Induction de lexiques bilingues

à partir de corpus comparables et parallèles.

Laurent Jakubina

14 Décembre 2017

Université de Montréal - Département d'informatique et de recherche opérationnelle - RALI

Plan

1. Introduction - Induction de lexiques Bilingues...

2. ...à partir de Corpus Parallèles

   Article n°3 (Jakubina and Langlais, 2016a)

3. ...à partir de Corpus Comparables

   via alignement de mots en corpus comparables

   Article n°1 (Jakubina and Langlais, 2015)

   comme tâche d'évaluation des embeddings interlingues

   Article n°2 (Jakubina and Langlais, 2016b)

   via reclassement supervisé de candidats à la traduction

   Article n°4 (Jakubina and Langlais, 2017)

4. Travaux futurs & Conclusion

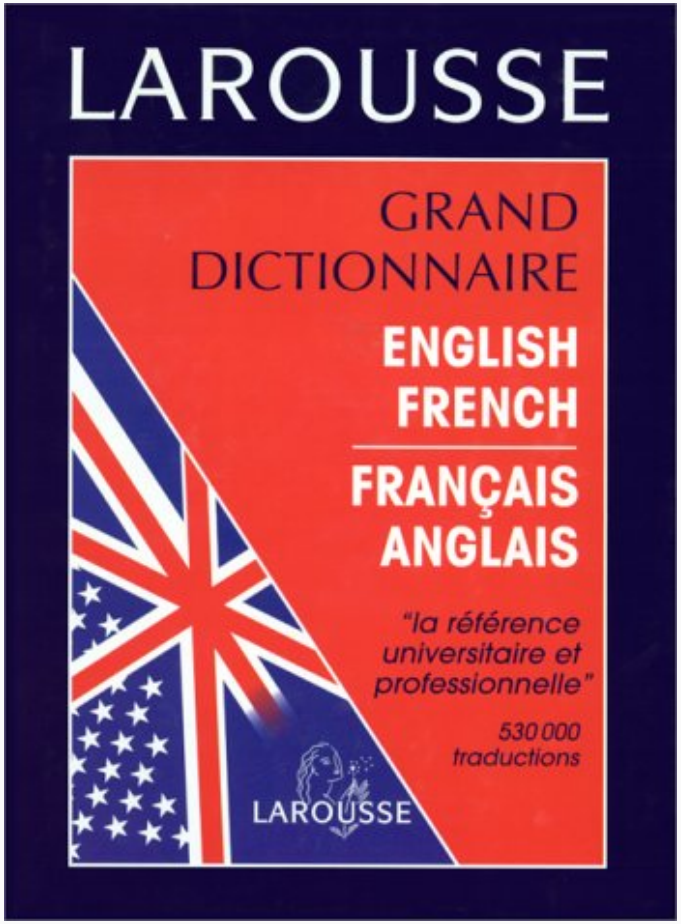
Introduction - Induction de lexiques Bilingues...

Lexiques Bilingues

Lexique Bilingue (LB) ? (syn: dictionnaire de traduction)

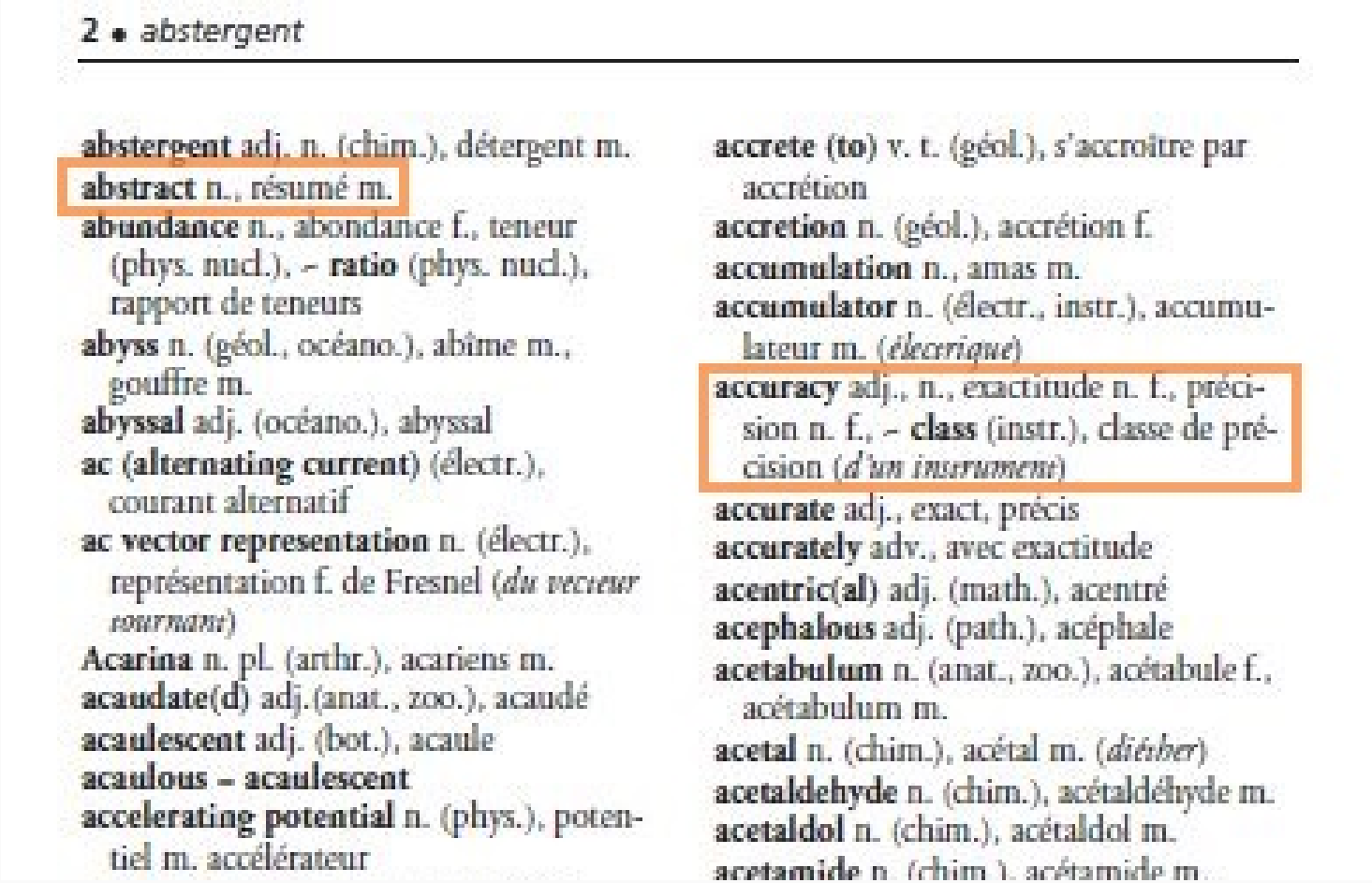
Lexiques Bilingues

Lexique Bilingue (LB) ? (syn: dictionnaire de traduction)



Lexiques Bilingues

Lexique Bilingue (LB) ? (syn: dictionnaire de traduction)



2 • *abstergent*

abstergent adj. n. (chim.), détergent m.
abstract n., résumé m.
abundance n., abondance f., teneur (phys. nucl.), – **ratio** (phys. nucl.), rapport de teneurs
abyss n. (géol., océano.), abîme m., gouffre m.
abyssal adj. (océano.), abyssal
ac (alternating current) (électr.), courant alternatif
ac vector representation n. (électr.), représentation f. de Fresnel (*du vecteur tournant*)
Acarina n. pl. (arthr.), acariens m.
acaudate(d) adj.(anat., zoo.), acaudé
acaulescent adj. (bot.), acaule
acaulous – acaulescent
accelerating potential n. (phys.), potentiel m. accélérateur
accrete (to) v. t. (géol.), s'accroître par accrétion
accretion n. (géol.), accrétion f.
accumulation n., amas m.
accumulator n. (électr., instr.), accumulateur m. (*électrique*)
accuracy adj., n., exactitude n. f., précision n. f., – **class** (instr.), classe de précision (*d'un instrument*)
accurate adj., exact, précis
accurately adv., avec exactitude
acentric(al) adj. (math.), acentré
acephalous adj. (path.), acéphale
acetabulum n. (anat., zoo.), acétabule f., acétabulum m.
acetal n. (chim.), acétal m. (*diéther*)
acetaldehyde n. (chim.), acétaldéhyde m.
acetaldol n. (chim.), acétaldol m.
acetamide n. (chim.), acétamide m.

Induction de Lexiques Bilingues (**ILB**) ?

Induction de Lexiques Bilingues (**ILB**) ?

Ressources (textuelles) Bilingues/Plurilingues

...		...	
patchouli	patchouli		
stickler	pointilleuse	rigoriste	
nightly	nocturne	nuitamment	
gel	gel	gélifier	prendre
astounding	abasourdissant	effarant	renversant
...		...	

Induction de Lexiques Bilingues

Induction de Lexiques Bilingues (**ILB**) ?

...	...
patchouli	?
stickler	?
nightly	?
gel	?
astounding	?
...	...

\+

Ressources (textuelles)

Bilingues/Plurilingues

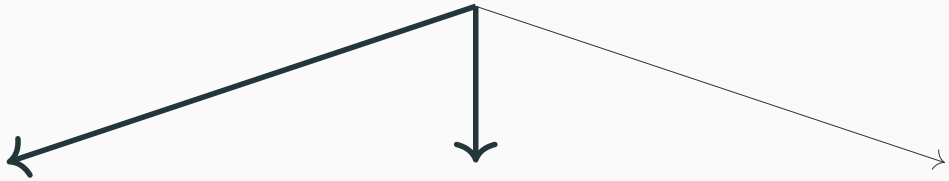
⇒

...
patchouli
pointilleuse
nuitament
prendre
effarant
...

Ressources Bilingues (/ Plurilingues)

Corpus (: ensemble de documents, regroupés selon une perspective précise) disponible en deux langues (source et cible)

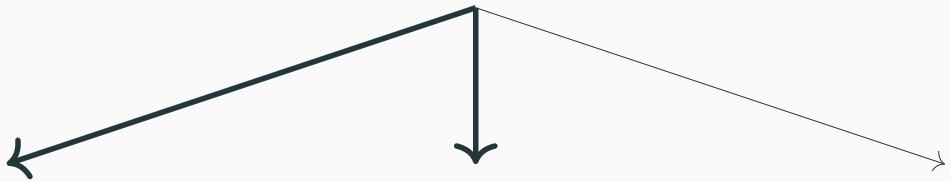
3 Types, selon le niveau de comparabilité source-cible
(Sharoff et al., 2013)



Ressources Bilingues (/ Plurilingues)

Corpus (: ensemble de documents, regroupés selon une perspective précise) disponible en deux langues (source et cible)

3 Types, selon le niveau de comparabilité source-cible
(Sharoff et al., 2013)



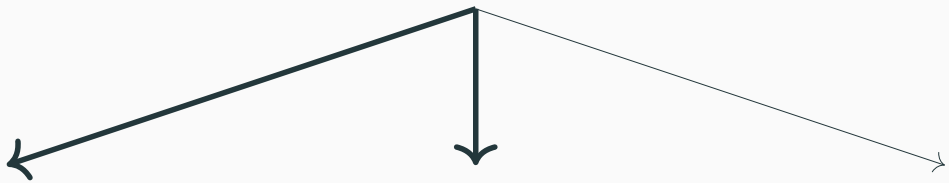
Parallèle

- Traduction
- Rare

Ressources Bilingues (/ Plurilingues)

Corpus (: ensemble de documents, regroupés selon une perspective précise) disponible en deux langues (source et cible)

3 Types, selon le niveau de comparabilité source-cible (Sharoff et al., 2013)



Parallèle

- Traduction
- Rare

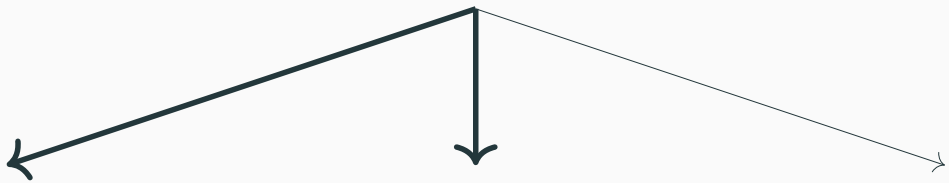
Comparable

- Domaine commun
- Fréquent

Ressources Bilingues (/ Plurilingues)

Corpus (: ensemble de documents, regroupés selon une perspective précise) disponible en deux langues (source et cible)

3 Types, selon le niveau de comparabilité source-cible (Sharoff et al., 2013)



Parallèle

- Traduction
- Rare

Comparable

- Domaine commun
- Fréquent

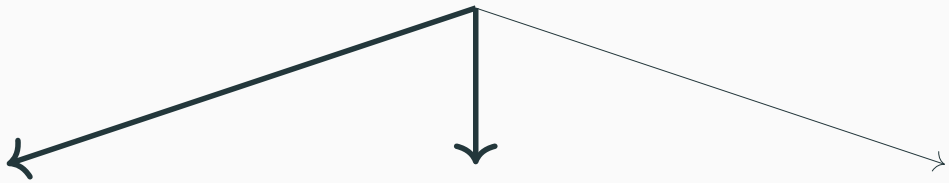
Indépendant

- ~~Relation~~
- Abondant

Ressources Bilingues (/ Plurilingues)

Corpus (: ensemble de documents, regroupés selon une perspective précise) disponible en deux langues (source et cible)

3 Types, selon le niveau de comparabilité source-cible (Sharoff et al., 2013)



Parallèle

- Traduction
- Rare

Comparable

- Domaine commun
- Fréquent

Indépendant

- ~~Relation~~
- Abondant

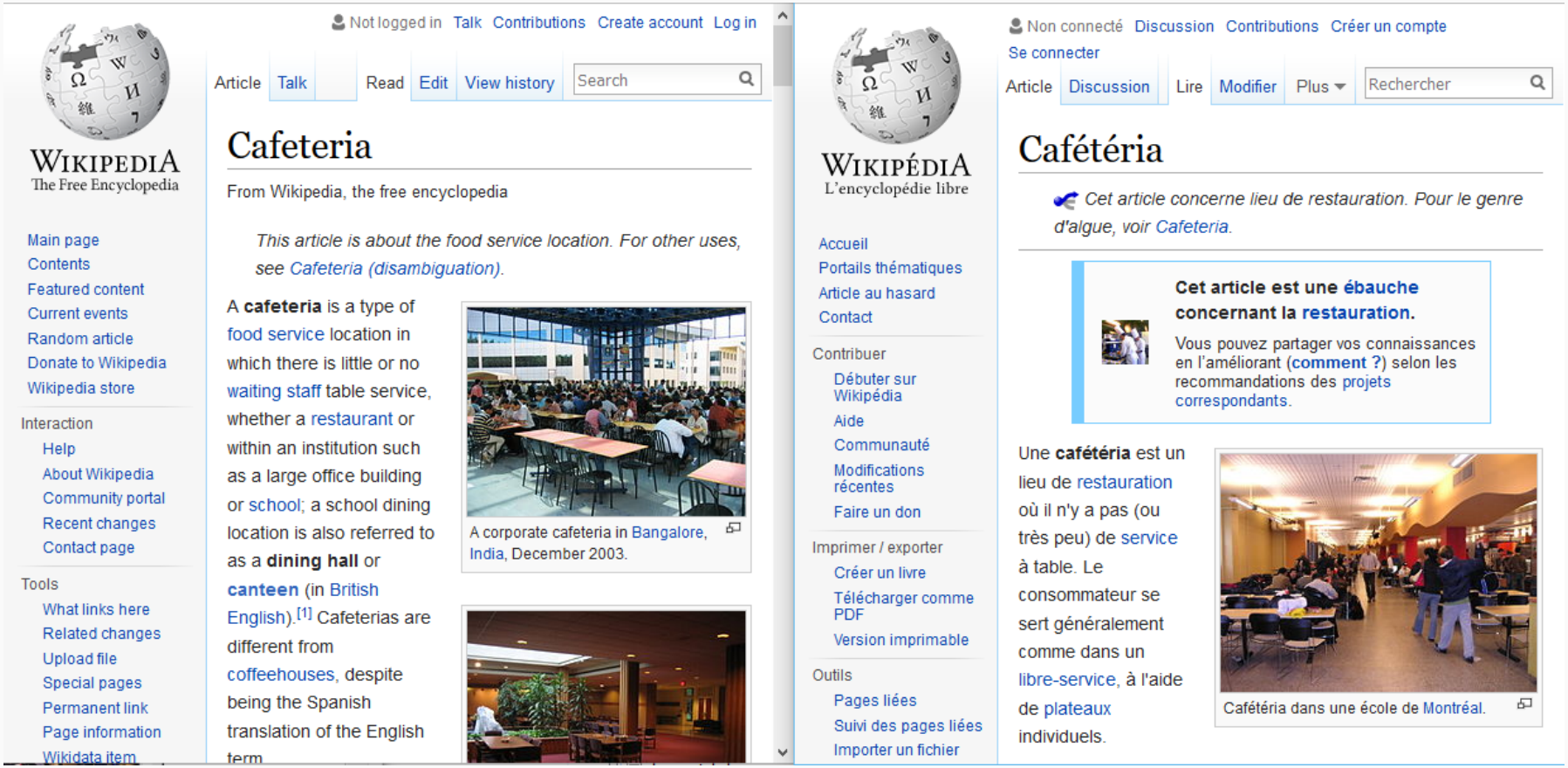
$\Rightarrow$ selon le type, la phase d'induction ($\Rightarrow$) de l'**ILB** diffère...

Corpus Parallèles : Exemple

Total household expenditures for this category reached P215.6 billion (B) in 2000 compared to only P125.6B in 1994. Among the major contributors are flour and noodles. NSO-FIES Total household expenditure on flour reached P1.04B in 2000 per NSO Family Income and Expenditures Survey (FIES). Bread consumption increased by 28% from P18B in 1997 to P23B in 2000. Spending on biscuits totaled P4.7B in 2000 up by P400 million (M) from 1997. Spending on noodles (categorized into bihon, miki, miswa, sotanghon, macaroni, etc.) reached P3.7B in 2000. Instant noodles, on the other hand, are valued at P6B in 2000, with growth estimated at 16% p.a. Household spending for flour is very small.

Les dépenses des ménages pour cette catégorie ont atteint 215,6 milliards de pesos en 2000, comparés à seulement 125,6 milliards de pesos en 1994. Parmi les principaux contributeurs, l'on retrouve la farine et les nouilles. NSO-FIES Selon l'enquête du Bureau national des statistiques portant sur le revenu familial et les dépenses (FIES), les dépenses totales des ménages consacrées à la farine ont atteint 1,04 milliard de pesos en 2000. La consommation de pain a augmenté de 28%, de 18 milliards de pesos en 1997 à 23 milliards en 2000. Les dépenses sur les biscottes ont totalisé 4,7 milliards de pesos en 2000, ayant connu une augmentation de 400 millions de pesos depuis 1997. Les dépenses au chapitre des nouilles (classées sous vermicelle chinois, nouilles de riz miki, de blé, asiatiques, macaroni, etc.) ont atteint 3,7 milliards de pesos en 2000. D'autre part, les dépenses sur les nouilles instantanées sont évaluées à 6 milliards de pesos en 2000, avec une croissance estimée de 16% p.a. Les dépenses des ménages concernant la farine sont vraiment minimes.

Corpus Comparables : Exemple



WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search

Cafeteria

From Wikipedia, the free encyclopedia

This article is about the food service location. For other uses, see Cafeteria (disambiguation).

A **cafeteria** is a type of food service location in which there is little or no waiting staff table service, whether a restaurant or within an institution such as a large office building or school; a school dining location is also referred to as a **dining hall** or **canteen** (in British English).[1] Cafeterias are different from coffeehouses, despite being the Spanish translation of the English term

A corporate cafeteria in Bangalore, India, December 2003.

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item

WIKIPÉDIA
L'encyclopédie libre

Non connecté Discussion Contributions Créer un compte Se connecter

Article Discussion Lire Modifier Plus Rechercher

Cafétéria

Cet article concerne lieu de restauration. Pour le genre d'algue, voir Cafeteria.

Cet article est une ébauche concernant la restauration.

Vous pouvez partager vos connaissances en l'améliorant (**comment ?**) selon les recommandations des projets correspondants.

Une **cafétéria** est un lieu de restauration où il n'y a pas (ou très peu) de service à table. Le consommateur se sert généralement comme dans un libre-service, à l'aide de plateaux individuels.

Cafétéria dans une école de Montréal.

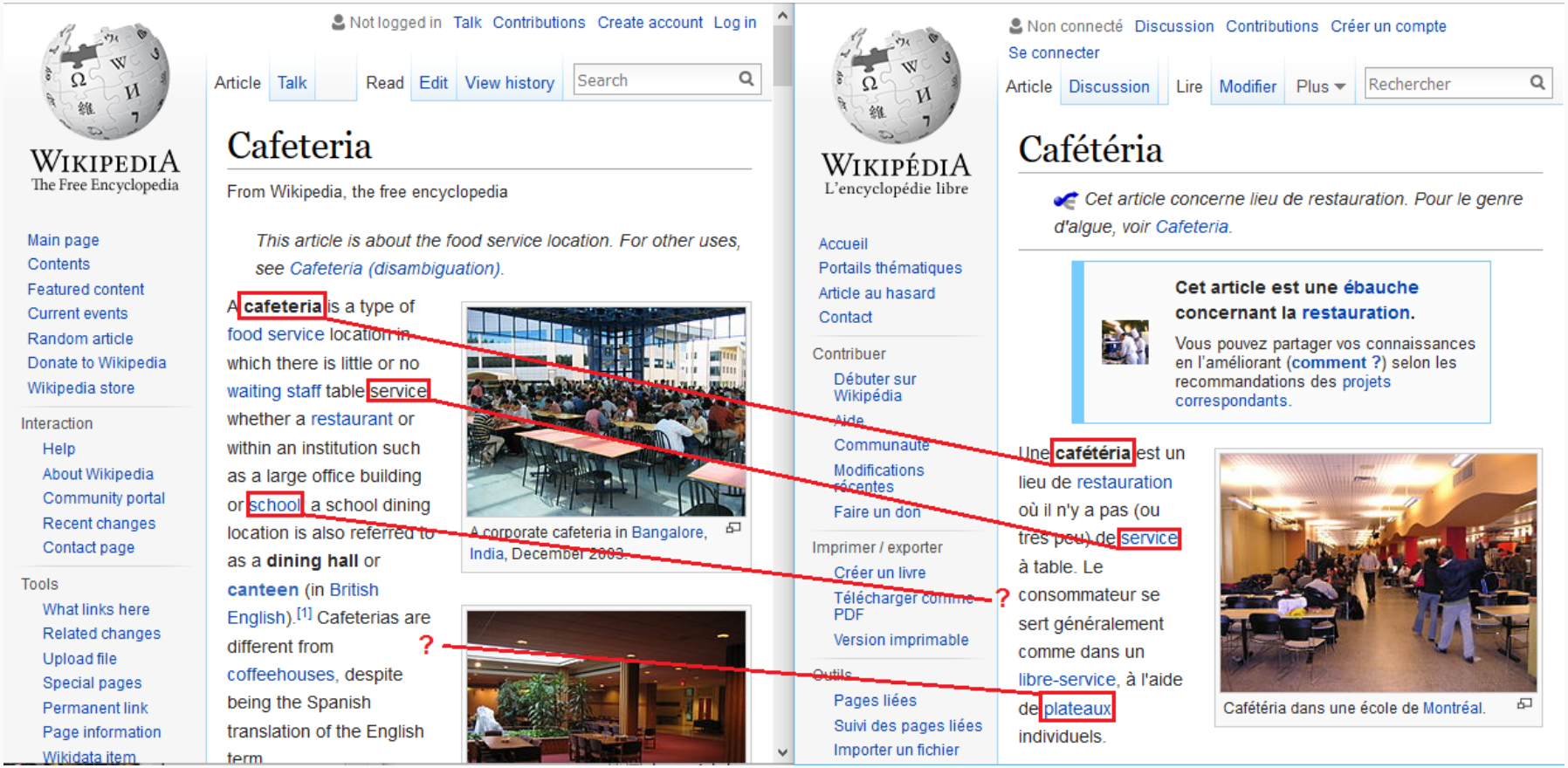
Accueil
Portails thématiques
Article au hasard
Contact

Contribuer
Débuter sur Wikipédia
Aide
Communauté
Modifications récentes
Faire un don

Imprimer / exporter
Créer un livre
Télécharger comme PDF
Version imprimable

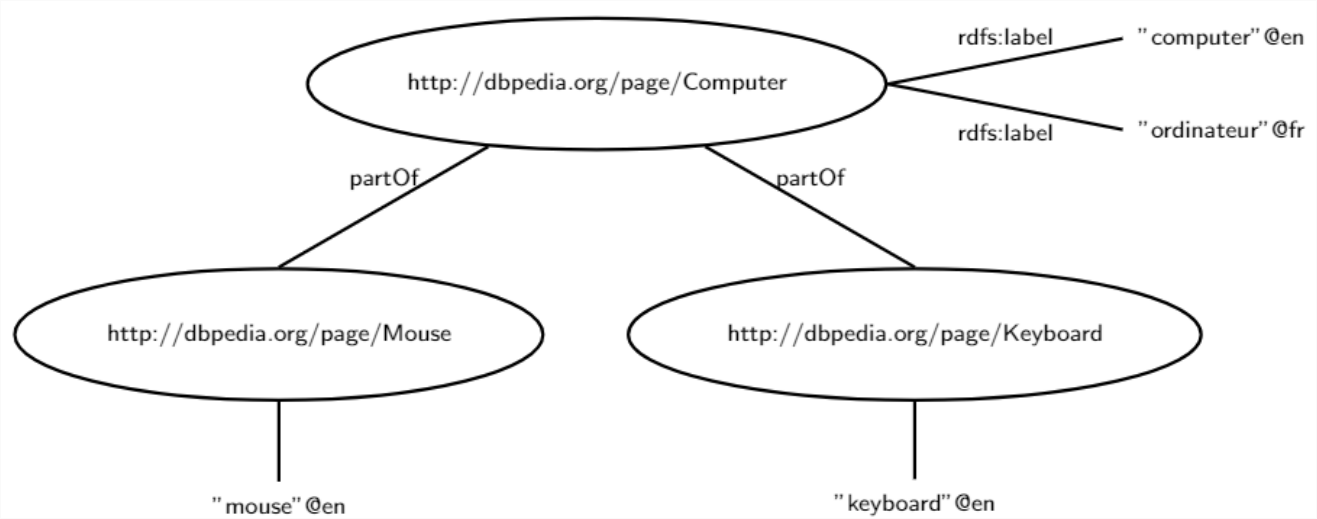
Outils
Pages liées
Suivi des pages liées
Importer un fichier

Corpus Comparables : Exemple



Contexte de recherche prédoc

Objectif: utilisation de l'**ILB** pour le Web Sémantique et **DBpedia**...



rdfs:label : Chaine de caractère "human-readable"

⇒ majoritairement en anglais (Gómez-Pérez et al., 2013)

⇒ En Français ? Seulement 1 concept sur 5 (20%)[1]

[1] wiki.dbpedia.org/Datasets/DatasetStatistics

Contexte de recherche prédoc

Solution (possible) : ”Induire” un label français pour les concepts de DBpedia, par identification d’une traduction au label anglais

- Concepts de **DBpedia** = Titres d’articles de **Wikipédia** anglais
- Pas de label français = pas de lien inter-langues dans **Wikipédia**

n-gram	Distribution [%]
1-gram	158 901 [30.44]
2-gram	185 975 [**35.63**]
3 – 4-gram	141 872 [27.16]
6 – 9-gram	26 855 [5.14]
10 – 28-gram	8 292 [1.58]

Distribution du nombre de mots qui composent les titres de Wikipedia

Corpus Parallèles / Corpus Comparables

...à partir de Corpus Parallèles

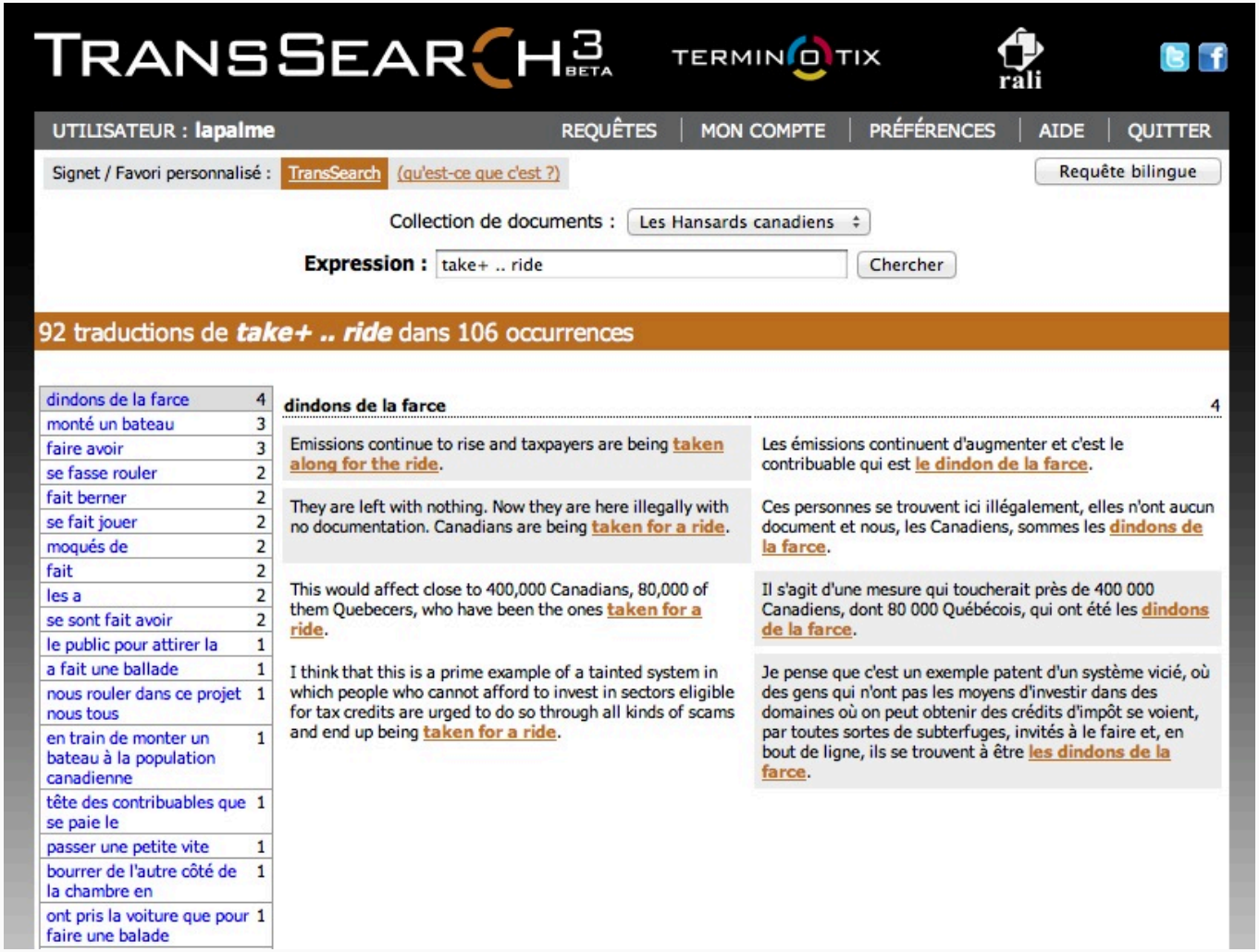
via les modèles IBM d'alignement de mots

ILB via les modèles IBM d'alignement de mots (Traduction Automatique Statistique) (Brown et al., 1990)

Exemple d'application en production: TransSearch[2]

[2] Développé au RALI.

TransSearch: Identificateur de traductions (en CP)



(Bourdaillet et al., 2010) : Précision de 78.3% (Transpot)

⇒ Évaluer la couverture du vocabulaire de nos ressources parallèles

Couverture de nos ressources parallèles - Référence

- = Liste de paires de titres (anglais; français) de **Wikipédia**
  - En relation de traduction = relié par lien inter-langue
  - Taille après filtres Entités nommées (et ponctuations) $\approx$ 200.000

Exemples

Greeting	Salutation
Activities of daily living	Acte de la vie quotidienne
Boston Port Act	Acte du port de Boston
Layout Versus Schematic	Logiciel de vérification de schéma
Output gap	Écart de production
South Sulawesi languages	Langues sulawesi du Sud

Couverture de nos ressources parallèles - Données

Corpus	$\#Paires^{Phrases}$	$\#Mots^{en}$	$\#Mots^{fr}$
	(en milliers)		
HEALTH	809	133	142
EUROPARL	2 007	132	142
HANSARDS	8 802	242	253
GIGAWORD	22 520	3 002	2 775

Couverture de nos ressources parallèles - Approche

GigaWord^{en}	GigaWord^{fr}
21677278: *French and English* greeting *translation cards* [...]	21677278: *Des fiches de* salutation *en anglais et en français* [...]

Exemple de phrases parallèles dans le corpus GigaWord respectant la contrainte d'alignement pour le terme *greeting* et sa traduction *salutation*.

Couverture de nos ressources parallèles - Résultats

	Hansards	Europarl	Health	GigaWord
n	%	%	%	%
1-4	1.1	0.9	1.2	4.0
5-10	0.4	0.3	0.4	1.5
11-100	0.7	0.5	0.6	2.9
>100	0.6	0.3	0.2	2.0
0	97.4	98.0	97.6	89.5

Couverture de nos ressources parallèles - Résultats

	Hansards	Europarl	Health	GigaWord
n	%	%	%	%
1-4	1.1	0.9	1.2	4.0
5-10	0.4	0.3	0.4	1.5
11-100	0.7	0.5	0.6	2.9
>100	0.6	0.3	0.2	2.0
0	97.4	98.0	97.6	89.5

Couverture de nos ressources parallèles - Résultats

	Hansards	Europarl	Health	GigaWord
n	%	%	%	%
1-4	1.1	0.9	1.2	4.0
5-10	0.4	0.3	0.4	1.5
11-100	0.7	0.5	0.6	2.9
>100	0.6	0.3	0.2	2.0
0	97.4	98.0	97.6	89.5

Couverture de nos ressources parallèles - Conclusion

$\Rightarrow$ Couverture insuffisante de nos corpus parallèles pour l'**ILB**

pistes de solution

- **ILB** sur Corpus Comparables : suite de la présentation
- Plus de données parallèles :

Article n°3

Jakubina, L. and Langlais, P. (2016a). Bad luc@wmt 2016: a bilingual document alignment platform based on lucene.

In *Proceedings of the First Conference on Machine Translation*, pages 703–709, Berlin, Germany. Association for Computational Linguistics

...à partir de Corpus Comparables

via l’alignement de mots en Corpus Comparables

Alignement par motifs de cooccurrence

ACL 1995 : (Rapp, 1995) et (Fung, 1995)

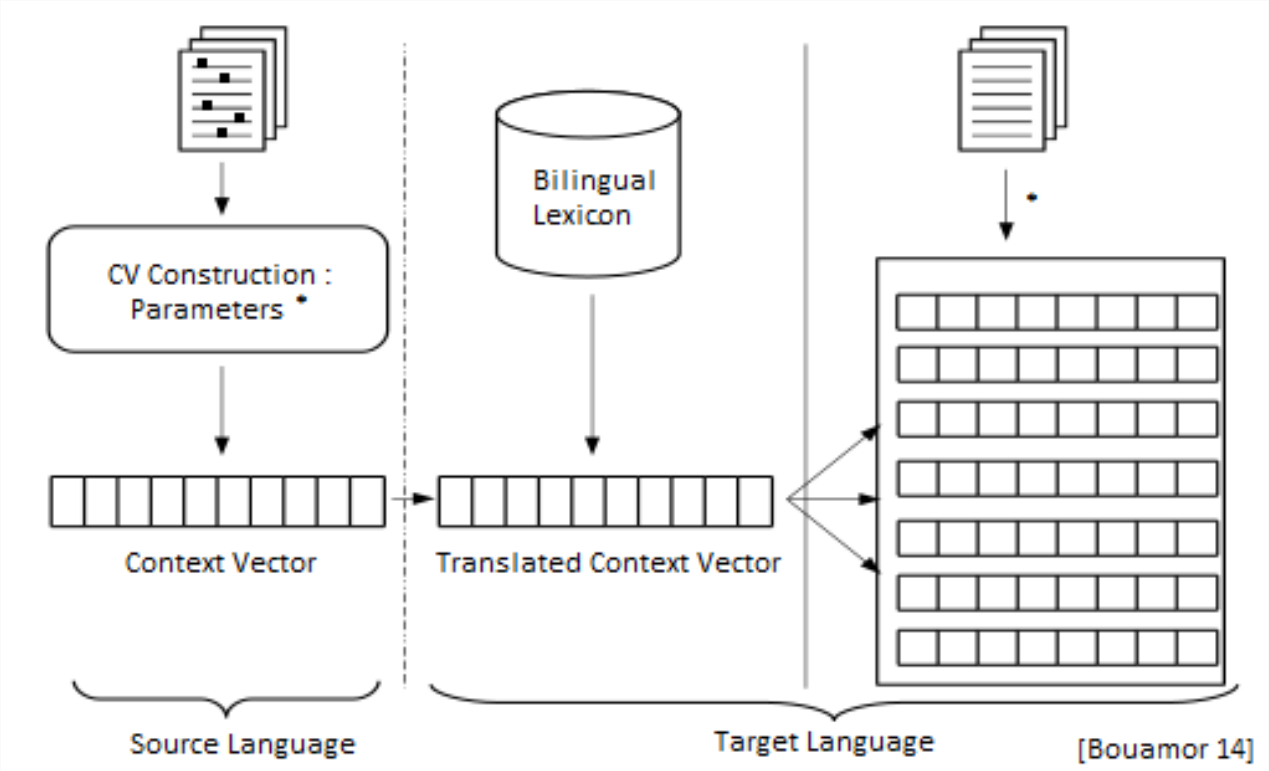
Hypothèse émise par (Rapp, 1995)

Si deux mots (A et B) cooccurrent plus souvent que par le hasard dans une langue source, alors leur traduction respective (A′ et B′) doivent cooccurrer plus souvent que par le hasard dans la langue cible.

⇒ Les mots en relations de traductions partagent des motifs de cooccurrence similaires

- Réalise une expérience de simulation pour le prouver
- Approche impraticable point de vue computationnel

(Rapp, 1999) : Approche Standard / RAPP



Étapes de l'**Approche Standard**

A attiré l'attention d'un grand nombre de chercheurs et de ce fait, a bénéficié de beaucoup d'améliorations (Sharoff et al., 2013)

RAPP et Cie. - Observation(s)

Échantillonage: espace de construction/recherche/voc réduit

Exemples

- (Vulić et al., 2011): "...only lemmatized noun forms ; voc ≈ 8.500"
- (Morin and Prochasson, 2011): "...words occurring less than twice discarded..."
- (Gaussier et al., 2004): "...frequency less than 5 discarded ; voc ≈ 40.000"

⇒ Quid d'une "mise en production", à l'ère du **`Big Data`** ?

Article n°1

Jakubina, L. and Langlais, P. (2015). Projective methods for mining missing translations in dbpedia.

In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 23–31, Beijing, China. Association for Computational Linguistics

(Jakubina and Langlais, 2015) - Protocole

Approches

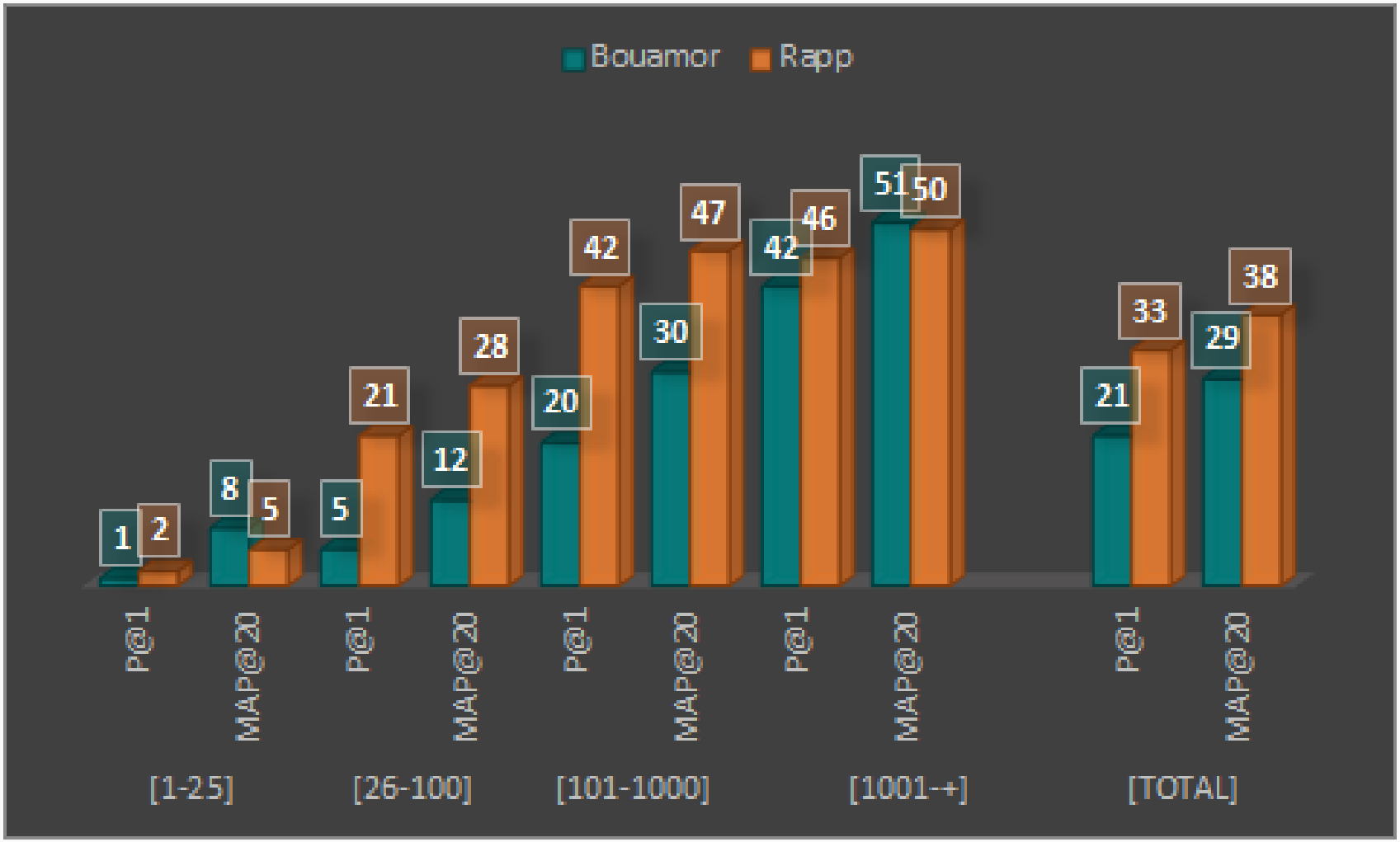
- Reproduction de **RAPP**
  - Implémentée pour le **non**-échantillonage
  - Espace de recherche = 2.942.067 mots cibles (français)
- Reproduction de (Bouamor et al., 2013)
  - Analyse Sémantique Explicite (Gabrilovich and Markovitch, 2009)
  - Supérieure à **RAPP**

Données

- **Wikipédia**[3] Anglais et Français 2013 non-échantillonné
- Liste de référence basée sur les titres 1-gramme de **Wikipédia**
  - Gammes de fréquences (très rare, rare, fréquent, très fréquent)

[3] DBpedia en est le silo RDF.

(Jakubina and Langlais, 2015) - Résultats



(Jakubina and Langlais, 2015) - Impacts

- Biais en faveur des mots fréquents
- Performances **RAPP** $\geq$ (Bouamor et al., 2013)
  - Protocoles différents: corpus spécialisés, pré-traitements, liste de référence ($\approx$ 100), ...
- Implémentation multi-threadée (8-16) avec **Lucene** et similarité vectorielle par morceaux
  - mise-à-mal par les mots hyper-fréquents (ex: France, freq=1.000.000)
  - $\Rightarrow$ échantillonnage toujours nécessaire
- Types d'erreurs (effet thésaurus)

(Jakubina and Langlais, 2015) - Exemples de sortie

RAPP

myringotomy [1-25]	permette	devra	scopie	nécessitait	...
syllabication [26-100]	modifier	suffit	vouloir	syllabique	...
numerology [101-1000]	numérologie	occultisme	ésotérisme	divinatoire	...
entertainment [1001+]	beatmakers	manglobe	spycraft	déduplication	...

(Bouamor et al., 2013)

myringotomy	laryngologie	oto	rhino	traitement	otite
syllabication	langues	consonne	langue	lettre	phonétique
numerology	oeuvre	gematria	angels	nombres	chire
entertainment	entertainment	divertissement	vidéo	sony	jeu

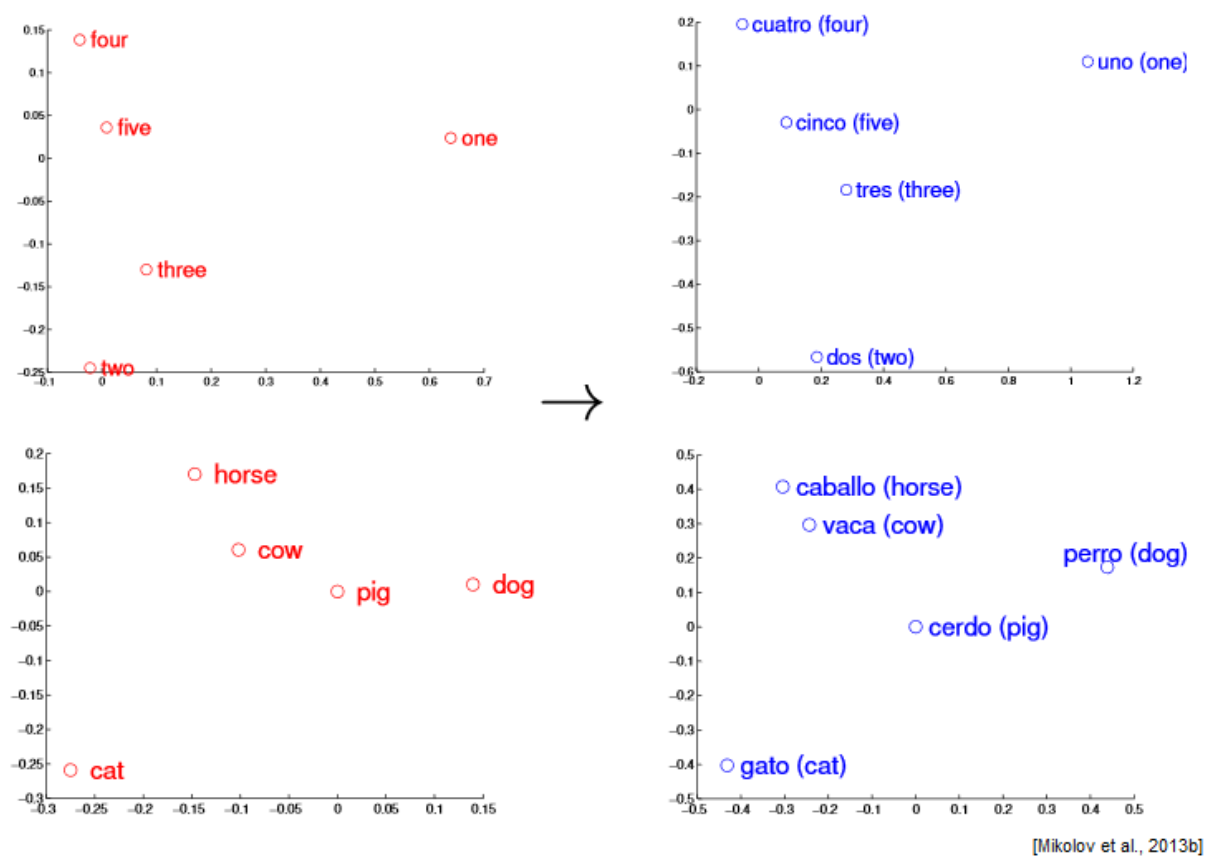
Idée: Combiner les listes de candidats[4] ?

[4] Potentielle traduction pour 431 termes sur 528 (81.6%) [100+...]

Comme tâche d’évaluation des embeddings interlingues

MIKO (Mikolov et al., 2013b) : ILB comme tâche d'évaluation

Nous observons des similarités dans la projection géométriques des embeddings de mots provenant de différentes langues.



[Mikolov et al., 2013b]

MIKO (Mikolov et al., 2013b) : Observation(s)

- Pas de positionnement vis-à-vis de **RAPP**
- Quid des **Mots Rares** ?
- Échantillonnage
  - "...vocabularies consist of the words that occurred at least five times in the corpus." + "...short phrases..."
  - Espace de recherche de **MIKO** $\approx$ 110.000 mots

Article n°2

Jakubina, L. and Langlais, P. (2016b). A comparison of methods for identifying the translation of words in a comparable corpus: Recipes and limits.

CICLing: International Conference on Computational Linguistics and Intelligent Text Processing

(Jakubina and Langlais, 2016b) - Protocole

Approches

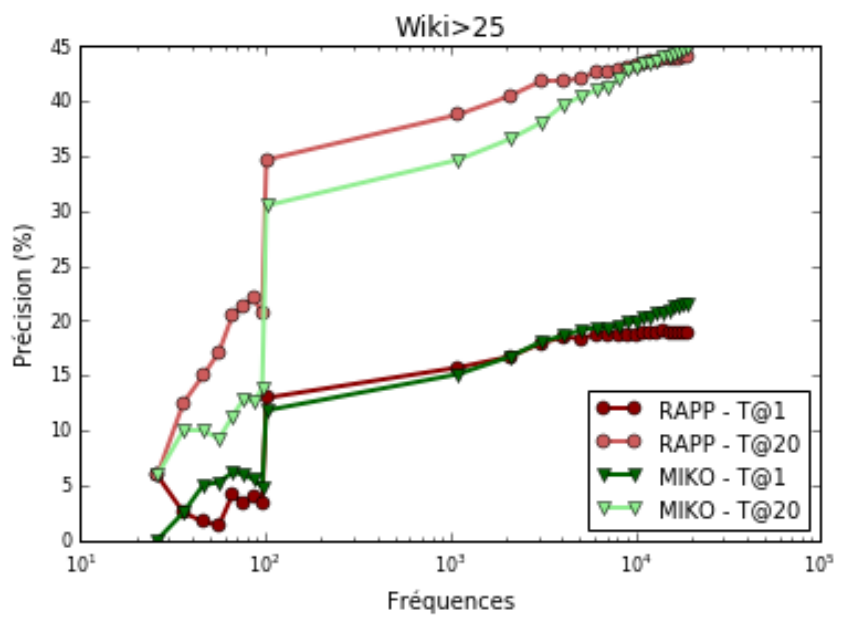
- **RAPP** (50 variantes[5])
- (Prochasson and Fung, 2011) (2 variantes)
  - Proposé pour gérer les **Mots Rares**
- **MIKO** (Mikolov et al., 2013b) (130 variantes)
  - Construction avec **Word2Vec** (Mikolov et al., 2013a)
  - Projection (Dinu and Baroni, 2014)

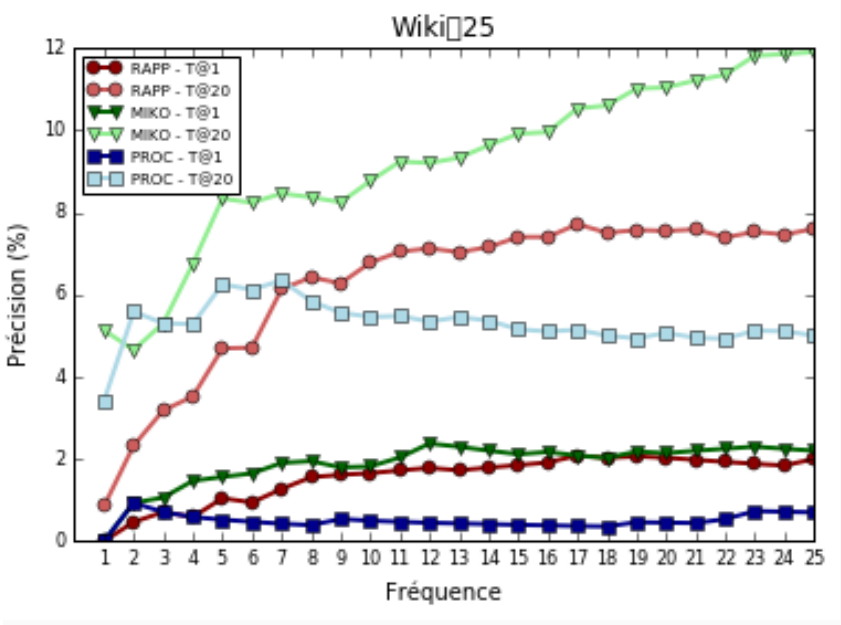
Données

- **Wikipédia** anglais et français 2013 non-échantillonné
- 2 Listes de référence: $Wiki_{>25}$ et $Wiki_{\leqslant 25}$ [titres $\Rightarrow$ mots]
  - 6.8 millions des mots (92%) de **Wikipédia** anglais apparaissent moins de 26 fois

[5] selon les valeurs des méta-paramètres ; Syn: signaux.

(Jakubina and Langlais, 2016b) - Résultats





RAPP

- Taille Fenêtre: 3
- Normalisation: PMI

MIKO

- Modèle CBOW
- Taille Fenêtre: 5

RAPP

- Taille Fenêtre: 15
- Normalisation: ORD

MIKO

- Modèle Skip-Gram
- Taille Fenêtre: 10

(Jakubina and Langlais, 2016b) - Impacts

- Performances **RAPP** ≈ Performances **MIKO** (biais de fréquence)
  - (Levy et al., 2015) prouve que les modèles sont proches
- "Recettes" : différents jeux de méta-paramètres selon la fréquence du terme à traiter
  - Meilleures performances sur les **Mots Rares**
- (Prochasson and Fung, 2011) : Résultats décevants
  - Problème des **Mots Rares** reste ouvert
- Complémentarité des approches validée (combinaison-oracle)
  - Performances doublées sur les **Mots Rares**

(Jakubina and Langlais, 2016b) - Exemples de sortie

Candidats corrects mais pas dans notre liste de référence

RAPP

donut (beigne)	aromatisé	donut	beignet
brilliantly (brillamment)	imaginatif	captivant	rusé
gentle (doux)	enjoué	serviable	affable
pathologically (pathologiquement)	cordonale	pathologique	diagnostiqué

MIKO

donut (beigne)	liper	babalous	savonette
brilliantly (brillamment)	éclatant	pathétique	émouvant
gentle (doux)	colérique	enjoué	espiègle
pathologically (pathologiquement)	psychosexuel	psychoaectif	piloerection

Idée: Approche basée sur la combinaison (pour les **Mots Rares**) ?

via reclassement supervisé de candidats à la traduction

RERANKER : Reclassement supervisé après Combinaison

Weighting [EN] : pondération [FR]					
Rapp		Miko		Faru	
maximise	0.069	**pondération**	**0.753**	**pondération**	**0.428**
quantifié	0.066	*pondéré*	*0.688*	calculé	0.389
adaptatif	0.065	dépendra	0.648	quantification	0.387
pondération	**0.066**	calculé	0.622	luminance	0.336
pondéré	*0.063*	maximise	0.62	*pondéré*	*0.319*
calculé	0.054	quantifié	0.58	quantifié	0.307
dépendra	0.054	luminance	0.567	maximise	0.268
liquidité	0.052	dénominateu	0.55	inférieure	0.267

Reclassement (avec features) après Combinaison

Résultats : **pondération (0.020)** ; maximise (0.0027) ;
pondéré (0.0021) ; quantifié (0,0019) ; calculé (0,0010) ; …

RERANKER : Observation(s)

- Axe de recherche prometteur mais peu exploré (Sharoff et al., 2013; Irvine and Callison-Burch, 2013)
- **Objectif :** Améliorer l'**`ILB`** sur les **`Mots Rares`**

Article n°4

Jakubina, L. and Langlais, P. (2017). Reranking translation candidates produced by several bilingual word similarity sources.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 605–611, Valencia, Spain. Association for Computational Linguistics

(Jakubina and Langlais, 2017) - Protocole

Approche

- **RERANKER** avec la librairie LTR **RANKLIB**[6]
  - 3 ensembles de features: rang (2), string (5), fréquence (4)
  - 700 entrées pour s'entraîner ; 300 pour tester

Données

- 3 Listes d'**ILB**: **RAPP**, **MIKO** et **FARU** (Faruqui and Dyer, 2014)
  - Listes adaptés selon la fréquence des mots sources[7]
  - 100 candidats par liste
- 3 Listes de référence: $Wiki_{>25}$, $Wiki_{\leqslant 25}$ et $Euro_{5-6k}$[8]

[6] https://sourceforge.net/p/lemur/wiki/RankLib/
[7] "Recettes" de l'article n°2.
[8] EUROPARL : Débats parlementaires européens.

(Jakubina and Langlais, 2017) - Résultats

	Individuel			Reclassé				Combiné & Reclassé		
	@1	@5	@20	@1	@5	@20		@1	@5	@20
Wiki$_{>25}$										oracle: 69.3
RAPP	20.0	33.0	43.0	$36.3^{2.5}$	$48.8^{1.9}$	$53.8^{1.9}$				
MIKO	17.0	32.6	41.6	$38.1^{1.9}$	$49.0^{1.5}$	$54.3^{1.3}$	R+M+F	**45.6**$^{2.2}$	**59.6**$^{1.1}$	**64.0**$^{1.8}$
FARU	13.3	26.0	33.3	$34.3^{1.5}$	$44.0^{2.6}$	$47.9^{2.1}$				
Wiki$_{\leqslant 25}$										oracle: 28.6
RAPP	2.6	4.3	7.3	$8.6^{1.2}$	$9.4^{0.8}$	$10.2^{1.0}$				
MIKO	1.6	4.6	10.6	$16.6^{2.2}$	$19.0^{1.5}$	$20.1^{1.4}$	R+M+F	**21.3**$^{1.9}$	**24.4**$^{1.7}$	**25.7**$^{1.9}$
FARU	1.6	2.6	5.0	$7.9^{2.2}$	$8.7^{2.5}$	$8.9^{2.7}$				
Euro$_{5-6k}$										oracle: 84.4
RAPP	16.6	31.8	41.2	$34.6^{5.7}$	$48.6^{1.2}$	$51.9^{1.2}$				
MIKO	42.0	59.0	67.8	$47.0^{2.3}$	$68.1^{2.7}$	$73.0^{1.7}$	R+M	**49.5**$^{3.7}$	**68.7**$^{1.5}$	$76.1^{1.0}$
FARU	30.6	47.7	59.8	$41.2^{3.9}$	$58.0^{3.5}$	$66.0^{3.5}$				

(Jakubina and Langlais, 2017) - Résultats

	Individuel			Reclassé				Combiné & Reclassé		
	@1	@5	@20	@1	@5	@20		@1	@5	@20
Wiki$_{>25}$										oracle: 69.3
RAPP	20.0	33.0	43.0	$36.3^{2.5}$	$48.8^{1.9}$	$53.8^{1.9}$				
MIKO	17.0	32.6	41.6	$38.1^{1.9}$	$49.0^{1.5}$	$54.3^{1.3}$	R+M+F	**45.6**$^{2.2}$	**59.6**$^{1.1}$	**64.0**$^{1.8}$
FARU	13.3	26.0	33.3	$34.3^{1.5}$	$44.0^{2.6}$	$47.9^{2.1}$				
Wiki$_{\leqslant 25}$										oracle: 28.6
RAPP	2.6	4.3	7.3	$8.6^{1.2}$	$9.4^{0.8}$	$10.2^{1.0}$				
MIKO	1.6	4.6	10.6	$16.6^{2.2}$	$19.0^{1.5}$	$20.1^{1.4}$	R+M+F	**21.3**$^{1.9}$	**24.4**$^{1.7}$	**25.7**$^{1.9}$
FARU	1.6	2.6	5.0	$7.9^{2.2}$	$8.7^{2.5}$	$8.9^{2.7}$				
Euro$_{5-6k}$										oracle: 84.4
RAPP	16.6	31.8	41.2	$34.6^{5.7}$	$48.6^{1.2}$	$51.9^{1.2}$				
MIKO	42.0	59.0	67.8	$47.0^{2.3}$	$68.1^{2.7}$	$73.0^{1.7}$	R+M	**49.5**$^{3.7}$	**68.7**$^{1.5}$	$76.1^{1.0}$
FARU	30.6	47.7	59.8	$41.2^{3.9}$	$58.0^{3.5}$	$66.0^{3.5}$				

(Jakubina and Langlais, 2017) - Résultats

	Individuel			Reclassé				Combiné & Reclassé		
	@1	@5	@20	@1	@5	@20		@1	@5	@20
Wiki$_{>25}$										oracle: 69.3
RAPP	20.0	33.0	43.0	$36.3^{2.5}$	$48.8^{1.9}$	$53.8^{1.9}$				
MIKO	17.0	32.6	41.6	$38.1^{1.9}$	$49.0^{1.5}$	$54.3^{1.3}$	R+M+F	**45.6**$^{2.2}$	**59.6**$^{1.1}$	**64.0**$^{1.8}$
FARU	13.3	26.0	33.3	$34.3^{1.5}$	$44.0^{2.6}$	$47.9^{2.1}$				
Wiki$_{\leqslant 25}$										oracle: 28.6
RAPP	2.6	4.3	7.3	$8.6^{1.2}$	$9.4^{0.8}$	$10.2^{1.0}$				
MIKO	1.6	4.6	10.6	$16.6^{2.2}$	$19.0^{1.5}$	$20.1^{1.4}$	R+M+F	**21.3**$^{1.9}$	**24.4**$^{1.7}$	**25.7**$^{1.9}$
FARU	1.6	2.6	5.0	$7.9^{2.2}$	$8.7^{2.5}$	$8.9^{2.7}$				
Euro$_{5-6k}$										oracle: 84.4
RAPP	16.6	31.8	41.2	$34.6^{5.7}$	$48.6^{1.2}$	$51.9^{1.2}$				
MIKO	42.0	59.0	67.8	$47.0^{2.3}$	$68.1^{2.7}$	$73.0^{1.7}$	R+M	**49.5**$^{3.7}$	**68.7**$^{1.5}$	$76.1^{1.0}$
FARU	30.6	47.7	59.8	$41.2^{3.9}$	$58.0^{3.5}$	$66.0^{3.5}$				

(Jakubina and Langlais, 2017) - Résultats

	Individuel			Reclassé				Combiné & Reclassé		
	@1	@5	@20	@1	@5	@20		@1	@5	@20
Wiki$_{>25}$										oracle: 69.3
RAPP	20.0	33.0	43.0	$36.3^{2.5}$	$48.8^{1.9}$	$53.8^{1.9}$				
MIKO	17.0	32.6	41.6	$38.1^{1.9}$	$49.0^{1.5}$	$54.3^{1.3}$	R+M+F	**45.6**$^{2.2}$	**59.6**$^{1.1}$	**64.0**$^{1.8}$
FARU	13.3	26.0	33.3	$34.3^{1.5}$	$44.0^{2.6}$	$47.9^{2.1}$				
Wiki$_{\leqslant 25}$										oracle: 28.6
RAPP	2.6	4.3	7.3	$8.6^{1.2}$	$9.4^{0.8}$	$10.2^{1.0}$				
MIKO	1.6	4.6	10.6	$16.6^{2.2}$	$19.0^{1.5}$	$20.1^{1.4}$	R+M+F	**21.3**$^{1.9}$	**24.4**$^{1.7}$	**25.7**$^{1.9}$
FARU	1.6	2.6	5.0	$7.9^{2.2}$	$8.7^{2.5}$	$8.9^{2.7}$				
Euro$_{5-6k}$										oracle: 84.4
RAPP	16.6	31.8	41.2	$34.6^{5.7}$	$48.6^{1.2}$	$51.9^{1.2}$				
MIKO	42.0	59.0	67.8	$47.0^{2.3}$	$68.1^{2.7}$	$73.0^{1.7}$	R+M	**49.5**$^{3.7}$	**68.7**$^{1.5}$	$76.1^{1.0}$
FARU	30.6	47.7	59.8	$41.2^{3.9}$	$58.0^{3.5}$	$66.0^{3.5}$				

(Jakubina and Langlais, 2017) - Résultats

	Individuel			Reclassé				Combiné & Reclassé		
	@1	@5	@20	@1	@5	@20		@1	@5	@20
Wiki$_{>25}$										oracle: 69.3
RAPP	20.0	33.0	43.0	$36.3^{2.5}$	$48.8^{1.9}$	$53.8^{1.9}$				
MIKO	17.0	32.6	41.6	$38.1^{1.9}$	$49.0^{1.5}$	$54.3^{1.3}$	R+M+F	**45.6**$^{2.2}$	**59.6**$^{1.1}$	**64.0**$^{1.8}$
FARU	13.3	26.0	33.3	$34.3^{1.5}$	$44.0^{2.6}$	$47.9^{2.1}$				
Wiki$_{\leqslant 25}$										oracle: 28.6
RAPP	2.6	4.3	7.3	$8.6^{1.2}$	$9.4^{0.8}$	$10.2^{1.0}$				
MIKO	1.6	4.6	10.6	$16.6^{2.2}$	$19.0^{1.5}$	$20.1^{1.4}$	R+M+F	**21.3**$^{1.9}$	**24.4**$^{1.7}$	**25.7**$^{1.9}$
FARU	1.6	2.6	5.0	$7.9^{2.2}$	$8.7^{2.5}$	$8.9^{2.7}$				
Euro$_{5-6k}$										oracle: 84.4
RAPP	16.6	31.8	41.2	$34.6^{5.7}$	$48.6^{1.2}$	$51.9^{1.2}$				
MIKO	42.0	59.0	67.8	$47.0^{2.3}$	$68.1^{2.7}$	$73.0^{1.7}$	R+M	**49.5**$^{3.7}$	**68.7**$^{1.5}$	$76.1^{1.0}$
FARU	30.6	47.7	59.8	$41.2^{3.9}$	$58.0^{3.5}$	$66.0^{3.5}$				

(Jakubina and Langlais, 2017) - Résultats

	Individuel			Reclassé			Combiné & Reclassé			
	@1	@5	@20	@1	@5	@20		@1	@5	@20
Wiki$_{>25}$										oracle: 69.3
RAPP	20.0	33.0	43.0	$36.3^{2.5}$	$48.8^{1.9}$	$53.8^{1.9}$				
MIKO	17.0	32.6	41.6	$38.1^{1.9}$	$49.0^{1.5}$	$54.3^{1.3}$	R+M+F	**45.6**$^{2.2}$	**59.6**$^{1.1}$	**64.0**$^{1.8}$
FARU	13.3	26.0	33.3	$34.3^{1.5}$	$44.0^{2.6}$	$47.9^{2.1}$				
Wiki$_{\leqslant 25}$										oracle: 28.6
RAPP	2.6	4.3	7.3	$8.6^{1.2}$	$9.4^{0.8}$	$10.2^{1.0}$				
MIKO	1.6	4.6	10.6	$16.6^{2.2}$	$19.0^{1.5}$	$20.1^{1.4}$	R+M+F	**21.3**$^{1.9}$	**24.4**$^{1.7}$	**25.7**$^{1.9}$
FARU	1.6	2.6	5.0	$7.9^{2.2}$	$8.7^{2.5}$	$8.9^{2.7}$				
Euro$_{5-6k}$										oracle: 84.4
RAPP	16.6	31.8	41.2	$34.6^{5.7}$	$48.6^{1.2}$	$51.9^{1.2}$				
MIKO	42.0	59.0	67.8	$47.0^{2.3}$	$68.1^{2.7}$	$73.0^{1.7}$	R+M	**49.5**$^{3.7}$	**68.7**$^{1.5}$	$76.1^{1.0}$
FARU	30.6	47.7	59.8	$41.2^{3.9}$	$58.0^{3.5}$	$66.0^{3.5}$				

(Jakubina and Langlais, 2017) - Résultats

	Individuel			Reclassé			Combiné & Reclassé			
	@1	@5	@20	@1	@5	@20		@1	@5	@20
Wiki$_{>25}$										oracle: 69.3
RAPP	20.0	33.0	43.0	36.3$^{2.5}$	48.8$^{1.9}$	53.8$^{1.9}$				
MIKO	17.0	32.6	41.6	38.1$^{1.9}$	49.0$^{1.5}$	54.3$^{1.3}$	R+M+F	**45.6**$^{2.2}$	**59.6**$^{1.1}$	**64.0**$^{1.8}$
FARU	13.3	26.0	33.3	34.3$^{1.5}$	44.0$^{2.6}$	47.9$^{2.1}$				
Wiki$_{\leqslant 25}$										oracle: 28.6
RAPP	2.6	4.3	7.3	8.6$^{1.2}$	9.4$^{0.8}$	10.2$^{1.0}$				
MIKO	1.6	4.6	10.6	16.6$^{2.2}$	19.0$^{1.5}$	20.1$^{1.4}$	R+M+F	**21.3**$^{1.9}$	**24.4**$^{1.7}$	**25.7**$^{1.9}$
FARU	1.6	2.6	5.0	7.9$^{2.2}$	8.7$^{2.5}$	8.9$^{2.7}$				
Euro$_{5-6k}$										oracle: 84.4
RAPP	16.6	31.8	41.2	34.6$^{5.7}$	48.6$^{1.2}$	51.9$^{1.2}$				
MIKO	42.0	59.0	67.8	47.0$^{2.3}$	68.1$^{2.7}$	73.0$^{1.7}$	R+M	**49.5**$^{3.7}$	**68.7**$^{1.5}$	76.1$^{1.0}$
FARU	30.6	47.7	59.8	41.2$^{3.9}$	58.0$^{3.5}$	66.0$^{3.5}$				

(Jakubina and Langlais, 2017) - Impacts

- **FARU** : performances similaires à **RAPP** et **MIKO** (biais de fréquence)
- Reclassement : ✓
- Combinaison (avec reclassement) : ✓
  - Améliorations de 15% en moyenne
  - Notamment, sur les **Mots Rares**
  - Avec peu de features
- Analyses complémentaires
  - Taille données d'entraînements
  - Impacts du reclassement et des features
  - Analyse d'erreurs plus détaillée[9]

[9] Morpho ; Relié (syno, anto, hypo, ...) ; etc.

(Jakubina and Langlais, 2017) - Exemples de sorties

Wiki$_{>25}$ - **RERANKER**

donut (beigne)	donuts	donut	grillé	baozi
brilliantly (brillamment)	brillant	éclatant	éblouissant	imaginatif
gentle (doux)	colérique	enjoué	insouciant	affable
pathologically (pathologiquement)	pathologique	sigmoïdite	fibromatose	nosophobie

Wiki$_{\leqslant 25}$

raillery	raillerie				
RERANKER	raillerie	railler	raille	railleries	1
RAPP	pourra	désire	...	raillerie	5
MIKO	reprocher	naïveté	...	raillerie	70
FARU	intercommunal	plaidoiries	fourrière	plaida	∅

schizoidism	schizoïdie				
RERANKER	schizoïdie	musicothérapeute	confusionnels	mentalisation	1
RAPP	refactorisat.	disiciples	knowlse	boutetage	∅
MIKO	atrophier	hypomania.	psychoneurol.	neurochimiques	∅
FARU	hypomania.	psychoneurol.	...	schizoïdie	65

(Jakubina and Langlais, 2017) - Exemples de sorties

Wiki$_{\leqslant 25}$

botanize	herboriser				
RERANKER	herborise	herboriser	prodigua	attardant	2
RAPP	aoutourou	délai	...	herboriser	6
MIKO	attardant	herboriser	prodigua	hastrel	2
FARU	compromissions	lorois	caboteur	pontée	∅

uncrushable	infroissable				
RERANKER	infroissable	incrustait	raquettistes	paludicroque	1
RAPP	senente	imbroyables	pulvérix	attriteur	∅
MIKO	nattées	paumage	infroissable	mouillettes	3
FARU	roninson	ospovat	talanov	mouraviova	∅

brotherliness	confraternité				
RERANKER	volonté	précepte	fraternel	désintéressé	3
RAPP	qoudous	tâche	...	fraternel	44
MIKO	joie	volonté	...	fraternel	68
FARU	observant	cueillies	concordent	moisson	∅

Travaux futurs & Conclusion

Axes de travaux futurs

- Combinaison - Reclassement
  - Reproduction sur d'autres paires de langues
  - Autres features / approches (Irvine and Callison-Burch, 2013)[10]
- Améliorations modèles / approches
  - Instanciation du modèle approprié (selon la fréquence)
  - Quid d'un modèle incluant les features de la combinaison ?
- Mots rares
  - Modèle(s) d'embeddings adapté(s)
- Réflexion méta sur l'**ILB**
  - Métriques d'évaluation
  - Échantillonnage(s) approprié(s) / intelligent(s)
- Retour sur... DBpedia ?

[10] contexte, temporel, orthographe, thématique et fréquence.

Reclassement - Autres paires de langues

Déjà réalisé[11] : Allemand, Espagnol et Roumain (vs Anglais)

⇒ Approche stable sur d'autres paires de langues

[11] Publication en cours de rédaction.

Mots rares - Modèle(s) d'embeddings adapté(s)

Améliorer les modèles des mots rares en combinant l'information de contexte avec d'autres informations:

- Rothe, S. and Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes
- Avraham, O. and Goldberg, Y. (2017). The interplay of semantics and morphology in word embeddings
- Pilehvar, M. T. and Collier, N. (2017). Inducing embeddings for rare and unseen words by leveraging lexical resources
- Bamler, R. and Mandt, S. (2017). Dynamic word embeddings
  - Embeddings intègrent le(s) sens du mot "dans le temps"
- Calixto, I., Liu, Q., and Campbell, N. (2017). Multilingual multi-modal embeddings for natural language processing
  - Embeddings améliorés en apprenant avec du texte **ET** des images

Possible aussi pour les mots fréquents

Conclusion

- Tâche d'Induction de Lexiques Bilingues sur deux types de corpus bilingues
- Corpus parallèles (rare): approche efficace mais couverture insuffisante
- Corpus comparables (fréquent): modèles basé sur le contexte
  - Performances convaincantes sur les mots fréquents
  - Trouvent leurs limites avec les mots rares
- Améliorations satisfaisantes de l'**ILB** grâce à la combinaison et au reclassement supervisé
- Nombreux axes de travaux futurs ; Intérêts portés vers le reclassement et les mots rares

Merci !

Questions ?

ESA BOUAMOR

- Adaptation de l'Analyse Semantique Explicite (ESA): utilise Wikipedia comme base de connaissance(Gabrilovich and Markovitch, 2007)
- Espace des mots $\rightarrow$ Espace des titres (vecteur)
- Dictionnaire $\rightarrow$ Lien interlangue Wikipedia

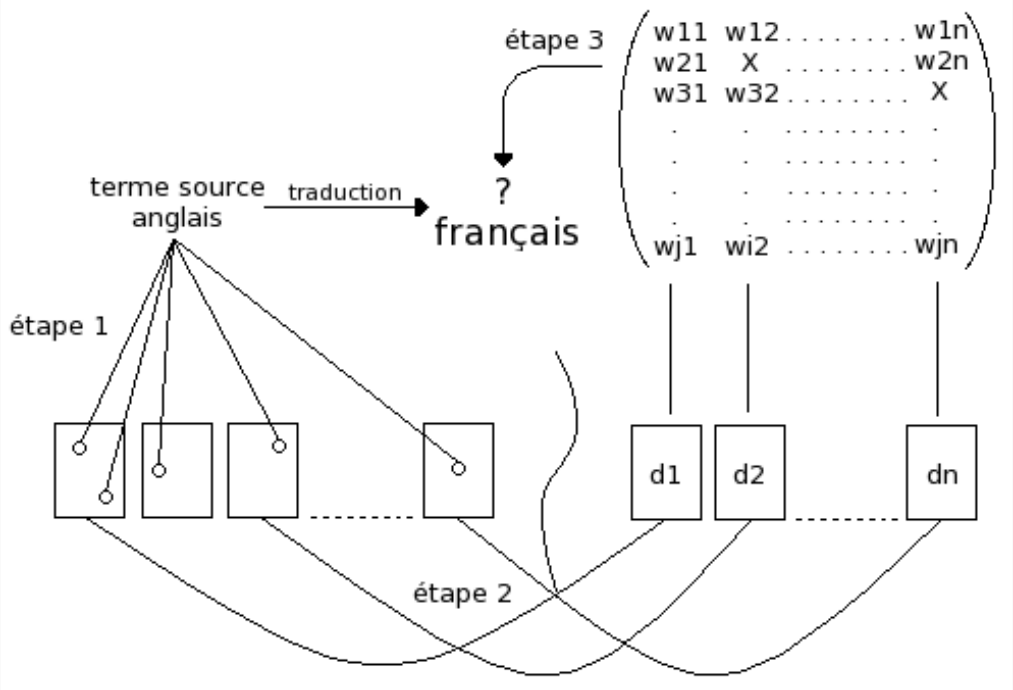


Figure 1: Approche ESA Bouamor

ILB comme tâche d'évaluation

Apprendre une projection des embeddings anglais vers les embeddings français

Embeddings Anglais

...	...	...	...
text	0.22	0.09	...
cafeteria	−0.32	−0.28	...
stereo	1.97	−0.30	...
dummy	0.28	0.24	...
one	−0.36	0.23	...
...	...	...	...

↔

Project° (Dinu and Baroni, 2014)

...	...	...
1.73	4.83	...
−7.21	−3.93	...
5.07	−3.41	...
5.42	8.51	...
1.57	−1.25	...
...	...	...

↔

Embeddings Français

...	...	...	...
texte	0.70	1.56	...
un	0.06	−1.44	...
cafeteria	0.00	0.36	...
ville	1.75	−1.43	...
saint	1.31	−0.03	...
...	...	...	...

(Apprendre **directement** des embeddings dits interlingues / bilingues)

Embeddings Bilingues

...	...	...	...	...	...
text / texte	−0.33	0.17	−0.43	0.92	...
cafeteria / cafeteria	−0.63	0.45	−0.09	−0.28	...
stereo / stereo	−0.32	0.04	−0.65	0.08	...
dummy / factice	−0.12	−0.27	0.03	0.26	...
one / un	−0.04	−0.12	0.05	−0.47	...
...	...	...	...	...	...

FARU (Faruqui and Dyer, 2014), BilBOWA (Gouws et al., 2015), (Coulmance et al., 2015; Ammar et al., 2016; Vulic et al., 2016)

Resultats Article n°2

- Best variant for each approach according to TOP@1.
- An *Oracle* shows that approaches are complementary.
- Disappointment for the poor performance of the document approach which was specifically designed to handle rare words.

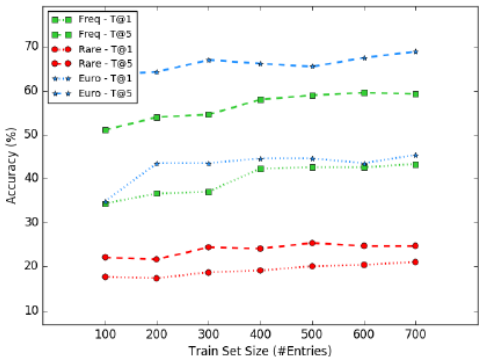
	$Wiki_{\leqslant 25}$			$Wiki_{>25}$		
	@1	@5	@20	@1	@5	@20
MIKO	**2,2**	**6,1**	**11,9**	**21,7**	**34,2**	**44,9**
RAPP	2,0	4,3	7,6	19,0	32,7	44,3
Doc	0,7	2,3	5,0	*10,0*	*19.0*	*24.0*
oracle	4,6	10,5	19,0	31,8	46,8	57,6

Analyse Reranker

Analysis

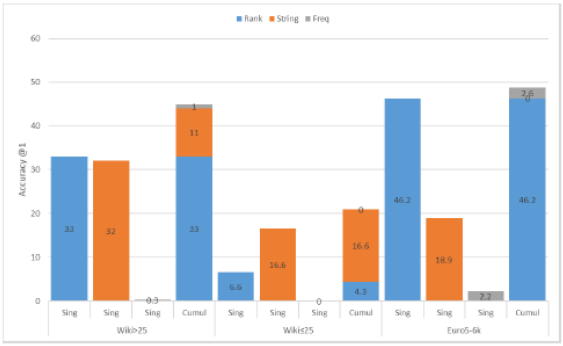
Training Size

- Test = 300 test words (fixed).
- More training examples → better results.
- But 200 training examples seem good enough.



Feature Selection

- Overall: rank >> string >> frequency
- Using all features is preferable.



Reranker Analysis

- Always keeps solutions proposed at @1 by (at least) two approaches.
- Rare that the reranker picks the reference solution at @1 if no approach did it.
- Still the average rank of the reference solution decreases.

	$\text{Wiki}_{>25}$	$\text{Wiki}_{\leq 25}$	Euro_{5-6k}
Rapp	12.7	19.6	16.2
Miko	16.3	30.0	7.5
Faru	20.4	35.5	11.3
list-oracle	12.3	9.1	7.1
reranker	5.6	4.0	4.9

Average rank of the reference translation

Error Analysis

- Manual inspection of the first candidate by our best reranker for the first 100 test forms for which the candidate translation differs from the reference one.
- % of JUNK errors is much higher on $\text{Wiki}_{\leq 25}$ → another illustration of the bias in favor of frequent terms.

	$\text{Wiki}_{>25}$	$\text{Wiki}_{\leq 25}$	Euro_{5-6k}
MORPHO	18	3	26
RELATED	16	4	23
synonyms	*15*	*1*	*19*
antonyms	*1*	*2*	*2*
hyponym			*1*
cohyponym		*1*	*1*
POLYSEMY	4	0	5
LOOSLY	14	15	20
ENGLISH	21	6	7
JUNK	27	72	19

References

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016).

Massively multilingual word embeddings.

arXiv preprint arXiv:1602.01925.

Avraham, O. and Goldberg, Y. (2017).

The interplay of semantics and morphology in word embeddings.

Bamler, R. and Mandt, S. (2017).

Dynamic word embeddings.

References ii

Bouamor, D., Popescu, A., Semmar, N., and Zweigenbaum, P. (2013).

Building specialized bilingual lexicons using large scale background knowledge.

In *EMNLP*, pages 479–489.

Bourdaillet, J., Huet, S., Langlais, P., and Lapalme, G. (2010).

TransSearch: from a bilingual concordancer to a translation finder.

Machine Translation, 24(3-4):241–271.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990).

A statistical approach to machine translation.

Computational linguistics, 16(2):79–85.

References iii

Calixto, I., Liu, Q., and Campbell, N. (2017).

Multilingual multi-modal embeddings for natural language processing.

Coulmance, J., Marty, J.-M., Wenzek, G., and Benhalloum, A. (2015).

Trans-gram, Fast Cross-lingual Word-embeddings.

In *Proceedings of EMNLP*, pages 1109–1113.

Dinu, G. and Baroni, M. (2014).

Improving zero-shot learning by mitigating the hubness problem.

CoRR.

References iv

Faruqui, M. and Dyer, C. (2014).

Improving Vector Space Word Representations Using Multilingual Correlation.

In *Proceedings of EACL*.

Fung, P. (1995).

Compiling bilingual lexicon entries from a non-parallel english-chinese corpus.

In *Third Workshop on Very Large Corpora*, pages 173–183.

Gabrilovich, E. and Markovitch, S. (2007).

Computing semantic relatedness using wikipedia-based explicit semantic analysis.

IJCAI'07, pages 1606–1611.

References v

Gabrilovich, E. and Markovitch, S. (2009).

Wikipedia-based semantic interpretation for natural language processing.

Journal of Artificial Intelligence Research, 34(2):443.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004).

A geometric view on bilingual lexicon extraction from comparable corpora.

In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 526. Association for Computational Linguistics.

References vi

Gouws, S., Bengio, Y., and Corrado, G. (2015).

BilBOWA: Fast Bilingual Distributed Representations without Word Alignments.

In *Proceedings of the 32nd ICML*, pages 748–756.

Gómez-Pérez, A., Vila-Suero, D., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G. (2013).

Guidelines for multilingual linked data.

WIMS '13, pages 3:1–3:12.

Irvine, A. and Callison-Burch, C. (2013).

Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals.

In *Proceedings of NAACL-HLT*, pages 518–523.

Jakubina, L. and Langlais, P. (2015).

Projective methods for mining missing translations in dbpedia.

In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 23–31, Beijing, China. Association for Computational Linguistics.

Jakubina, L. and Langlais, P. (2016a).

Bad luc@wmt 2016: a bilingual document alignment platform based on lucene.

In *Proceedings of the First Conference on Machine Translation*, pages 703–709, Berlin, Germany. Association for Computational Linguistics.

Jakubina, L. and Langlais, P. (2016b).

A comparison of methods for identifying the translation of words in a comparable corpus: Recipes and limits.

CICLing: International Conference on Computational Linguistics and Intelligent Text Processing.

Jakubina, L. and Langlais, P. (2017).

Reranking translation candidates produced by several bilingual word similarity sources.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 605–611, Valencia, Spain. Association for Computational Linguistics.

Levy, O., Goldberg, Y., and Dagan, I. (2015).

Improving distributional similarity with lessons learned from word embeddings.

Transactions of ACL, pages 211–225.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).

Efficient estimation of word representations in vector space.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b).

Exploiting similarities among languages for machine translation.

CoRR.

References x

Morin, E. and Prochasson, E. (2011).

Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora.

In *Proceedings of the 4th workshop on building and using comparable corpora: comparable corpora and the web*, pages 27–34.

Pilehvar, M. T. and Collier, N. (2017).

Inducing embeddings for rare and unseen words by leveraging lexical resources.

Prochasson, E. and Fung, P. (2011).

Rare word translation extraction from aligned comparable documents.

HLT '11, pages 1327–1335.

Rapp, R. (1995).

Identifying Word Translations in Non-parallel Texts.

In *Proceedings of the 33rd ACL*, pages 320–322.

Rapp, R. (1999).

Automatic identification of word translations from unrelated English and German corpora.

In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Rothe, S. and Schütze, H. (2015).

Autoextend: Extending word embeddings to embeddings for synsets and lexemes.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013).

Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora.

In *Building and Using Comparable Corpora*, pages 1–17. Springer Berlin Heidelberg.

Vulić, I., De Smet, W., and Moens, M.-F. (2011).

Identifying word translations from comparable corpora using latent topic models.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 479–484. Association for Computational Linguistics.

Vulic, I., Kiela, D., Clark, S., and Moens, M.-F. (2016).

Multi-modal representations for improved bilingual lexicon learning.

In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 188.