Utilisation du Web comme ressource bilingue

Traduction de termes complexes français/anglais

Stéphanie LEON
Équipe DELIC, Université de Provence

Traduction automatique

- <u>Un problème essentiel</u> : polysémie / homonymie

  caisse$_1$ (usage BANQUE) > fund
  caisse$_2$ (usage MUSIQUE) > drum
  …
  8 traductions recensées!

- Un exemple du traducteur *Systran*

  caisse centrale ⇒ central **case**
  ⇒ central **fund**

Objectif

Corpus parallèles

Ensemble de textes alignés avec leur traduction au niveau du paragraphe, de la phrase, des expressions ou des mots.

Corpus comparables

Corpus de langues différentes traitant du même domaine mais non parallèles.

Ressources rares + domaines restreints

✦ Web : mega base de données lexicale multilingue

1- Le Web : un « corpus » partiellement parallèle

- Textes parallèles ou traductions ponctuelles sur le Web :

"*Further support was guaranteed by large loans from the World Bank, the Saudi Fund, France's Central Fund for Economic Cooperation (Caisse Centrale de Coopération Economique--CCCE)*"

- Un corpus « parallèle » ou « partiellement parallèle » :

(Ma & Liberman, 1999); (Nie, 1999); (Nie et al, 2000);

(Nagata, 2001); (Resnik, 1999); (Resnik & Smith, 2002), etc.

2- Le Web, une source d'informations quantitatives

✦ Fréquence plus élevée des traductions existantes

✦ Une source d'informations quantitatives pour la traduction :

-*(Grefenstette, 1999)*
-*(Cao & Li, 2002)*
-*(Léon & Millon, 2005)*

Un exemple

✦ (Léon & Millon, 2005) : fréquence sur le Web des traductions candidates. Exemple :

Appareil numérique > digital **aeroplane**? Digital **camera**? …

Difficulté

- Pas de test d'équivalence entre terme complexe source et sa traduction candidate. Exemple :

Cours de formation > group rate *(> tarif de groupe)*

Traduction existante, mais non équivalente

« Mondes lexicaux »…

Co-occurrences fréquentes d'un mot ou d'un terme complexe (phrase, paragraphe).

(Véronis, 2003; 2004)

✦ Mondes lexicaux sur le Web relativement proches entre le français et l'anglais. Exemple :

Appareil numérique

canon, photographie, nikon, informatique, produits, accessoires, digital, mémoire, kodak, pc, etc.

Digital camera

photography, film, computer, kodak, technology, olympus, canon, right, zoom, sony, etc.

Méthodologie (1)

Termes complexes français

Exalead

Traductions candidates

Dictionnaire Électronique

Filtre : requêtes couples français/anglais

API Yahoo

Méthodologie (2)

2 décisions possibles :

1) Cas des têtes sémantiques polysémiques :

Comparaison
des mondes lexicaux

API Yahoo

2) Cas des têtes sémantiques peu polysémiques :

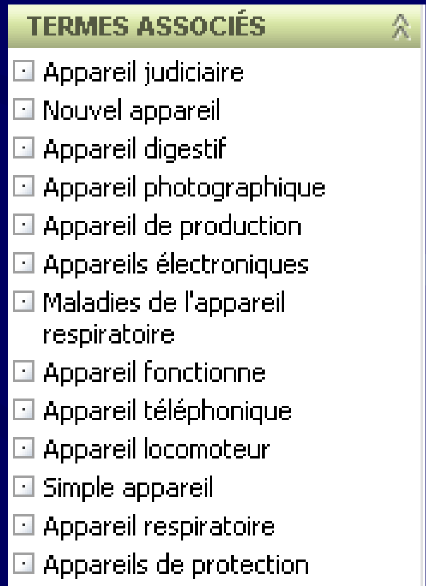
Fréquence des traductions

API Yahoo

1- Acquisition de termes complexes français

- Noms simples français (Lexique *Multext)*

- « Termes associés » : moteur de recherche *Exalead*



- Longueur 2

2- Génération automatique des traductions candidates

Appareil	**numérique**
aeroplane *aircraft* *appliance* etc.	*numerical* *digital*

Traductions candidates	**de *appareil numérique***
numerical aeroplane *numerical aircraft* *numerical appliance* etc.	*digital aeroplane* *digital aircraft* *digital appliance* etc.

3- Un « corpus » partiellement parallèle (1)

✦ Requêtes des couples français/anglais (documents bilingues) (*Yahoo*) :

Terme français/Traduction candidate
"absorption atomique" "atomic absorption"
"absorption atomique" "nuclear absorption"
"accès approprié" "appropriate access"
"accès approprié" "appropriate approach"
"accès approprié" "appropriate approaches"
"accès approprié" "appropriate attack"
"accès approprié" "appropriate bout"
"accès approprié" "appropriate fit"
"accès approprié" "appropriate means of access"
"accès approprié" "appropriate outbreak"
"accès approprié" "appropriate outburst"
"accès approprié" "appropriate spasm"
"accès approprié" "apt access"
"accès approprié" "apt approach"
"accès approprié" "apt approaches"
"accès approprié" "apt attack"

…

3- Un « corpus » partiellement parallèle (2)

- Filtre de fréquence : >=1

COUPLE	FREQUENCE COUPLE
"banque mondiale" "world bank"	131000
"musique contemporaine" "contemporary music"	127000
"secteur public" "public sector"	58800
"double face" "double face"	35977
"transport international" "international transport"	34600
"transport maritime" "maritime transport"	25000
"vin rouge" "red wine"	16800
"vin blanc" "white wine"	16000
"peau douce" "soft skin"	13900
"noble art" "noble art"	12049
"casino virtuel" "virtual casino"	6600
"meilleur casino" "top casino"	6074
"roi lion" "lion king"	4930
"meilleur casino" "best casino"	4162
"horoscope gratuit" "free horoscope"	3376
"chirurgie plastique" "plastic surgery"	3170
"océan indien" "Indian ocean"	3164
"jardin botanique" "botanical garden"	3153
"dictionnaire français" "French dictionary"	2922

. . .

2 décisions de traitements possibles (1)

- Critère : nombre de traductions de la tête sémantique française

- Cas I : Têtes sémantiques polysémiques (nombre de traductions >2)

- Exemple :

  Appareil judiciaire > appareil (11 traductions) *(>aeroplane, aircraft, appliance, brace, etc.)*

2 décisions de traitements possibles (2)

✦ Cas II : Têtes sémantiques peu polysémiques (nombre de traductions <=2)

NOM SOURCE	NB TRADUCTIONS	TERME COMPLEXE SOURCE	TRADUCTION CANDIDATE
secteur	2	secteur public	public sector/district
transport	1	transport international	international transport
vin	1	vin rouge	red wine
casino	1	casino virtuel	virtual casino
roi	1	roi lion	lion king
horoscope	1	horoscope gratuit	free horoscope
chirurgie	1	chirurgie plastique	plastic surgery
océan	1	océan indien	Indian ocean
jardin	1	jardin botanique	botanical garden
dictionnaire	1	dictionnaire français	French dictionary
pilote	2	pilote automatique	automatic pilot/driver
restaurant	1	restaurant gastronomique	gastronomic restaurant
guitare	1	guitare classique	classical guitar

…

Cas I-
Comparaison de mondes lexicaux (têtes sémantiques polysémiques)

Cas I- Construction de mondes lexicaux français (1)

- *API Yahoo* : acquisition automatique des résumés

  « appareil numérique » -« appareils numériques »
  « appareils numériques » -« appareil numérique »
  « appareil numérique » +« appareils numériques »

- Liste des 50 mots les plus fréquents

Appareil Numérique pour €59, 00 Fixer vos souvenirs **Appareils** photos, caméras, vous trouverez tout l´équipement moderne chez l´Homme Moderne. Commandez les en ligne. Paiement sécurisé. Objets originaux du monde.

Mondes lexicaux français

Cas I- Construction de mondes lexicaux français (2)

- Exemples : « appareil numérique » (20 premiers mots)

  canon, photographie, nikon, informatique, produits, accessoires, digital, mémoire, kodak, pc, olympus, flash, zoom, cartes, argentique, prises, gamme, reflex, prise, matériel…

  « appareil judiciaire »

  justice, droits, homme, politique, droit, enfants, parents, administratif, privé, policiers, police, pays, démocratie, discours, nationale, revue, travail, femmes, collaboration, situation…

  « appareil respiratoire »

  sécurité, santé, protection, isolants, air, incendie, formation, pompiers, autonomes, gaz, service, isolant, travail, masques, maladies, autonome, médecine, feu, plongée, circuit…

Cas I- Construction de mondes lexicaux anglais

« digital camera »

photography, film, computer, kodak, technology, olympus, canon, right, zoom, sony, memory, resolution, lens, imaging, fuji, ratings, pc, brands, product, battery…

« judicial apparatus »

law, government, code, administration, reform, civil, political, guidelines, police, power, human, criminal, respect, court, courts, strengthen, justice, democracy, spirit, entrench…

« respiratory apparatus »

anatomy, human, illness, tuscany, air, lungs, splanchnology, arthritises, function, skin, organs, respiration, siena, larynx, rheumatisms, gray, cares, anatomical, practices, flow

Cas I- Comparaison des mondes lexicaux (1)

✦ On comptabilise le nombre de mots en commun via le dictionnaire :

Monde lexical français :	mémoire
Monde lexical anglais :	memory

✦ Cas des mots absents du dictionnaire présents dans les 2 mondes lexicaux (noms propres) : ils sont comptabilisés.

Nikon, Sony, etc.

Cas I- Comparaison des mondes lexicaux (2)

- Indice de Jacquard (* 1000) :

  | inter(X,Y) | / | union (X,Y) | * 1000

- Filtre : indice >= 100. Exemples :

TERME COMPLEXE SOURCE	TRADUCTION CANDIDATE	INDICE DE JACQUARD
droit naturel	natural law	340,00
droit naturel	natural right	333,33
rayon lumineux	light ray	306,93
conseil canadien	Canadian council	300,00
couleur jaune	yellow colour	300,00
fleur rouge	red flower	297,03
comportement sexuel	sexual behaviour	295,92
charge électrique	electric charge	292,93
bec rouge	red beak	278,35
histoire contemporaine	contemporary history	278,35
histoire humaine	human history	272,73
couleur rose	rose colour	270,00
grand escalier	wide staircase	270,00
histoire politique	politics history	270,00

Cas II- Fréquence des traductions (têtes sémantiques non polysémiques)

Cas II- Traductions sans mondes lexicaux

- Filtre : fréquence de la traduction candidate sur Yahoo (>=5000).

FR	EN	FREQUENCE TRADUCTION
force aérienne	air force	62 000 000
ingénierie informatique	computer science	49 000 000
musique populaire	popular music	41 000 000
école primaire	elementary school	37 000 000
alarme piscine	swimming pool alarm	36 000 000
secteur public	public sector	31 000 000
ancienne école	old school	28 000 000
océan indien	Indian ocean	25 000 000
petit temps	short time	24 000 000
rôle important	important part	23 000 000
médecine alternative	alternative medicine	22 000 000
chirurgie plastique	plastic surgery	21 000 000
propre risque	own risk	20 000 000
rôle important	important role	20 000 000
banque mondiale	world bank	19 000 000
peuple américain	American people	17 000 000
communauté indigène	local community	15 000 000
école primaire	primary school	13 000 000

Résultats,
limites et perspectives

Résultats…

Termes complexes français	Traductions candidates	*Filtre automatique*			*Validation manuelle*		
		Filtre "couples"	Nb avec mondes lexicaux conservés	Nb sans mondes lexicaux conservés	Traductions correctes avec mondes lexicaux	Traductions correctes sans mondes lexicaux	Total traductions correctes
3512	23 275	3541	699	1382	650	1357	2007

- Précision globale : 96.44 % (2007 traductions correctes sur 2081 traductions proposées)

Limites (1)

- Qualité des données :

- locuteurs non natifs/non spécialistes

- view detailed information about flights which includes flight number, name of airliner, origin, destination, departure time and date, arrival time and date, and price. Departure and arrival times are in 24 hour format, e.g. 15:30. All times are PST. Departure and arrival dates are of integer type between 1 - 31 assuming the travel will only be made in December. This date restriction applies to Rental Car and Hotel Reservations too.
- select and book (commit) a flight by the flight number.
- book multiple flights but one at a time. Do not worry about location conflicts.
- view his or her flight schedules. Cancellation is not allowed.

- spams, traductions automatiques de sites

Limites (2)

- Analyse syntaxique lors des requêtes

  restauration complète > complete (Adj, V) restoration
  → « a complete restoration » OR « the complete restoration »

- Construction des mondes lexicaux

  aspirateur > vacuum cleaner

Analyseur syntaxique

Conclusion

- Fréquences sur le Web : filtre préalable mais non suffisant pour réduire les ambiguïtés lexicales pour la traduction

- Entourage linguistique : mondes lexicaux

- Perspective principale : ajout d'un niveau de traitement syntaxique

Merci pour votre attention…