UNIVERSITÉ Grenoble Alpes

Journées LTT 2018
28 septembre

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

exemple de mise en oeuvre automatique avec une architecture légère en Perl

Olivier Kraif

Introduction

- Un constat
  - les **corpus parallèles alignés** connaissent une surreprésentation dans certains genres et domaines :
    - institutions nationales et internationales, ONG : textes de loi, réglementations, comptes-rendus de débats, rapports
    - traductions collaboratives : logiciels libres, Wiki, sous-titres, conférences, sites d'actualité, p.ex. *Project Syndicate* (Opus : http://opus.nlpl.eu/)
    - littérature : essentiellement avant le XXe http://www.farkastranslations.com/bilingual_books.php)

Introduction

- Littérature scientifique
  - des corpus mono/multilingues mais non parallèles :
    - MENELAS (Zweigenbaum et al., 1995)
    - Scientext (Tutin et Grossmann, 2012)
    - Kiap (Fløttum et al., 2013)
  - des besoins importants en traduction, terminologie, didactique

Introduction

- Pourquoi si peu de textes scientifiques parallèles ?
  - beaucoup de publications en anglais
  - les auteurs pratiquent l'auto-traduction / réécriture
  - peu de traductions libres de droit et/ou accessibles sur le Web
- Une exception : dans le domaine des SHS, de nombreuses revues font le choix du plurilinguisme
  - https://journals.openedition.org/

Introduction

- Partenariats entre revues
  - *L'Année psychanalytique internationale* (*The International Journal of Psychoanalysis*)
  - *ReS Futurae* (*Science Fiction Studies*)
  - *Women, Gender, History* (*Clio*)

- Revues bilingues ou plurilignues
  - *Revue de géographie alpine*
  - *Champs pénal*
  - *Perspectives chinoises*
  - *Revue internationale de politique de développement*
  - *Anthropologie et développement*
  - *Via tourism*
  - *...*

Introduction

- Comment tirer parti de ces ressources en lignes ?
- Comment constituer des corpus parallèles par *crawling* du Web ? (pour copie privée ou usage scientifique sans reproduction/diffusion)

Quels outils ?

Outils

- Aspiration de sites web
  - Web copier (shareware)
  - Httrack (libre)
- Aligneurs
  - You Align (Terminotix)
  - WinAlign (Trados)
  - LF Aligner (libre), Yasa (libre), Alinea (libre)
  - Interface multimoteurs : http://turing3.u-grenoble3.fr/webAlignToolkit/
- Aspiration + chaine de traitement de corpus
  - AlignFactory (Terminotix)
  - Gromoteur (libre)

Outils

- Ce qui manque :
  - outil libre
  - intégrant aspiration et alignement
  - souple et paramétrable

Outils

- Spécificités d'un aspirateur « parallèle » :
  - associer les textes et leur traduction
  - adopter des conventions de nommage pour l'association des fichiers :

    p.ex. :

    `article58.fr.html`
    `article58.en.html`

Outils

- Différentes stratégies d'association :

1) présence d'un lien vers une traduction depuis la page de l'article original

2) transformation d'url :

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

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

3) page d'index associant les deux URLs

Via
TOURISM REVIEW

La fabrique des concepts touristiques

Jean-Michel Decroly et Anya Diekmann

Traduction(s) :

The production of tourism concepts [en]
La fábrica de los conceptos turísticos [es]
La fabbricazione dei concetti turistici [it]
A fabricação dos conceitos turísticos [pt]
La fàbrica de conceptes turístics [ca]
Die Entstehung von touristischen Konzepten [de]

Plan | Texte | Bibliographie | Notes | Citation | Auteurs

Plan

Vous consultez

Traumas et catastrophe aujourd'hui

Éditorial

par **Christian Hoffmann**

Psychanalyste

…

Raccourcis

Plan de l'article →
Citer cet article →
Sommaire du numéro →

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

Recherches en psychanalyse
2015/2 (n° 20)

Pages : 98
Affiliation : Revue précédemment éditée par L'Esprit du temps

DOI : 10.3917/rep.020.0098
Éditeur : Association Recherches en psychanalyse

À propos de cette revue →
Site de la revue →

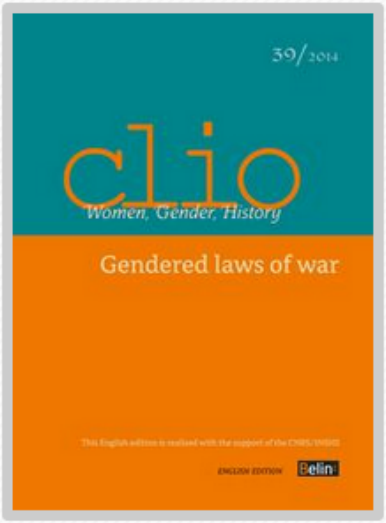
Alertes e-mail

Notre actualité évoque régulièrement de nouvelles catastrophes, et le mot traumatisme fait désormais partie de notre vocabulaire ordinaire. Comme le remarquait déjà Freud, il n'est plus possible à l'homme contemporain de s'appuyer sur une vision du monde, faisant récit et produisant du sens dans un espace et un temps perçus comme homogènes. 1

I - Problématique contemporaine

Traditionnellement la catastrophe était le dénouement d'un drame. Aujourd'hui les catastrophes viennent à un rythme de plus en plus rapide trouer les écrans de notre actualité, sans que nous puissions facilement les intégrer sous un sens 2

PDF



CLIO

No. 39, 2014/1

Gendered Laws of War

PAGES : 336

PUBLISHER : Belin

LIST OF ISSUES

PREVIOUS ISSUE **NEXT ISSUE**

TABLE OF CONTENTS

Outils

- Comment capturer ces associations ?

French
English : Free

Outils

- Avec des expressions régulières !

/<a href="(?<href1>.*?)" class="button-blue2 w49 right">French<.a>
<a href="(?<href2>.*?)"/

Le toolkit PCP

Perl Corpus Processor (PCP)

- Outil dédié à la mise en place de **chaines de traitement** sur des corpus de textes :
  - 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 XIP (à installer séparément)
  - extraction de concordances, d'index hiérarchiques, de tableau de cooccurrences, de segments répétés
  - Crawling de sites web
  - Alignement (Yasa, Alinea, Jam)
- Téléchargeable depuis ma page personnelle

Perl Corpus Processor (PCP)

- Ecriture des pipelines :
  - pas de programmation
  - maîtrise des expressions régulières
    - nommage des fichiers et répertoires
    - définition du crawling
    - rechercher / remplacer
  - connaissance sommaire des formats de fichier : txt, xml, html, csv
  - connaissance sommaire de la ligne de commande :
    - `perl runPipeline.pl nomDuPipeline.pl`

Perl Corpus Processor (PCP)

- Exemple de pipeline :

```
sourceId='viaTourism'
targetDir="./data/viaTourism/html"
url='http://journals.openedition.org/viatourism/'
urlPattern=qr/http:..journals.openedition.org.viatourism.\d+$/
sourceLanguage="fr"
downloadIfAligned=1
alignedUrlWithContextPatterns={en=>qr/<link title=".*?" type="text\/html" rel="alternate" hreflang="en" href="(.*?)"/}
contentPattern=qr/<meta http-equiv="Content-language" content="(?:fr|en)" \/>.*<!-- #widgets -->(.*)<\/div><!-- .text wResizable -->/s
namePattern=[qr/<title>(.*?)<\/title>/]
nameBase='content'
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']

->addSource()
->runCrawler()
```

Constitution d’un corpus parallèle

Perl Corpus Processor (PCP)

- Stage de 2 mois (mars-avril 2017)
- Etudiant de L2 Sciences du langage (option « Métier des humanités numériques »)
- Prise en main rapide de PCP

Perl Corpus Processor (PCP)

- Revues téléchargées :
  - OpenEdition.org : Architecture Beyond Europe, Brussel Studies, Champ Pénal, Communiquer, Environnement urbain, PISTES, Revue Internationale de Politique de Développement, Revue de Géographie Alpine, Signata, Témoigner, ViaTourism
  - Cairn : Revue de recherche en psychanalyse
  - Journal du CNRS (vulgarisation)
  - Domaines : architecture, info-com, géographie, psychanalyse, sciences politiques, sociologie, tourisme, urbanisme...

Perl Corpus Processor (PCP)

- Bilan :
  - OpenEdition.org : 558 articles
  - Cairn : 144 articles
  - Journal du CNRS : 180 articles
  - Environ 4 000 000 de mots dans chaque langue
  - Formats : XML-TEI + étiquetage, TMX

Perspectives

Perspectives

- Le public d'un tel outil existe !
  - spécialistes de la linguistique de corpus
  - ingénieurs linguistes
  - linguistes-talistes
- Reste à faire un gros travail :
  - de documentation
  - de formation (tutoriels)
- En chantier :
  - interface web permettant la construction guidée des pipelines (assistant)

/(Merci|Thanks) !/