEncoding='utf8'
metadataPatterns=[

 {label=>'authors',base=>'content',search=>qr/<meta name="citation_authors" content="(.*?)"-->/s},
 {label=>'publicationDate',base=>'content',search=>qr/<meta name="citation_publication_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/(.*)/,'$1.meta']
maxRecordedPages=50000
sourceLanguage="fr"
->addSource()
->runCrawler()
#****************************** Step 2 :cleaning boilerplate html
inputDir="./data/viaTourism/html"
outputDir="./data/viaTourism/html2"
filePattern=qr/html$/
outputFileName=[qr/[.]html/,'.html']
searchReplacePatterns=[
[qr/.*?<\/span><\/p><\/div>/""],
#[,]
]
->findAndReplace({overwriteOutput=>1})
#****************************** Step 3 :converting html to text
inputDir="./data/viaTourism/html"
outputDir="./data/viaTourism/txt"
filePattern=qr/html$/
outputFileName=[qr/[.]html/,'.txt']
fileEncoding="utf8"
->html2txt({overwriteOutput=>1})
->next()
#****************************** Step 4 :adding tags
outputDir="./data/viaTourism/para"
filePattern=qr/txt$/
outputFileName=[qr/(.*)[.]txt$/i,'$1.txt']
noEmptyPara=1
escapeMeta2entities=1
```

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

```
->addParaTag({overwriteOutput=>1})
->next()
#****************************** step 5 :tokenize raw txt files
outputDir="./data/viaTourism/tok"
language='fr'
filePattern=qr/fr[.]txt$/
outputFileName=[qr/(.*)[.]txt$/,'$1.tok']
newLineTag="
->tokenize({overwriteOutput=>1})
language='en'
filePattern=qr/en[.]txt$/
outputFileName=[qr/(.*)[.]txt$/,'$1.tok']
newLineTag="
->tokenize({overwriteOutput=>1})
->next()
# ****************************** tagging files
filePattern=qr/(.*).fr.tok$/
treetaggerLanguage="french-utf8"
outputDir="./data/viaTourism/ttg"
outputFileName=[qr/(.*).tok$/i,'$1.ttg']
tokenize=0
supprSpcTag=1
sentMark="SENT"
addSentTag=1
->runTreetagger({overwriteOutput=>1})
filePattern=qr/(.*).en.tok$/
treetaggerLanguage="english-utf8"
->runTreetagger({overwriteOutput=>1})
->next()
# ****************************** converting to TEI
template='./grm/tei.revue.tpl'
filePattern=qr/(.*).ttg$/
outputDir="./data/viaTourism/tei"
outputFileName=[qr/(.*).ttg$/i,'$1.xml']
dataFilePattern=[qr/ttg\/(.*extract.\w\w)[.]ttg/,'html/$1.html.meta']
data={
'h.title'=>"Corpus Bilingue via Tourism",
respType=>"Téléchargé en parallèle par",
respName=>"Olivier Kraif & Dorian Bellanger",
```

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

```
pubDate=>“2018”,
samplingDecl=>“full paper”,
segmentation=>“paragraph and sentence level”,
}
->applyTemplate({overwriteOutput=>1})
# ****************************** extracting stats
inputDir='./data/viaTourism/ttg'
filePattern=qr/(.*)fr.ttg$/
outputDir='./data/viaTourism/stats.fr'
inputFormat=“ttg”
outputFormat=“csv”
global=1
->anaText({overwriteOutput=>1})
filePattern=qr/(.*)en.ttg$/
outputDir='./data/viaTourism/stats.en'
inputFormat=“ttg”
outputFormat=“csv”
global=1
->anaText({overwriteOutput=>1})
```

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454