Traitement automatique des langues et instrumentation du multilinguisme

Pierre Zweigenbaum

LIMSI, CNRS, Orsay
http://www.limsi.fr/~pz/

ERTIM, INALCO, Paris

Journée PIRSTEC *Informatique multilingue*
6 octobre 2009

Introduction

- Multilinguisme
  - Besoins en traduction, recherche translingue, etc.
  - Besoins en ressources langagières diverses
    - lexiques, terminologies, corpus multilingues
- Traitement automatique des langues
  - Outils informatiques : instrumenter certaines de ces tâches
    - aide à la production de ressources lexicales et terminologiques multilingues
    - traduction automatique
  - Besoin pour ce faire de ressources similaires
- Panorama de méthodes d'aide à la constitution de ressources multilingues
  - Constitution de corpus parallèles et comparables
  - Production ou extension de lexiques ou terminologies bilingues

Introduction

- Multilinguisme
  - Besoins en traduction, recherche translingue, etc.
  - Besoins en ressources langagières diverses
    - lexiques, terminologies, corpus multilingues
- Traitement automatique des langues
  - Outils informatiques : instrumenter certaines de ces tâches
    - aide à la production de ressources lexicales et terminologiques multilingues
    - traduction automatique
  - Besoin pour ce faire de ressources similaires
- Panorama de méthodes d'aide à la constitution de ressources multilingues
  - Constitution de corpus parallèles et comparables
  - Production ou extension de lexiques ou terminologies bilingues

Introduction

- Multilinguisme
  - Besoins en traduction, recherche translingue, etc.
  - Besoins en ressources langagières diverses
    - lexiques, terminologies, corpus multilingues
- Traitement automatique des langues
  - Outils informatiques : instrumenter certaines de ces tâches
    - aide à la production de ressources lexicales et terminologiques multilingues
    - traduction automatique
  - Besoin pour ce faire de ressources similaires
- Panorama de méthodes d'aide à la constitution de ressources multilingues
  - Constitution de corpus parallèles et comparables
  - Production ou extension de lexiques ou terminologies bilingues

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources

2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables

3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources

2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables

3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

Traitements informatiques du multilinguisme

- Détection de la langue d'un document (d'un segment de texte) (...)
- Traduction automatique (Google, Reverso)
  - Y compris traduction de parole (Quæro)
- Recherche d'information translingue (Google, HON, CISMeF)
  - $\supset$ recherche translingue de réponses à des questions
  - $\sim$ catégorisation de textes translingue
- Environnements d'aide à la traduction humaine
  - Accès à des lexiques et thésaurus multilingues (Alexandria)
  - Mémoires de traduction (Trados...)
  - Suggestion de traductions (Sharoff)
  - Traduction collaborative (Jibiki / Lydia)

Ressources langagières

Lexiques et terminologies multilingues

- (Grand Dictionnaire Terminologique)

Bases de phrases et textes traduits

- corpus multilingues parallèles (Europarl)

Bases de textes similaires dans une autre langue

- corpus comparables
  - synchrones : journaux d'une période donnée (NYT / Le Monde aujourd'hui)
  - documents sur un même thème, etc.

Constitution automatisée de ressources

Un schéma productif

1. Entrée : Ressources disponibles
   -
2. Traitement automatisé
   -
3. Sortie : Ressources nouvelles
   -

Constitution automatisée de ressources

Un schéma productif

1. Entrée : Ressources disponibles
   - Corpus parallèle
2. Traitement automatisé
   - Alignement automatique de phrases et de mots
3. Sortie : Ressources nouvelles
   - Lexique bilingue

Constitution automatisée de ressources

Un schéma productif

1. Entrée : Ressources disponibles
   - Corpus parallèle et analyseur syntaxique source
2. Traitement automatisé
   - Alignement automatique de phrases, mots et arbres
3. Sortie : Ressources nouvelles
   - Analyseur syntaxique cible

Constitution automatisée de ressources

Un schéma productif

1. Entrée : Ressources disponibles
   - Liste de mots source, lexique bilingue partiel
2. Traitement automatisé
   - Traducteur par apprentissage / par analogie
3. Sortie : Ressources nouvelles
   - Lexique bilingue plus complet

Constitution automatisée de ressources

Un schéma productif

1. Entrée : Ressources disponibles
   - ...
2. Traitement automatisé
   - ...
3. Sortie : Ressources nouvelles
   - ...

Construction automatique de ressources multilingues : Un exemple

Un exemple prototypique de tâche

- Constitution de lexiques ou terminologies bilingues
- Par alignement
- À partir de corpus bilingues (parallèles, comparables)

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources

2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables

3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

Corpus

- Un ensemble de textes sélectionnés possédant des caractéristiques contrôlées :

Sinclair (1996)

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language

Corpus parallèles et comparables

Degrés de parallélisme

Corpus parallèles : fort parallélisme

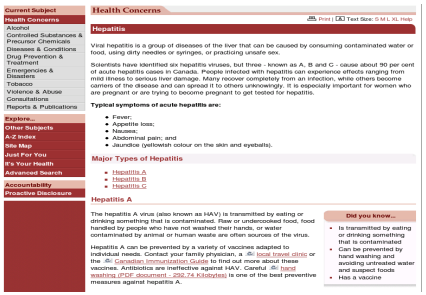
Corpus comparables : faible parallélisme

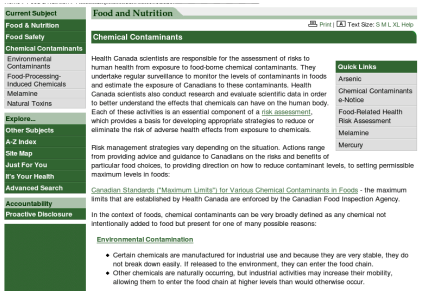
D'autres caractéristiques sont également mises en jeu (voir plus bas)

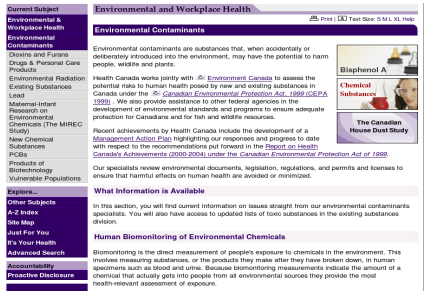
Corpus parallèle

Un corpus de textes en relation de traduction

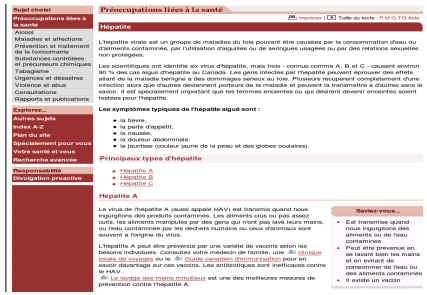
Langue A (*EN*)







Langue B (*FR*)

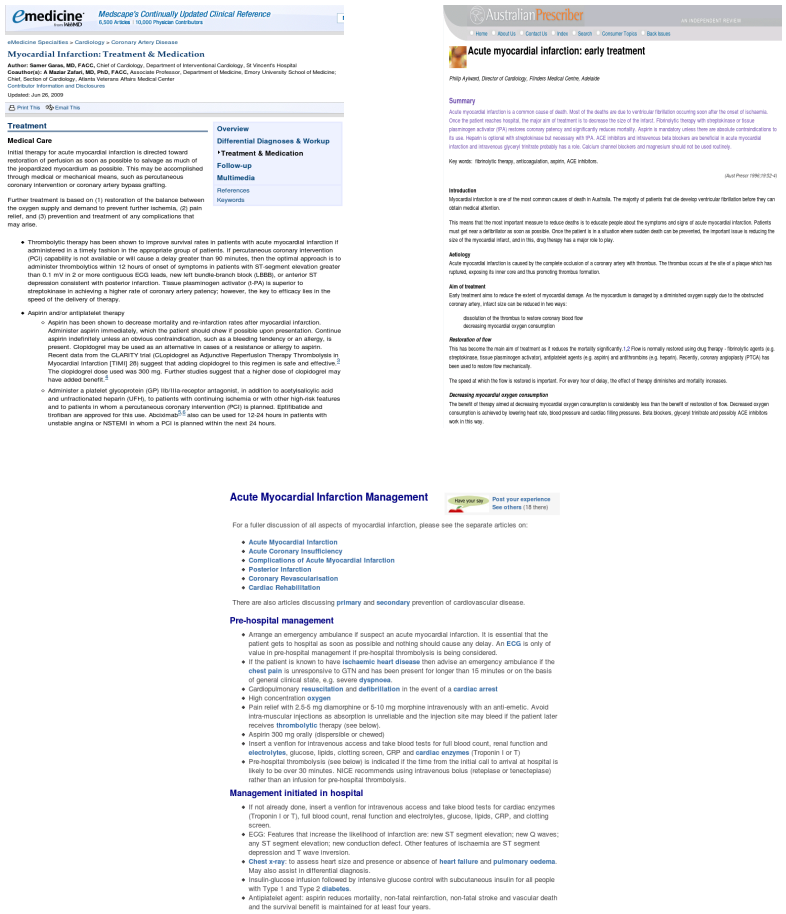


D'après Deléger (2009)

Corpus comparable

Un corpus de textes du même domaine, genre, etc.

Langue A (*EN*)



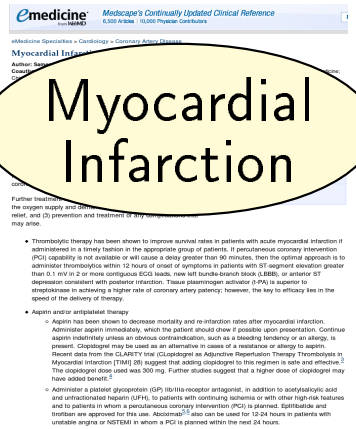
Langue B (*FR*)

D'après Deléger (2009)

Corpus comparable

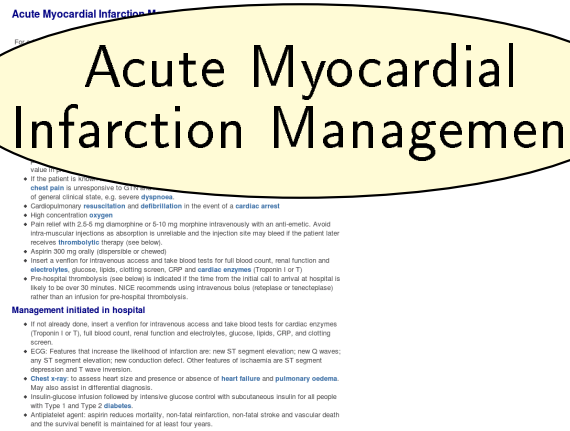
Un corpus de textes du même domaine, genre, etc.

Langue A (*EN*)



Myocardial Infarction

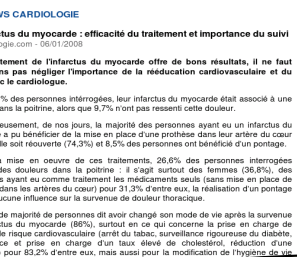
Acute Myocardial Infarction: early treatment



Acute Myocardial Infarction Management

Langue B (*FR*)

Prévention secondaire après infarctus du myocarde



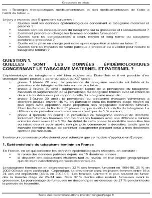
Infarctus du myocarde

D'après Deléger (2009)

Corpus comparable

Ici, textes d'une même langue avec deux variétés de discours

Discours spécialisé



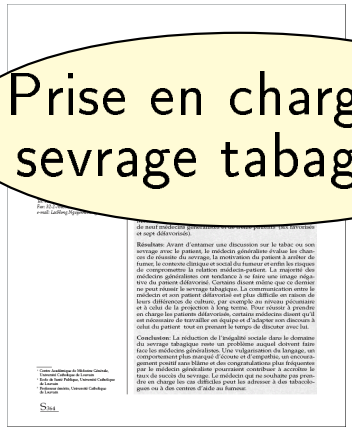
Discours grand public

D'après Deléger (2009)

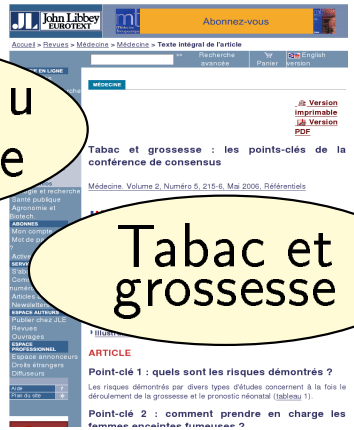
Corpus comparable

Ici, textes d'une même langue avec deux variétés de discours

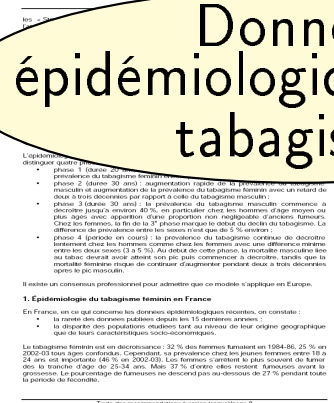
Discours spécialisé



Prise en charge du sevrage tabagique



Tabac et grossesse



Données épidémiologiques sur le tabagisme

Discours grand public

Les dangers du tabagisme passif

Médicaments pour arrêter de fumer

D'après Deléger (2009)

Corpus parallèles

Un corpus de textes et le corpus de leurs traductions

source		cible
texte s_1	$\longleftrightarrow$	texte c_1
texte s_2	$\longleftrightarrow$	texte c_2
texte s_3	$\longleftrightarrow$	texte c_3
...		...
texte s_n	$\longleftrightarrow$	texte c_n

- Dans deux langues différentes (ou la même langue)
- Textes traduits

Corpus comparables

Deux corpus de textes de même domaine, genre, etc.

source
texte s_1
texte s_2
texte s_3
...
texte s_n

cible
texte c_a
texte c_b
...
texte c_m

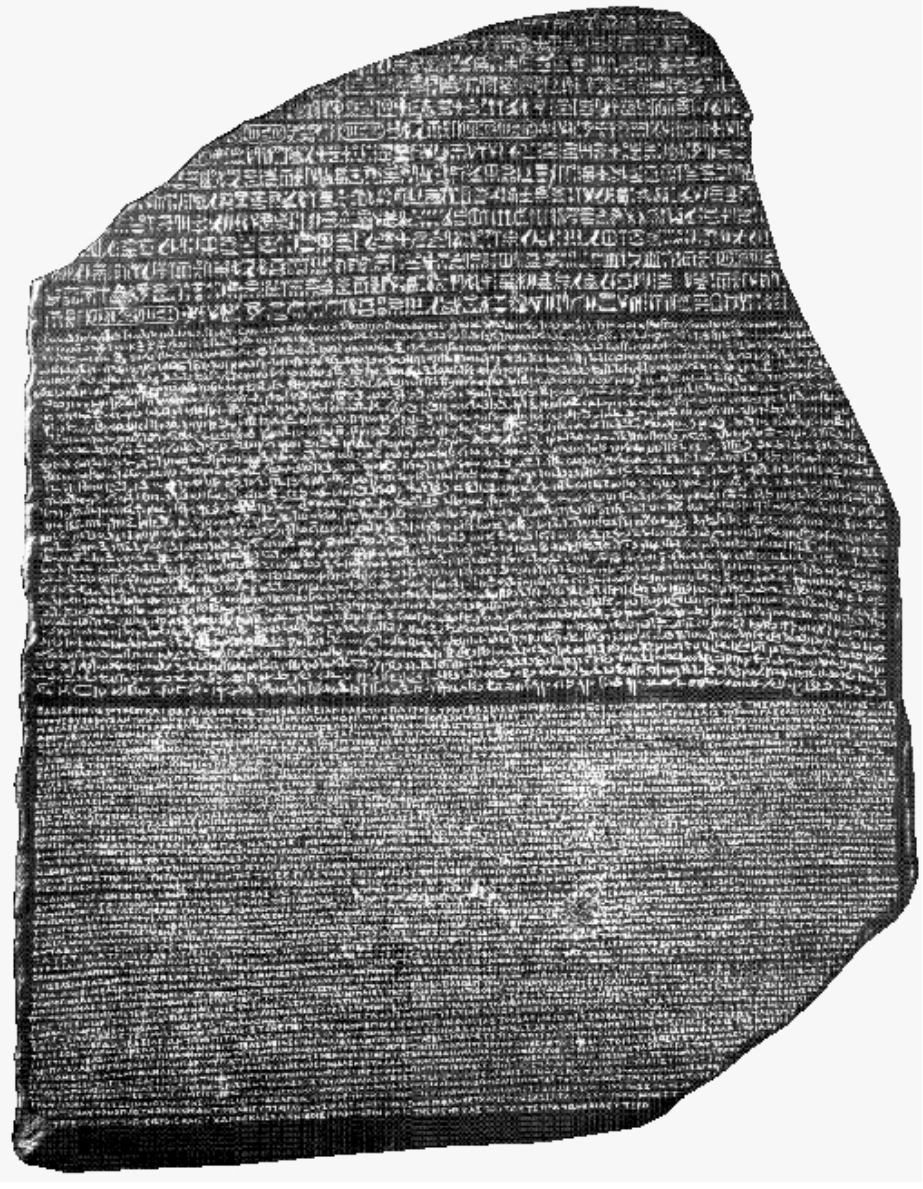
- Dans deux langues différentes (ou la même langue)
- Les textes des deux langues peuvent être originaux (non traduits)

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources

2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables

3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

Où trouver des corpus parallèles ?



Corpus disponibles (exemples)

Ouvrages traduits

- Textes religieux : La Bible, le Coran, etc.
- Déclaration des droits de l'homme
- Romans
- Documentation : manuels techniques

Débats, textes législatifs multilingues

- Parlement canadien (Hansard : français, anglais, inuktitut)
- Parlement européen (Europarl : français, italien, espagnol, portugais, anglais, allemand, néerlandais, danois, suédois, grec, finnois)
- Parlement de Hong Kong (anglais, chinois)
- Nations Unies
- Acquis communautaire (JR Acquis)

Constituer un corpus parallèle de documents web

Limitations des corpus parallèles disponibles

- Taille
- Domaine, genre
- Langues représentées

Obtenir d'autres types de documents

- Page web et sa traduction
- Plus largement, un site web multilingue
  - Exemple : Santé Canada

Pages web parallèles

Santé Canada Health Canada Canada

English | Contactez-nous | Aide | Recherche | Site du Canada
Vie saine | Soins de santé | Maladies et affections | Protection de la santé | Salle des médias

Santé Canada
ON VOUS INFORME En direct

Santé Canada Accueil
Communiqué
Retour aux communiqués
Information
Retour au réaction au rapport du vérificateur général

novembre 1999

Information

Réponse de Santé Canada au rapport du vérificateur général

Gestion des poussées d'intoxication alimentaire

Rôles et responsabilités

Quand il s'agit de salubrité des aliments, Santé Canada travaille en étroite collaboration avec l'Agence canadienne d'inspection des aliments (ACIA). Par le truchement du Programme des produits alimentaires, Santé Canada établit les politiques et les normes relatives à la sécurité et à la qualité nutritive des aliments vendus au Canada. Santé Canada a aussi pour mandat d'évaluer l'efficacité des activités de l'ACIA relatives à la sécurité des aliments.

Lorsque survient une poussée d'intoxication d'origine alimentaire, le Laboratoire de lutte contre la maladie de Santé Canada coopère avec l'ACIA et d'autres ordres de gouvernement aux activités d'enquête et de circonscription rapide de la poussée. L'ACIA est chargée de l'application des mesures se rapportant aux situations d'urgence liées aux aliments et joue un rôle prépondérant dans les

Health Canada Santé Canada Canada

Français | Contact Us | Help | Search | Canada Site
Healthy Living | Health Care | Diseases & Conditions | Health Protection | Media Room

Health Canada
KEEPING YOU INFORMED Online

Health Canada Home
News Release
Back to Release

November 1999

Health Canada's response to the Auditor General's Report

Managing Food-Borne Disease Outbreaks

Roles and Responsibilities

When it comes to food safety, Health Canada works closely with the Canadian Food Inspection Agency (CFIA). Through the Food Program, Health Canada establishes the policies and standards for the safety and nutritional quality of food sold in Canada. It is also responsible for assessing the effectiveness of CFIA's activities related to food safety.

In the event of a food-borne disease outbreak, Health Canada's Laboratory Centre for Disease Control collaborates with CFIA and other levels of government to investigate and control the outbreak in a timely way. The CFIA is responsible for enforcement actions in food-related emergencies and the Agency takes a lead role in investigations and coordination of food safety emergency responses. PDF Provincial and local medical officers of health have a legislative mandate to investigate disease outbreaks, and provincial laboratories provide laboratory services

Méthode générale de collecte

- Téléchargement du site entier
- Repérage de couples de pages HTML par leurs liens
- Vérifications : langue, taille du texte
  (puis de la qualité de l'alignement des phrases)

$\rightsquigarrow$ Résultat : $\sim$ 10 000 couples de pages FR – EN

Plus largement : indices de parallélisme

Métainformations

- Faire partie du même site (!)
- Noms de fichiers (URL)
- Liens entre documents (hyperliens)

Être écrit dans deux langues différentes

Similarité du contenu

- Longueur des fichiers (en caractères, mots, paragraphes)
- Similarité de la structure
  - Séquence des balises principales
  - Séquence des longueurs des phrases
- Similarité des mots
  - En direct : cognats
  - À travers un lexique bilingue : % mots traduisibles
- Qualité de l'alignement des phrases (a posteriori)

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources
2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables
3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

Constitution de corpus comparables

- Sélection de la langue
  - Détecteur de langue (Grefenstette & Nioche, 2000)
- Sélection du thème
  - Catégorisation automatique (nombreux travaux) (Sebastiani, 2002)
- Sélection du genre, du type de discours
  - Classification et catégorisation automatique (travaux moins nombreux)
    - (Karlgren, 1999; Santini *et al.*, 2006; Goeuriot *et al.*, 2009; Ke & Zweigenbaum, 2009)

Une large part du travail de constitution peut rester manuelle

Exemples de corpus comparables

Corpus synchrones

- Corpus de nouvelles synchrones
  - Wall Street Journal (en), Nikkei Financial News (ja), 1993–1994 (Fung & McKeown, 1997)
  - Frankfurter Allgemeine Zeitung (de, 1993–1996), Guardian (en, 1990–1994) (Rapp, 1999)

Exemples de corpus comparables

Autres critères de comparabilité

- Corpus Comparable CISMeF-CliniWeb (C4) (Chiao, 2004)
  - Documents web indexés par le même ensemble de descripteurs MeSH
- Corpus du projet DECO (Goeuriot *et al.*, 2008)
  - Thème = Diabète ; Langue = fr, ja, ru ; scientifique, populaire
- Corpus du projet C-Mantic (`http://www.c-mantic.org/`)
  - Thème = tabac ; Langue = fr, en, zh ; spécialisé, grand public, pro, anti...
- Corpus médicaux (Deléger, 2009)
  - Thème = tabac, diabète, cancer ; Langue = fr, en ; spécialisé, grand public

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources

2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables

3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

Méthodes externes et méthodes internes

Contexte *vs.* constitution

Méthodes externes : *contexte d'usage* d'un mot

- Usage dans des corpus parallèles ou comparables

Méthodes internes : *forme* d'un mot

- Similarité de forme entre mots en relation de traduction

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources
2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables
3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

Alignement dans des corpus parallèles

Enchaînement habituel de traitements

- Alignement (appariement) de documents (voir plus haut)
- Alignement de phrases
- Alignement de mots ou expressions

Alignement de phrases

- Le cas idéal : alignement 1-1

anglais	français
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.

d'après *Gale & Church (1993)*

Le parallélisme n'est pas toujours strict : 2-1

- Alignement 2-1

anglais	français
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Le parallélisme n'est pas toujours strict : 2-2

- Alignement 2-2

anglais	français
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.

Principes d'alignement de phrases

Comment savoir quelles phrases se correspondent ?

- Similarité de structure des textes
  - structure hiérarchique (paragraphes...)
  - régularité de l'ordre des phrases
- Similarité des phrases
  - forme : longueur
  - contenu lexical : ponctuations, nombres, *cognats*, mots en relation de traduction (à travers lexique bilingue)

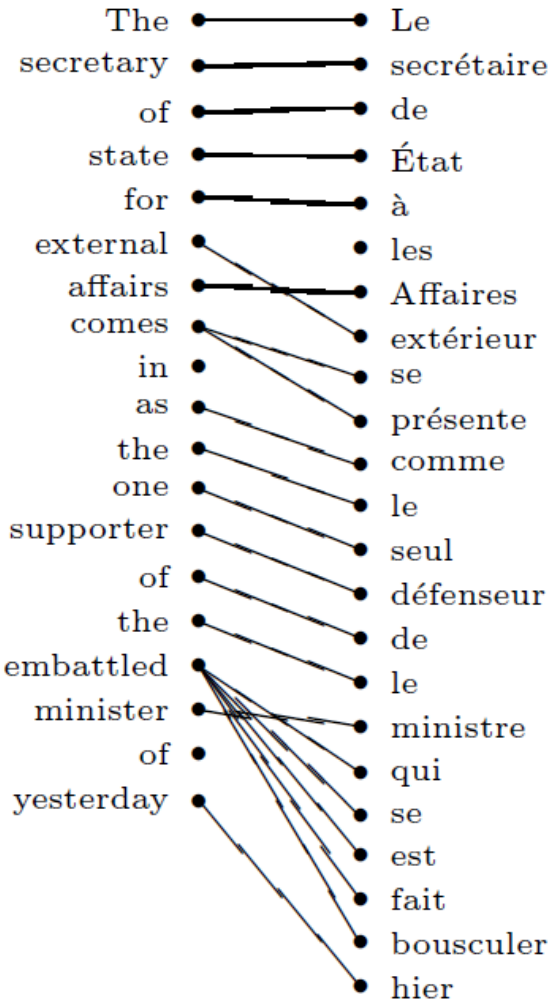
Quelques systèmes d'alignement de phrases

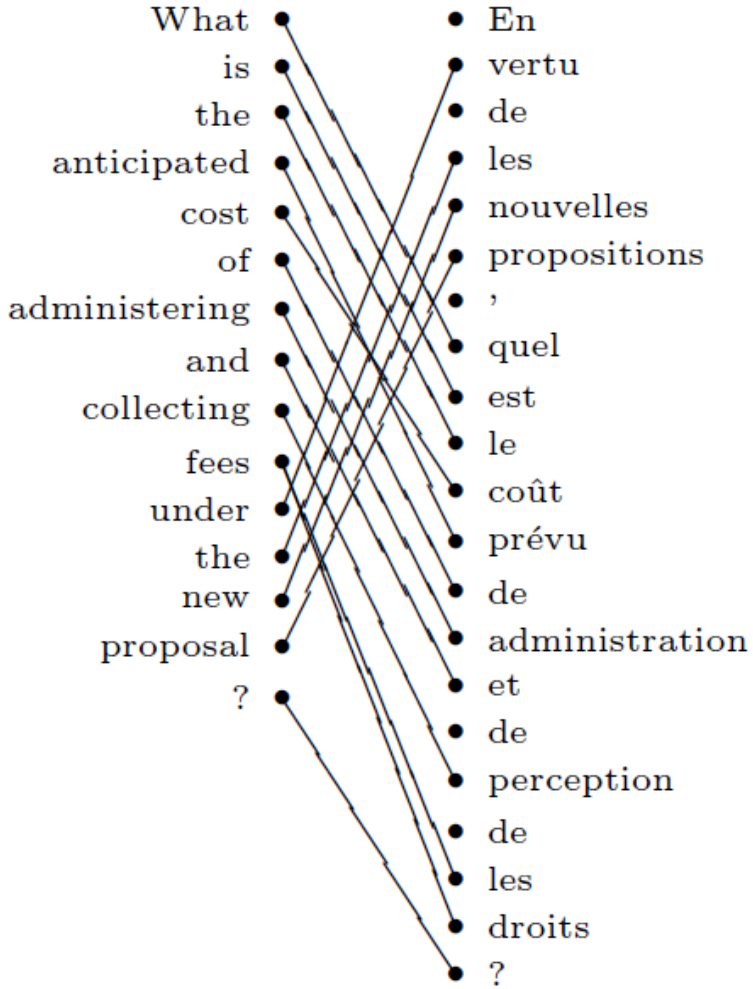
Char_align (Gale & Church, 1993) : longueur des phrases

GMA/GSA (Melamed, 1999) : mixte, avec cognats, lexique

(Moore, 2002) : mixte, sans lexique externe

Alignement de mots





d'après (*Macklovitch & Langlais, 2004*)

Principes d'alignement de mots

Comment savoir quels mots se correspondent ?

- Co-occurrence fréquente dans des phrases alignées
- Probabilité de traduction d'un mot par un autre (hors contexte), apprise ou trouvée dans un dictionnaire
- Position des mots
- Possibilié de traduction n:n (*fertilité*)
- Dépendance (syntaxique) entre mots
- ...

Quelques systèmes d'alignement de mots

- GIZA++ : modèle statistique (Och & Ney, 2003)
  - `http://www.fjoch.com/GIZA++.html`
    `http://code.google.com/p/giza-pp/`
- Méthodes heuristiques (Melamed, 1999)
- I-Tools : mixte statistique et linguistique (Ahrenberg *et al.*, 2003)
- Inversion Transduction Grammar : linguistique+ (Wu, 1995)

1. Faire se rencontrer ressources langagières et traitements automatiques
   - Traitements informatiques du multilinguisme
   - Ressources langagières
   - Constitution automatisée de ressources
2. Constitution automatique de corpus multilingues
   - Constitution de corpus parallèles
   - Constitution de corpus comparables
3. Construction automatique de lexiques multilingues
   - Alignement dans des corpus parallèles
   - Alignement dans des corpus comparables
   - Méthodes internes : génération de traductions

Alignement dans des corpus comparables

Enchaînement habituel de traitements

- Analyse distributionnelle monolingue
- Analyse distributionnelle translingue
- Similarité distributionnelle translingue

Donnée : lexique bilingue partiel (amorçage)

Principe : analyse distributionnelle translingue

- Hypothèse distributionnelle : le sens d'un mot est déterminé par l'ensemble de ses usages (Firth, 1957; Harris, 1991)
  - → Deux mots d'emplois similaires ont des sens proches
- Extension translingue :
  - Deux mots de deux langues différentes qui ont des emplois similaires dans leur langue respective ont des sens proches
  - Les mots qui ont les emplois les plus similaires sont potentiellement en relation de traduction

Comment comparer les emplois des mots d'une langue à l'autre ?

Principe : analyse distributionnelle translingue

- Hypothèse distributionnelle : le sens d'un mot est déterminé par l'ensemble de ses usages (Firth, 1957; Harris, 1991)
  - → Deux mots d'emplois similaires ont des sens proches
- Extension translingue :
  - Deux mots de deux langues différentes qui ont des emplois similaires dans leur langue respective ont des sens proches
  - Les mots qui ont les emplois les plus similaires sont potentiellement en relation de traduction

Comment comparer les emplois des mots d'une langue à l'autre ?

Analyse distributionnelle monolingue

- Un mot est caractérisé par sa force d'association avec chaque autre mot
- Représentation : vecteur, dimension = nombre de mots du corpus
- En pratique : réduit la dimension au nombre de mots du lexique d'amorçage

en français	score	converti en anglais
adénome	(11.8)	adenoma
cellule	(8.9)	cell
examen	(5.9)	test
hyperplasie	(14.2)	hyperplasia
lésion	(8.8)	lesion
nucléole	(17.4)	nucleolus
photographie	(13.9)	photograph
prolifération	(11.9)	proliferation
prostate	(9.1)	prostate
...		

Analyse distributionnelle monolingue

- Un mot est caractérisé par sa force d'association avec chaque autre mot
- Représentation : vecteur, dimension = nombre de mots du corpus
- En pratique : réduit la dimension au nombre de mots du lexique d'amorçage

en français	score	converti en anglais
adénome	(11.8)	adenoma
cellule	(8.9)	cell
examen	(5.9)	test
hyperplasie	(14.2)	hyperplasia
lésion	(8.8)	lesion
nucléole	(17.4)	nucleolus
photographie	(13.9)	photograph
prolifération	(11.9)	proliferation
prostate	(9.1)	prostate
...		

Analyse distributionnelle translingue

On connaît la traduction de chaque mot du lexique d'amorçage

Un profil distributionnel construit dans une langue peut donc se lire dans l'autre langue

Les profils distributionnels des mots des deux corpus peuvent ainsi être comparés

Similarité distributionnelle translingue

Comparaison de vecteurs: mesures classiques

- Cosinus : angle entre deux vecteurs
- Jaccard : intersection / union
- Manhattan : somme des distances sur chaque dimension

Les mots cibles dont les profils distributionnels sont les plus proches d'un mot source sont candidats à sa traduction
Mots anglais dont le profil est le plus similaire à foie

français	anglais	similarité
foie	lung	.270294
foie	liver	.231073
foie	pain	.174125
foie	patient	.162746
foie	tumor	.137852
foie	disease	.136998
foie	primary	.119938
foie	treatment	.119257
foie	brain	.109586
foie	cancer	.105038
foie	bone	.104870

Recherche symétrique

(Sadat *et al.*, 2003; Chiao *et al.*, 2004)

foie → ?

français	anglais	similarité
foie	lung	.270294
foie	liver	.231073
foie	pain	.174125
foie	patient	.162746
foie	tumor	.137852
foie	disease	.136998
foie	primary	.119938
foie	treatment	.119257
foie	brain	.109586
foie	cancer	.105038
foie	bone	.104870

? ← liver

anglais	français	similarité
liver	foie	.365169
liver	rare	.309686
liver	associée	.292330
liver	alzheimer	.284989
liver	transmissible	.269096
liver	fréquente	.263598
liver	pathologie	.257709
liver	cardiovasculaire	.250468
liver	cardio-vasculaire	.248039
liver	creutzfeldt-jakob	.243688
liver	hépatique	.242475
liver	origine	.240563

foie ↔

candidats	rang_{FrEn}	rang_{EnFr}	MH	nouveau rang
lung	1	4	1.60	2
liver	2	1	1.33	1
pain	3	31	5.48	4

Méthodes internes : génération de traductions

Méthode interne : utilise la forme d'un mot

Deux exemples de méthodes :

1. Génération de règles de transduction
2. Traduction par analogie formelle

Entrée : lexique bilingue partiel

Sortie : lexique bilingue étendu

Génération de règles de transduction

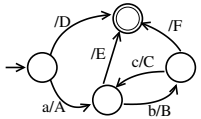
English	French
zirconium	zirconium
...	...
ophthalmotoxin	ophtalmotoxine
ophthalmologist	ophtalmologiste
...	...
oscheitis	oschéite
...	...

- Exemples : paires {mot source, mot cible} du lexique bilingue partiel
- Inférence d'un transducteur qui représente les correspondances source⟶cible
- Application du transducteur sur d'autres mots source

(Claveau & Zweigenbaum, 2005)

Génération de règles de transduction

English	French
zirconium	zirconium
...	...
ophthalmotoxin	ophtalmotoxine
ophthalmologist	ophtalmologiste
...	...
oscheitis	oschéite
...	...

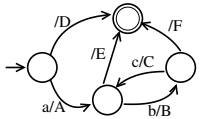
- Exemples : paires {mot source, mot cible} du lexique bilingue partiel
- Inférence d'un transducteur qui représente les correspondances source⟶cible 

- Application du transducteur sur d'autres mots source

(Claveau & Zweigenbaum, 2005)

Génération de règles de transduction

English	French
zirconium	zirconium
...	...
ophthalmotoxin	ophtalmotoxine
ophthalmologist	ophtalmologiste
...	...
oscheitis	oschéite
...	...

- Exemples : paires {mot source, mot cible} du lexique bilingue partiel
- Inférence d'un transducteur qui représente les correspondances source⟶cible 

- Application du transducteur sur d'autres mots source

(Claveau & Zweigenbaum, 2005)

Traduction par analogie formelle

Donnée : lexique bilingue partiel

Entrée : mot source

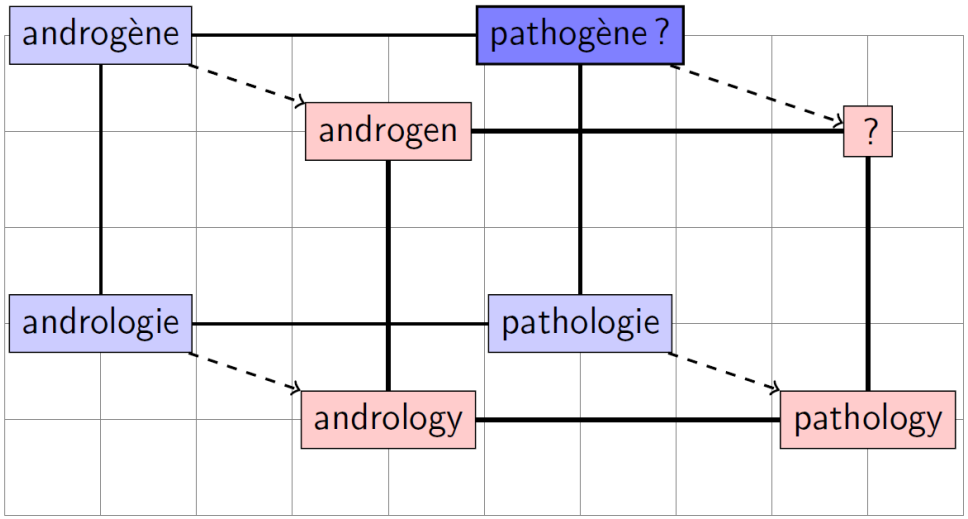
Transfert d'analogies formelles

- Recensement d'analogies formelles qui produisent le mot de départ en langue source
- Transfert de ces analogies en langue cible
- Résolution des analogies formelles en langue cible

(Langlais et al., *2009)*

Traduction par analogie formelle

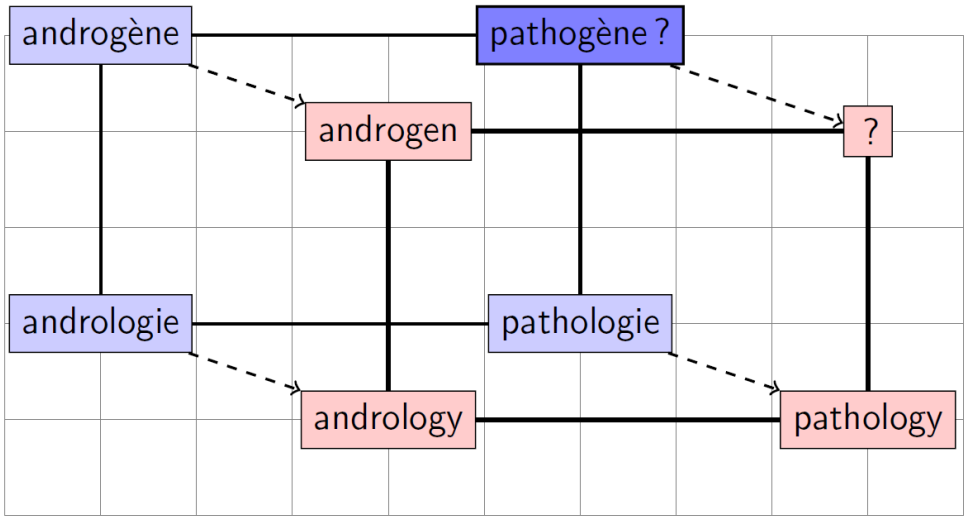
Traduire *pathogène* : carré analogique source, équation analogique cible



on voit : [androlologie : androgène :: pathologie : pathogène]

Traduction par analogie formelle

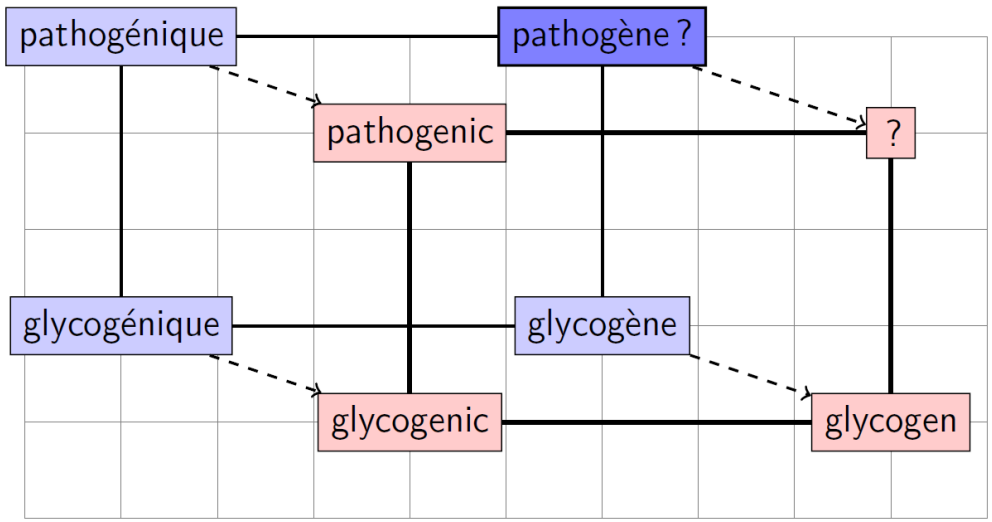
Traduire *pathogène* : carré analogique source, équation analogique cible



on voit : [andrologie : androgène :: pathologie : pathogène]
résoudre : [andrology : androgen :: pathology : ?]

Traduction par analogie formelle

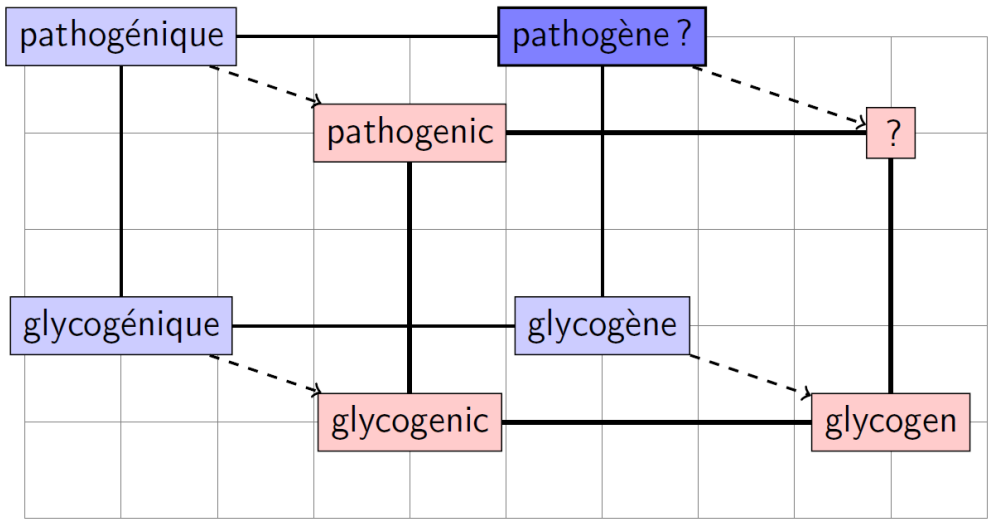
Traduire *pathogène* : carré analogique source, équation analogique cible



on voit : [glycogénique : pathogénique :: glycogène : pathogène]

Traduction par analogie formelle

Traduire *pathogène* : carré analogique source, équation analogique cible



on voit : [glycogénique : pathogénique :: glycogène : pathogène]

résoudre : [glycogenic : pathogenic :: glycogen : ?]

Pour finir

Conclusion

- Intérêt et limites des corpus parallèles
  - Meilleure précision, meilleur rendement
  - Volume borné
- Potentiel et difficultés des corpus comparables
  - Langue plus naturelle, volume potentiel plus grand
  - Précision plus faible des propositions de traduction
- Indications et limites des méthodes internes
  - Génération de traductions non vues
  - Repose sur la similarité de construction des mots ou sur la proximité des langues
- Méthodes automatiques vs intervention humaine
  - Besoin de pilotage des méthodes automatiques
  - Besoin de validation des ressources constituées

Bibliographie I

Ahrenberg L., Merkel M. & Petterstedt M. (2003). Interactive word alignment for language engineering. In A. Copestake & J. Hajic, Eds., *Proceedings EACL 2003*, p. 49–52, Budapest.

Chiao Y.-C. (2004). *Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue*. Thèse de doctorat, informatique médicale, Université Paris 6.

Chiao Y.-C., Sta J.-D. & Zweigenbaum P. (2004). A novel approach to improve word translations extraction from non-parallel, comparable corpora. In *Proceedings International Joint Conference on Natural Language Processing*, Hainan, China: AFNLP.

Claveau V. & Zweigenbaum P. (2005). Traduction de termes biomédicaux par inférence de transducteurs. In *Proceedings Traitement automatique des langues naturelles (Traitement automatique des langues naturelles)*, Dourdan.

Bibliographie II

Deléger L. (2009). *Exploitation de corpus parallèles et comparables pour la détection de correspondances lexicales : application au domaine médical*. Thèse de doctorat, informatique médicale, Université Pierre et Marie Curie.

Firth J. R. (1957). *Papers in Linguistics, 1934–1951*. London: Oxford University Press.

Fung P. & McKeown K. (1997). Finding terminology translations from parallel corpora. In *Proceedings Fifth Annual Workshop on Very Large Corpora*, p. 192–202: ACL.

Gale W. & Church K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, **19**(3), 75–102.

Bibliographie III

Goeuriot L., Grabar N. & Daille B. (2008). Characterization of scientific and popular science discourse in French, Japanese and Russian. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis & D. Tapias, Eds., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Goeuriot L., Morin E. & Daille B. (2009). Reconnaissance du type de discours dans des corpus comparables spécialisés. In *Proceedings CORIA 2009*: ARIA. Ce volume.

Grefenstette G. & Nioche J. (2000). Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, p. 237–246, Paris, France: C.I.D.

Bibliographie IV

Harris Z. S. (1991). *A theory of language and information. A mathematical approach*. Oxford: Oxford University Press.

Karlgren J. (1999). Stylistic experiments in information retrieval. In T. Strzalkowski, Ed., *Natural language information retrieval*, volume 7 of *Text, speech and language technology*, chapter 6, p. 147–166. Dordrecht & Boston: Kluwer Academic Publishers.

Ke G. & Zweigenbaum P. (2009). Catégorisation automatique de pages web chinoises : documents spécialisés *vs* grand public sur le tabagisme. In *Proceedings CORIA 2009*, p. 203–128: ARIA.

Langlais P., Yvon F. & Zweigenbaum P. (2009). Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Proceedings 12th Conference of the European Chapter of the ACL (EACL 2009)*, p. 487–495, Athens, Greece: Association for Computational Linguistics.

Bibliographie V

Macklovitch E. & Langlais P. (2004). Le bi-texte et ses applications. In P. Blache, Ed., *Proceedings of TALN 2004 (Traitement automatique des langues naturelles)*, Fès, Maroc: ATALA LPL. Tutoriel.

Melamed I. D. (1999). Bitext maps and alignments via pattern recognition. *Computational Linguistics*, **25**(1), 107–130.

Moore R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Machine Translation: From Research to Real Users*, p. 135–244, Heidelberg, Germany: Springer-Verlag. Proceedings 5th Conference of the Association for Machine Translation in the Americas.

Och F. J. & Ney H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

Rapp R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th ACL*, College Park, Maryland.

Bibliographie VI

Sadat F., Yoshikawa M. & Uemura S. (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval: Hybrid statistics-based and linguistics-based approach. In J. Adachi & K.-F. Wong, Eds., *Proceedings Sixth International Workshop on Information Retrieval with Asian Languages*, p. 57–64.

Santini M., Power R. & Evans E. (2006). Implementing a characterization of genre for automatic genre identification of Web pages. In *Proceedings COLING/ACL 2006 Main Conference Poster Sessions*, p. 699–706, Sydney.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

Wu D. (1995). Grammarless extraction of phrasal translation examples from parallel texts. In *In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, p. 354–372.