GREYC

Université de Caen
Basse-Normandie

Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle

Adrien Lardilleux

GREYC, université de Caen Basse-Normandie

14 septembre 2010

Cadre : traduction automatique empirique

Paradigme de traduction automatique dans lequel
toute la connaissance nécessaire au processus de traduction
est tirée de textes parallèles.

Exemple de texte parallèle extrait du BTEC (Takezawa et coll., 2002)

この辺りにパブはありませんか。	↔	Is there a pub around here ?
ビール一杯飲みたくて死にそう。	↔	I 'm dying for a swig of beer .
このピザ、美味しい。	↔	This pizza is delicious .
⋮		⋮

Cadre : traduction automatique empirique

Paradigme de traduction automatique dans lequel
toute la connaissance nécessaire au processus de traduction
est tirée de textes parallèles.

Exemple de texte parallèle extrait du BTEC (Takezawa et coll., 2002)

この辺りにパブはありませんか。	↔	Is there a pub around here ?
ビール一杯飲みたくて死にそう。	↔	I 'm dying for a swig of beer .
このピザ、美味しい。	↔	This pizza is delicious .
⋮		⋮

このビール、美味しい。 →

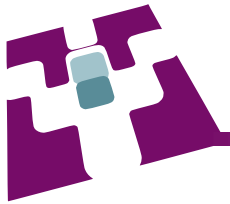
Cadre : traduction automatique empirique

Paradigme de traduction automatique dans lequel
toute la connaissance nécessaire au processus de traduction
est tirée de textes parallèles.

Exemple de texte parallèle extrait du BTEC (Takezawa et coll., 2002)

この辺りにパブはありませんか。	↔	Is there a pub around here ?
ビール一杯飲みたくて死にそう。	↔	I 'm dying for a swig of beer .
このピザ、美味しい。	↔	This pizza is delicious .
⋮		⋮

このビール、美味しい。 →



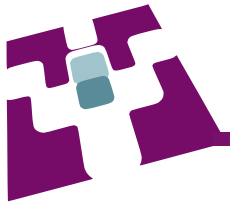
Cadre : traduction automatique empirique

Paradigme de traduction automatique dans lequel
toute la connaissance nécessaire au processus de traduction
est tirée de textes parallèles.

Exemple de texte parallèle extrait du BTEC (Takezawa et coll., 2002)

この辺りにパブはありませんか。	↔	Is there a pub around here ?
ビール一杯飲みたくて死にそう。	↔	I 'm dying for a swig of beer .
このピザ、美味しい。	↔	This pizza is delicious .
⋮		⋮

このビール、美味しい。 →



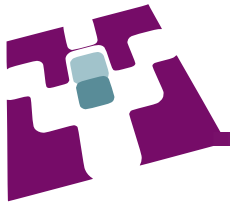
Cadre : traduction automatique empirique

Paradigme de traduction automatique dans lequel
toute la connaissance nécessaire au processus de traduction
est tirée de textes parallèles.

Exemple de texte parallèle extrait du BTEC (Takezawa et coll., 2002)

この辺りにパブはありませんか。	↔	Is there a pub around here ?
ビール一杯飲みたくて死にそう。	↔	I 'm dying for a swig of beer .
このピザ、美味しい。	↔	This pizza is delicious .
⋮		⋮

このビール、美味しい。 → This beer is delicious .



Cadre : traduction automatique empirique

Paradigme de traduction automatique dans lequel
toute la connaissance nécessaire au processus de traduction
est tirée de textes parallèles.

Exemple de texte parallèle extrait du BTEC (Takezawa et coll., 2002)

この辺りにパブはありませんか。	↔	Is there a pub around here ?
ビール一杯飲みたくて死にそう。	↔	I 'm dying for a swig of beer .
このピザ、美味しい。	↔	This pizza is delicious .
⋮		⋮

Objectif : production de tables de traductions

Japonais	Anglais	Scores
ビール	beer	1,0
ピザ	pizza	0,5
ピッツァ	pizza	0,5
ビール 一 杯 飲み	a swig of beer	0,7
この _ 、 美味しい 。	This _ is delicious	0,4
⋮	⋮	⋮

Processus endogène et non supervisé

Objectif : production de tables de traductions

Japonais	Anglais	Scores
ビール	beer	1,0
ピザ	pizza	0,5
ピッツァ	pizza	0,5
ビール 一 杯 飲み	a swig of beer	0,7
この _ 、 美味しい 。	This _ is delicious	0,4
⋮	⋮	⋮

Processus endogène et non supervisé

Objectif : production de tables de traductions

Japonais	Anglais	Scores
ビール	beer	1,0
ピザ	pizza	0,5
ピッツァ	pizza	0,5
ビール 一 杯 飲み	a swig of beer	0,7
この _ 、 美味しい 。	This _ is delicious	0,4
⋮	⋮	⋮

Processus endogène et non supervisé

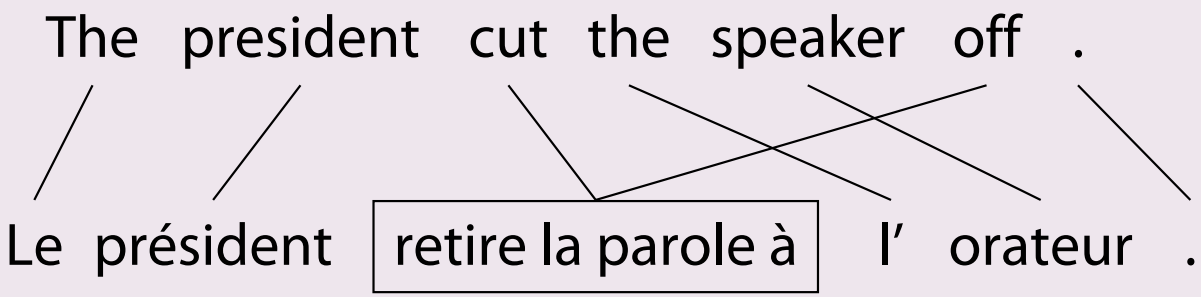
Objectif : production de tables de traductions

Japonais	Anglais	Scores
ビール	beer	1,0
ピザ	pizza	0,5
ピッツァ	pizza	0,5
ビール 一 杯 飲み	a swig of beer	0,7
この _ 、 美味しい 。	This _ is delicious	0,4
⋮	⋮	⋮

Processus endogène et non supervisé

Approche 1/2 : estimative (Brown et coll., 1988)

Un modèle, des paramètres, des liens



(× plusieurs milliers)

Problème de maximisation globale

⇓

Une table de traductions

$e \leftrightarrow f$	$P(f\|e)$	$P(e\|f)$
Le ↔ the	0,2	0,3
La ↔ the	0,1	0,2

(× plusieurs milliers)

Approche 2/2 : associative (Gale et Church, 1991)

Des candidats, un test d'indépendance

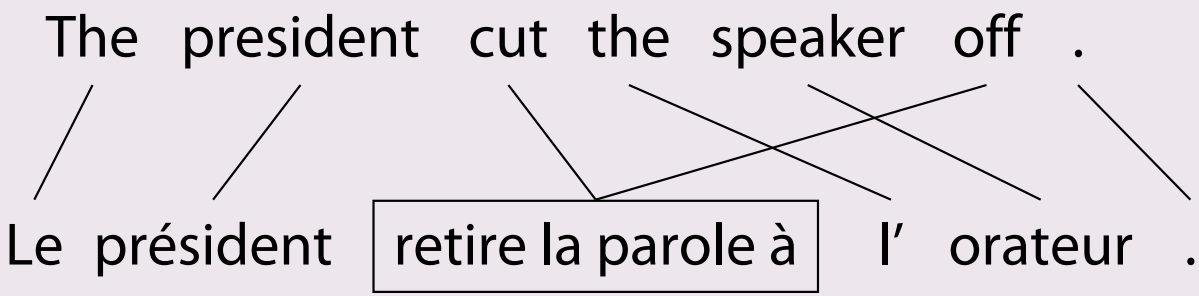
$e \leftrightarrow f$	$\varphi(e,f)$
Le ↔ the	0,6
La ↔ the	0,5

(× plusieurs milliers)

$\Downarrow$

Problème de maximisation locale

Éventuellement : des liens



(× plusieurs milliers)

L'outil de référence : GIZA++ (Och et Ney, 2003)

- Implémente les modèles IBM (Brown et coll., 1993) et le HMM (Vogel et coll., 1996)
- Produit des résultats de très bonne qualité
- Parfaitement couplé au système de TA statistique Moses (Koehn et coll., 2007)

L'outil de référence : GIZA++ (Och et Ney, 2003)

- Implémente les modèles IBM (Brown et coll., 1993) et le HMM (Vogel et coll., 1996)
- Produit des résultats de très bonne qualité
- Parfaitement couplé au système de TA statistique Moses (Koehn et coll., 2007)

Critiques possibles :

- 6 modèles différents, de complexité croissante
  $\Rightarrow$ temps d'exécution important, passage à l'échelle difficile
- Modèles compliqués $\Rightarrow$ outil compliqué
  ($\approx$ 30 000 lignes de code, 58 options, $\approx$ 25 fichiers temporaires)
- Modèles bilingues et asymétriques

Nos objectifs

Réponses possibles aux critiques précédentes

- Passage à l'échelle au cœur de la méthode
- Outil simple ⇐ un seul modèle simple
- Modèle multilingue non directionnel

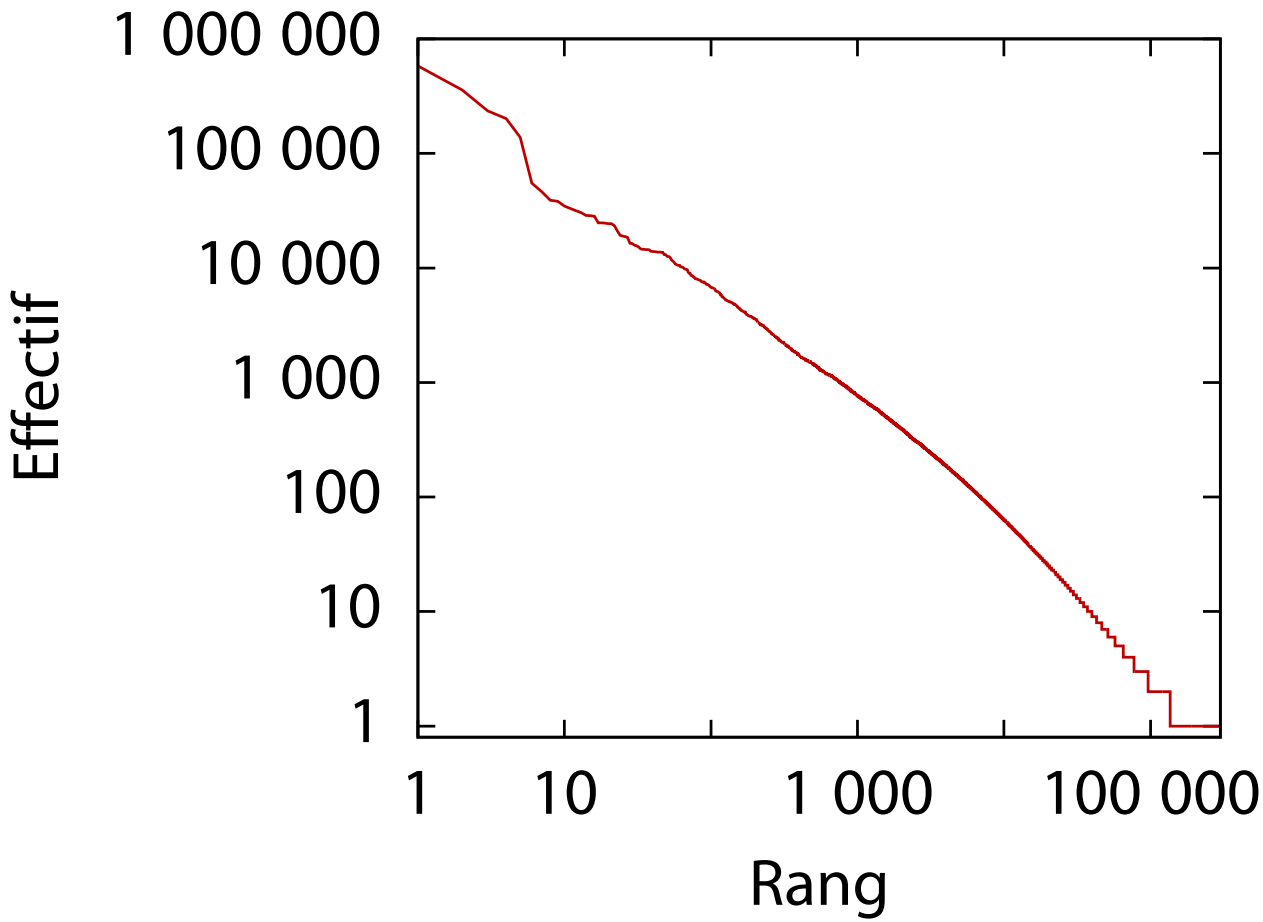
1. *Contribution à l'étude des mots rares*

2. *Contribution à l'alignement de séquences de mots*
   - Ébauche de la méthode : comment, quoi ?
   - Extraction d'alignements à partir de corpus multilingues
   - Calcul des scores
   - Optimisation de la méthode

3. *Évaluation et perspectives*

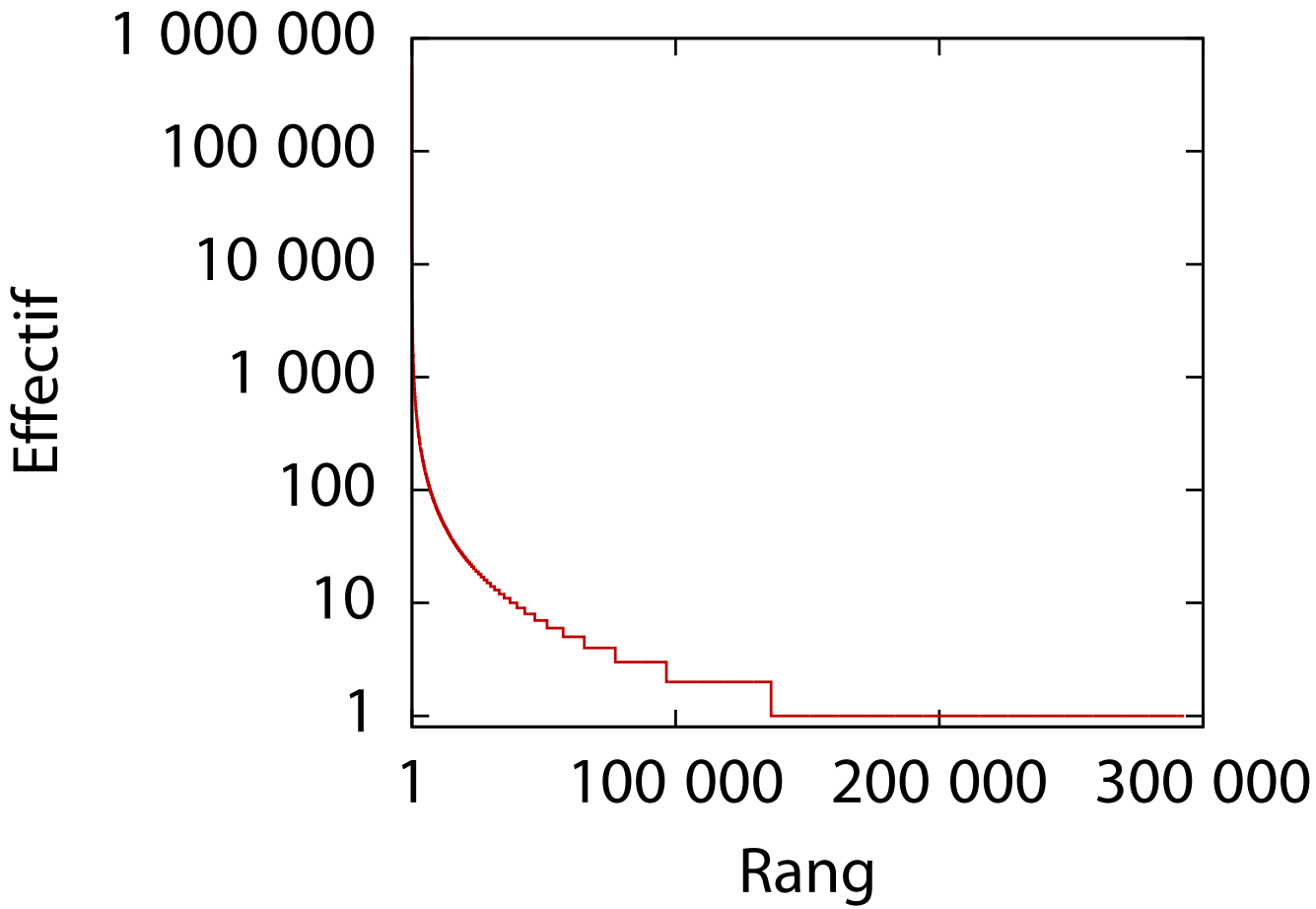
Loi d'Estoup-Zipf

Échelle logarithmique :

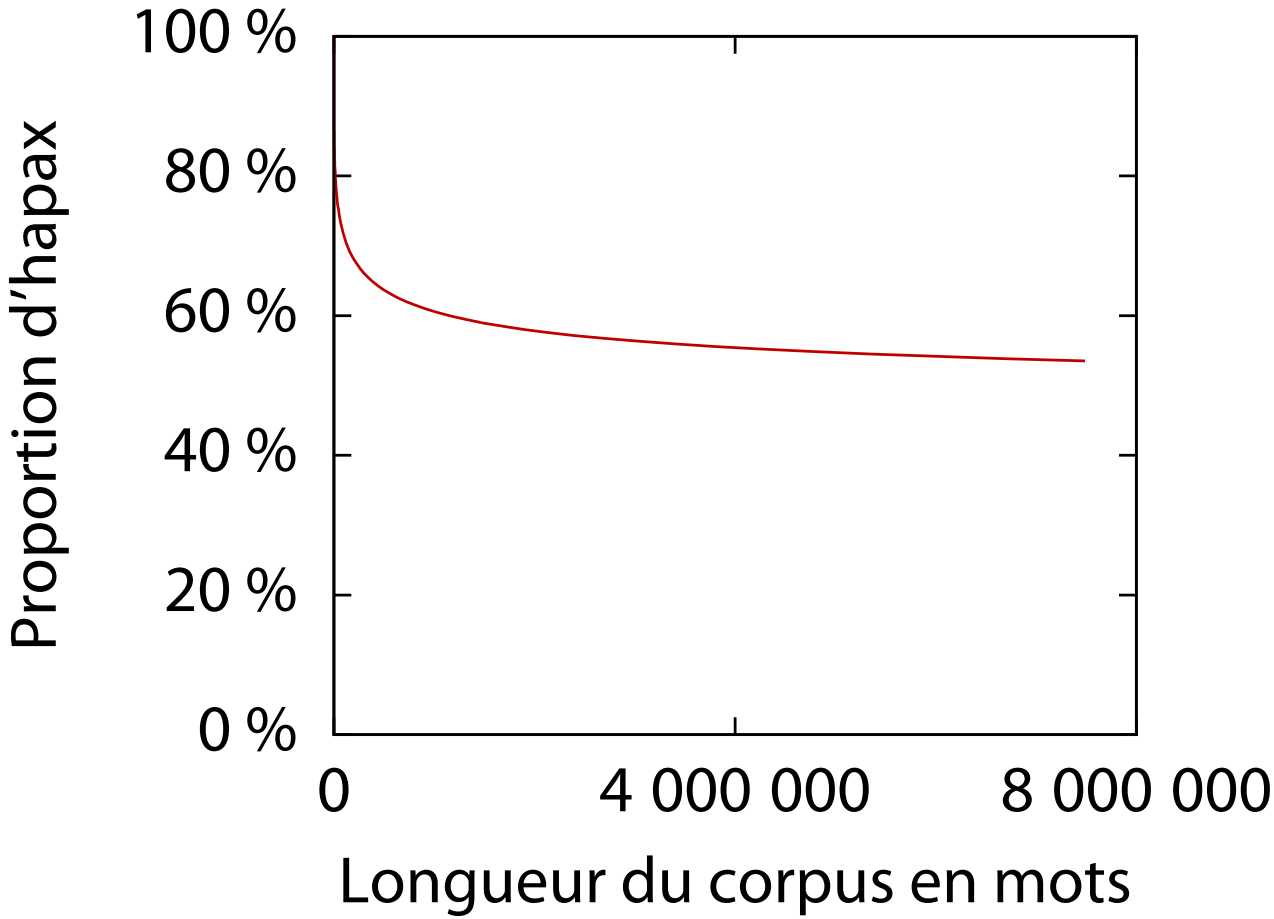


Loi d'Estoup-Zipf

Échelle semi-logarithmique :



Incompressibilité du nombre d'hapax



Expérience d'alignement bilingue entre mots

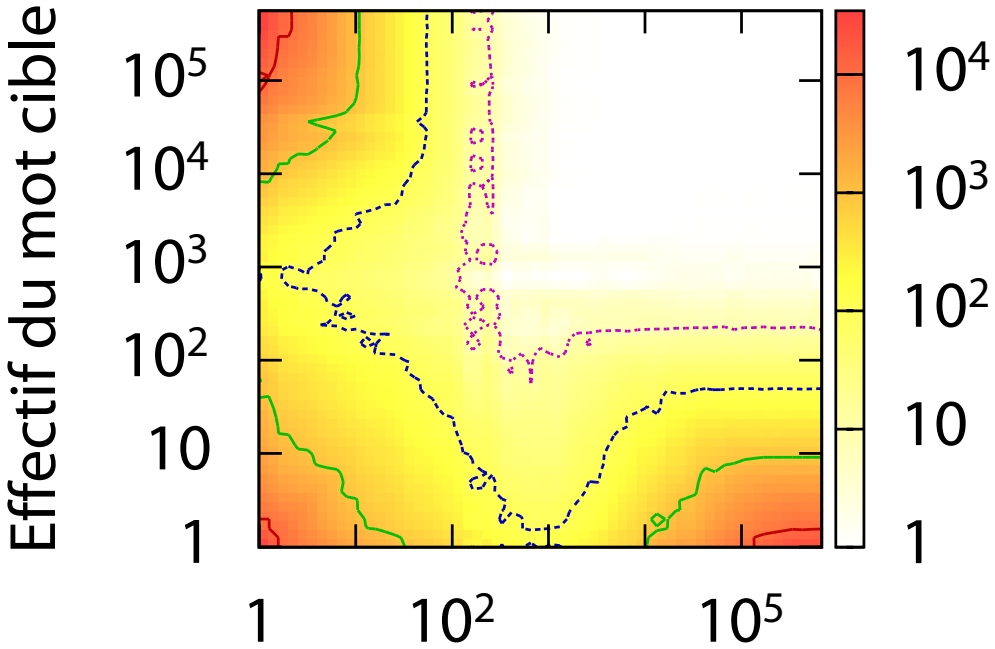
- Une méthode associative : méthode du cosinus
- On associe à chaque mot d'un corpus parallèle un vecteur dont les éléments sont le nombre de fois que le mot apparaît dans chacune des phrases
- On calcule pour chaque couple de mots (*source*, *cible*) l'angle formé par leurs vecteurs respectifs

$$\Downarrow$$

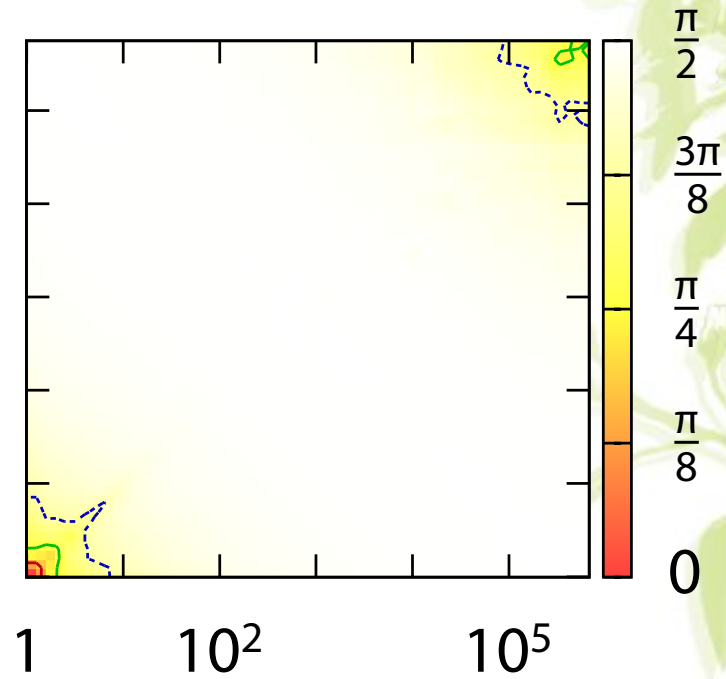
$m \times n$ couples de mots
et leurs angles $\in [0 ; \pi/2]$

Nombre d'alignements et angles (≈ qualité)

Quantité d'alignements



Angles des alignements



Alignements d'hapax

- 80 à 90 % des alignements d'angle nul = couples d'hapax
- Seulement 12 % des phrases contiennent au moins un hapax
- Parmi ces 12 %, plus de 80 % n'en contiennent qu'un seul

L'alignement d'hapax est simple
et produit des résultats de bonne qualité :

桃	↔	peach
風景画	↔	landscape
ミサ	↔	mass
宝くじ	↔	lottery
アルバム	↔	album

スイート	↔	suite
ユーフォー	↔	UFO
ベーグル	↔	bagels
わっ	↔	crocodile
エイズ	↔	AIDS

1 *Contribution à l'étude des mots rares*

2 *Contribution à l'alignement de séquences de mots*

- Ébauche de la méthode : comment, quoi ?
- Extraction d'alignements à partir de corpus multilingues
- Calcul des scores
- Optimisation de la méthode

3 *Évaluation et perspectives*

Rappel des objectifs

Résultats sur les mots rares

- Multiplicité 1-1 : mot-à-mot
- Bilingue
- On ne sait bien aligner que les hapax

Nos objectifs

- Multiplicité m-n : séquences de mots
- « Alingue » (= multilingue)
- Nous saurons aligner tous les mots

Notre proposition

Renversement des termes de la doxa du domaine :

- Pour aligner des mots rares
  par une méthode fondée sur les mots fréquents,
  ajouter des données $\rightarrow \infty$

Notre proposition

Renversement des termes de la doxa du domaine :

- Pour aligner des mots rares
  par une méthode fondée sur les mots fréquents,
  ajouter des données $\rightarrow \infty$

- Pour aligner des mots fréquents
  par une méthode fondée sur les mots rares,
  supprimer des données $\rightarrow 0$

Notre proposition

Renversement des termes de la doxa du domaine :

- Pour aligner des mots rares
  par une méthode fondée sur les mots fréquents,
  ajouter des données $\rightarrow \infty$

- Pour aligner des mots fréquents
  par une méthode fondée sur les mots rares,
  supprimer des données $\rightarrow 0$

Cœur de la méthode

Traiter de multiples sous-corpus

Avantages de notre proposition

Modèle :

- significativité : « *Less data is more data !* »
- simplicité : principe du tout ou rien

Linguistique :

- désambiguïsation gratuite : un hapax = un sens
- traitement réellement multilingue

Informatique :

- moins de mémoire nécessaire
- parallélisation facile, rapidité

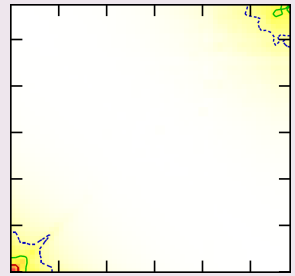
Problème des hautes fréquences

- Les mots très fréquents sont toujours très fréquents, quelle que soit la taille des sous-corpus
  ⇒ Il est impossible de les aligner en les rendant rares (Exemple : le point)
- Peut-on concilier hautes et basses fréquences ?

Problème des hautes fréquences

- Les mots très fréquents sont toujours très fréquents, quelle que soit la taille des sous-corpus
  ⇒ Il est impossible de les aligner en les rendant rares (Exemple : le point)
- Peut-on concilier hautes et basses fréquences ?

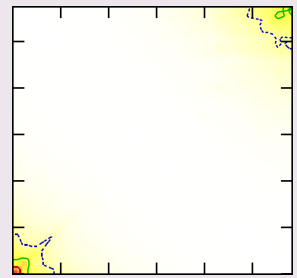
Hautes et basses fréquences donnent
de bons alignements par des méthodes reposant
sur la similarité de leurs distributions



Problème des hautes fréquences

- Les mots très fréquents sont toujours très fréquents, quelle que soit la taille des sous-corpus
  ⇒ Il est impossible de les aligner en les rendant rares (Exemple : le point)
- Peut-on concilier hautes et basses fréquences ?

Hautes et basses fréquences donnent
de bons alignements par des méthodes reposant
sur la similarité de leurs distributions



⇓

Nous alignerons des mots partageant la même distribution,
quelles que soient leurs fréquences

Ébauche d'algorithme

Répéter :

Tirer un sous-corpus

En extraire les séquences de mots de même distribution

Fin répéter

Calculer les scores des alignements

Ébauche d'algorithme

Répéter :

Tirer un sous-corpus

En extraire les séquences de mots de même distribution

Fin répéter

Calculer les scores des alignements

$\Rightarrow$ « *Anytime* »

Le nombre de sous-corpus traités n'influe pas sur la qualité mais sur la quantité et la significativité des résultats

Le multilingue par l'« alingue »

Corpus parallèle multilingue

1	قهوة ، من فضلك .	↔ Un café , s'il vous plaît .	↔ One coffee , please .
2	هذه قهوة ممتازة .	↔ Ce café est excellent .	↔ This coffee is excellent .
3	شاي ثقيل .	↔ Un thé fort .	↔ One strong tea .
4	قهوة ثقيلة .	↔ Un café fort .	↔ One strong coffee .

Le multilingue par l'« alingue »

Corpus parallèle multilingue

1 قهوة ، من فضلك . ↔ Un café , s'il vous plaît . ↔ One coffee , please .

2 هذه قهوة ممتازة . ↔ Ce café est excellent . ↔ This coffee is excellent .

3 شاي ثقيل . ↔ Un thé fort . ↔ One strong tea .

4 قهوة ثقيلة . ↔ Un café fort . ↔ One strong coffee .

⇓

Transformation en corpus « alingue »

1 قهوة$_1$ ،$_1$ من$_1$ فضلك$_1$.$_1$ Un$_2$ café$_2$,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ One$_3$ coffee$_3$,$_3$ please$_3$.$_3$

2 هذه$_1$ قهوة$_1$ ممتازة$_1$.$_1$ Ce$_2$ café$_2$ est$_2$ excellent$_2$.$_2$ This$_3$ coffee$_3$ is$_3$ excellent$_3$.$_3$

3 شاي$_1$ ثقيل$_1$.$_1$ Un$_2$ thé$_2$ fort$_2$.$_2$ One$_3$ strong$_3$ tea$_3$.$_3$

4 قهوة$_1$ ثقيلة$_1$.$_1$ Un$_2$ café$_2$ fort$_2$.$_2$ One$_3$ strong$_3$ coffee$_3$.$_3$

Extraction des alignements (1/2)

Sélection d'un sous-corpus

1 1، 1 فضلك 1 قهوة 1 من 1. ،2 ،3 .2 .3 One$_3$ Un$_2$ café$_2$ coffee$_3$ plaît$_2$ please$_3$ s'il$_2$ vous$_2$

2 1 قهوة 1 ممتازة 1 هذه 1. .2 .3 Ce$_2$ This$_3$ café$_2$ coffee$_3$ est$_2$ excellent$_2$ excellent$_3$ is$_3$

3 1 ثقيل 1 شاي 1. .2 .3 One$_3$ Un$_2$ fort$_2$ strong$_3$ tea$_3$ thé$_2$

Extraction des alignements (1/2)

Sélection d'un sous-corpus

1 ١، ١ فضلك ١ قهوة ١ من ١. ،٢ ،٣ .٢ .٣ One$_3$ Un$_2$ café$_2$ coffee$_3$ plaît$_2$ please$_3$ s'il$_2$ vous$_2$

2 ١ قهوة ١ ممتازة ١ هذه ١. .٢ .٣ Ce$_2$ This$_3$ café$_2$ coffee$_3$ est$_2$ excellent$_2$ excellent$_3$ is$_3$

3 ١ ثقيل ١ شاي ١. .٢ .٣ One$_3$ Un$_2$ fort$_2$ strong$_3$ tea$_3$ thé$_2$

⇓

Calcul des vecteurs d'apparition des mots

	١، ١ ثقيل ١ شاي ١ فضلك ١ قهوة ١ ممتازة ١ من ١ هذه ١ ،٢ ،٣ ١. .٢ .٣	Ce$_2$	One$_3$	This$_3$	Un$_2$	café$_2$	coffee$_3$	est$_2$	excellent$_2$	excellent$_3$	fort$_2$	is$_3$	plaît$_2$	. . .
1	1 0 0 1 1 0 1 0 1 1 1 1 1	0	1	0	1	1	1	0	0	0	0	0	1	. . .
2	0 0 0 0 1 1 0 1 0 0 1 1 1	1	0	1	0	1	1	1	1	1	0	1	0	. . .
3	0 1 1 0 0 0 0 0 0 0 1 1 1	0	1	0	1	0	0	0	0	0	1	0	0	. . .

Extraction des alignements (1/2)

Sélection d'un sous-corpus

1 1، 1 فضلك 1 قهوة 1 من 1 .1 ،2 ،3 .2 .3 One$_3$ Un$_2$ café$_2$ coffee$_3$ plaît$_2$ please$_3$ s'il$_2$ vous$_2$

2 1 قهوة 1 ممتازة 1 هذه 1 .1 .2 .3 Ce$_2$ This$_3$ café$_2$ coffee$_3$ est$_2$ excellent$_2$ excellent$_3$ is$_3$

3 1 ثقيل 1 شاي 1 .1 .2 .3 One$_3$ Un$_2$ fort$_2$ strong$_3$ tea$_3$ thé$_2$

⇓

Calcul des vecteurs d'apparition des mots

	1،	1 ثقيل	1 شاي	1 فضلك	1 قهوة	1 ممتازة	1 من	1 هذه	،2	،3	.1	.2	.3	Ce$_2$	One$_3$	This$_3$	Un$_2$	café$_2$	coffee$_3$	est$_2$	excellent$_2$	excellent$_3$	fort$_2$	is$_3$	plaît$_2$	. . .
1	1	0	0	1	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	0	0	0	1	. . .
2	0	0	0	0	1	1	0	1	0	0	1	1	1	1	0	1	0	1	1	1	1	1	0	1	0	. . .
3	0	1	1	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	1	0	0	. . .

Extraction des alignements (1/2)

Sélection d'un sous-corpus

1 ١، ١ فضلك ١ قهوة ١ من ١ ٠. ،٢ ،٣ .٢ .٣ One$_3$ Un$_2$ café$_2$ coffee$_3$ plaît$_2$ please$_3$ s'il$_2$ vous$_2$

2 ١ قهوة ١ ممتازة ١ هذه ١ ٠. .٢ .٣ Ce$_2$ This$_3$ café$_2$ coffee$_3$ est$_2$ excellent$_2$ excellent$_3$ is$_3$

3 ١ ثقيل ١ شاي ١ ٠. .٢ .٣ One$_3$ Un$_2$ fort$_2$ strong$_3$ tea$_3$ thé$_2$

⇓

Calcul des vecteurs d'apparition des mots

	١، ١ ثقيل ١ شاي ١ فضلك ١ قهوة ١ ممتازة ١ من ١ هذه ١ ،٢ ،٣ ٠. .٢ .٣	Ce$_2$	One$_3$	This$_3$	Un$_2$	café$_2$	coffee$_3$	est$_2$	excellent$_2$	excellent$_3$	fort$_2$	is$_3$	plaît$_2$	. . .
1	1 0 0 1 1 0 1 0 1 1 1 1 1	0	1	0	1	1	1	1	0	0	0	0	1	. . .
2	0 0 0 0 1 1 0 1 0 0 1 1 1	1	0	1	0	1	1	1	1	1	0	1	0	. . .
3	0 1 1 0 0 0 0 0 0 0 1 1 1	0	1	0	1	0	0	0	0	1	0	0	0	. . .

⇓

Tri des mots par vecteurs

	٠. .٢ .٣	١ قهوة	café$_2$	coffee$_3$	One$_3$	Un$_2$	١، ١ فضلك ١ من ١ ،٢ ،٣	plaît$_2$	please$_3$	s'il$_2$	vous$_2$	١ ممتازة ١ هذه	Ce$_2$	This$_3$	est$_2$	excellent$_2$	excellent$_3$	. . .
1	1 1 1	1	1	1	1	1	1 1 1 1 1	1	1	1	1	0 0 0	0	0	0	0	0	. . .
2	1 1 1	1	1	1	0	0	0 0 0 0 0	0	0	0	0	1 1 1	1	1	1	1	1	. . .
3	1 1 1	0	0	0	1	1	0 0 0 0 0	0	0	0	0	0 0 0	0	0	0	0	0	. . .

Extraction des alignements (2/2)

Extraction des séquences de même distribution et de leurs contextes

Les mots :	*apparaissent dans les phrases :*	*d'où nous extrayons :*
قهوة_1 café_2 coffee_3	1	قهوة_1 café_2 coffee_3 فضلك_1 من_1 ، Un_2 _ ,$_2$ s'il_2 vous_2 plaît_2 .$_2$ One_3 _ ,$_3$ please_3 .$_3$
	2	قهوة_1 café_2 coffee_3 ممتازة_1 _ هذه_1 Ce_2 _ est_2 excellent_2 .$_2$ This_3 _ is_3 excellent_3 .$_3$
		⋮

Extraction des alignements (2/2)

Extraction des séquences de même distribution et de leurs contextes

Les mots :	*apparaissent dans les phrases :*	*d'où nous extrayons :*
قهوة$_1$ café$_2$ coffee$_3$	1	قهوة$_1$ café$_2$ coffee$_3$ Un$_2$ _ ,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ One$_3$ _ ,$_3$ please$_3$.$_3$ ، من$_1$ فضلك$_1$.$_1$
	2	قهوة$_1$ café$_2$ coffee$_3$ Ce$_2$ _ est$_2$ excellent$_2$.$_2$ This$_3$ _ is$_3$ excellent$_3$.$_3$ هذه$_1$ _ ممتازة$_1$.$_1$

⋮

⇓

Collecte des alignements et décompte

Arabe	*Français*	*Anglais*	*Décompte*
قهوة	café	coffee	2
، من فضلك .	Un _ , s'il vous plaît .	One _ , please .	1
هذه _ ممتازة .	Ce _ est excellent .	This _ is excellent .	1

Transformation en table de traductions

Calcul des scores ⇒ table de traductions complète

Arabe	*Français*	*Anglais*	P(*fr,en*|*ar*)	P(*ar,en*|*fr*)	P(*ar,fr*|*en*)	. . .
قهوة	café	coffee	1,0	0,5	1,0	. . .
من فضلك ، .	Un _ , s'il vous plaît .	One _ , please .	0,9	0,5	1,0	. . .
هذه _ ممتازة .	Ce _ est excellent .	This _ is excellent .	0,4	0,3	0,7	. . .
⋮	⋮	⋮		⋮		

Attribution de scores aux alignements

Probabilités de traduction : à partir des décomptes des alignements

$$P(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_L | s_i) = \frac{C(s_1, \ldots, s_L)}{C(s_i)}$$

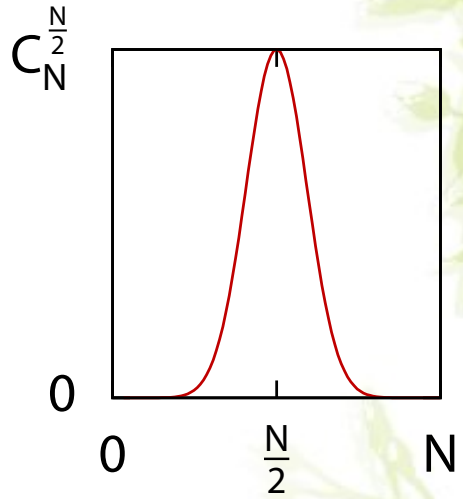
Poids lexicaux : éventuellement, à partir des occurrences de mots dans le corpus de départ
(calcul non standard adapté de Koehn, 2003)

$$L(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_L | s_i) = \prod_{m_i \in s_i} \max_{m_j \in \cup_{i \neq j} s_j} D(m_j | m_i)$$

où $D(m_j|m_i) = \frac{C(m_i, m_j)}{C(m_i)}$ avec $1 \leqslant i \leqslant L$

Stratégie de sélection des sous-corpus

- $2^N - 1$ sous-corpus possibles
- Toutes les tailles ne présentent pas le même intérêt



Nous constituerons des sous-corpus en échantillonnant selon une distribution a priori

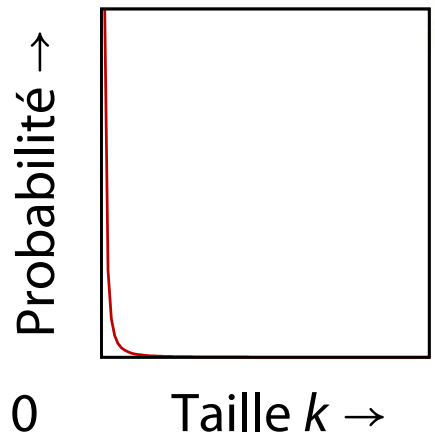
Fonction d'échantillonnage

Objectif imposé à la distribution

Couvrir au mieux le vocabulaire du corpus de départ

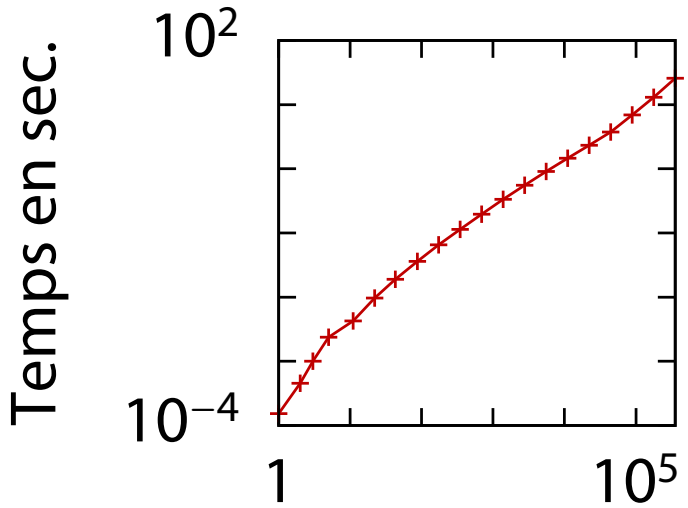
$\Rightarrow$ Couvrir au mieux les phrases

$$p(k) = \frac{-1}{k \log (1 - k/N)}$$

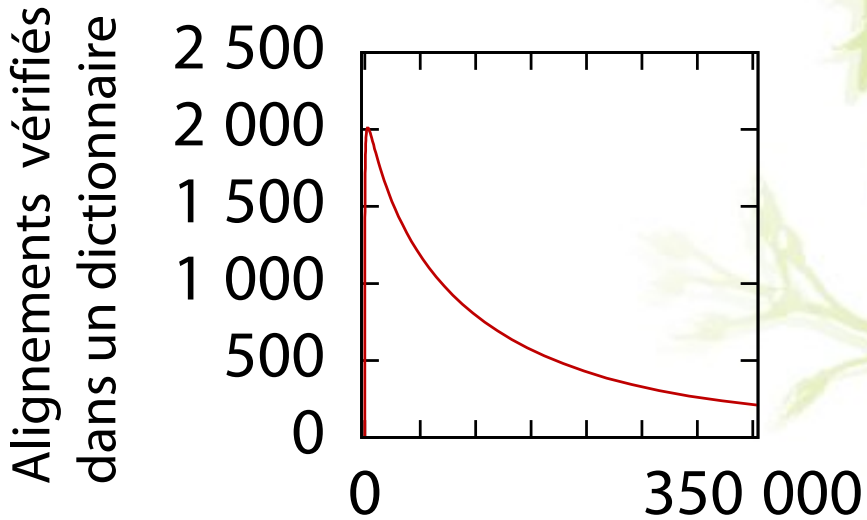


Influence de la taille des sous-corpus (1/2)

Temps de traitement

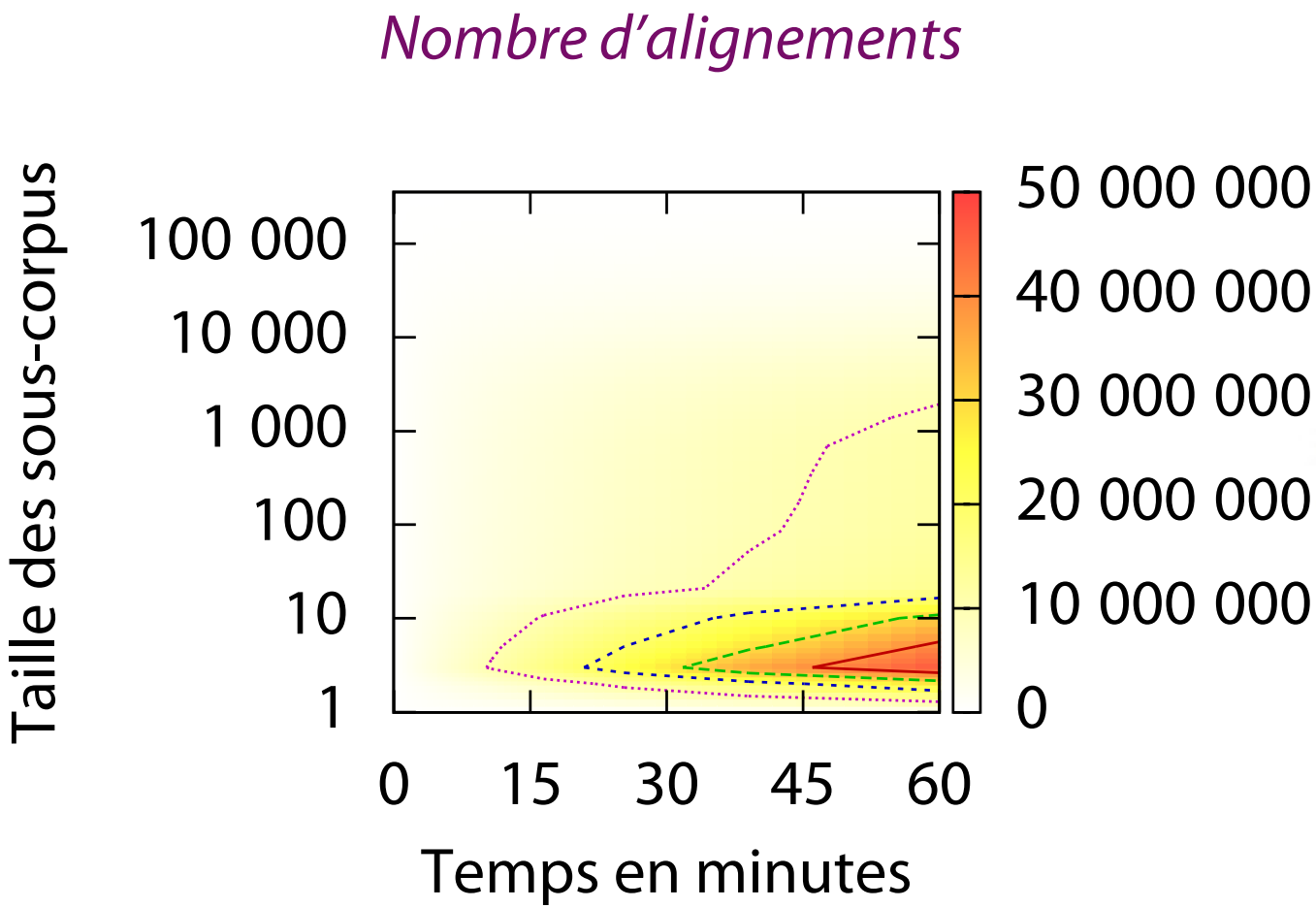


Qualité des alignements



Taille des sous-corpus

Influence de la taille des sous-corpus (2/2)



Algorithme complet

Répéter :

Tirer un sous-corpus

En extraire les séquences de mots de même distribution

Fin répéter

Calculer les scores des alignements

Algorithme complet

Transformer le corpus parallèle multilingue en corpus alingue

Répéter :

Tirer un sous-corpus

En extraire les séquences de mots de même distribution

Fin répéter

Calculer les scores des alignements

Algorithme complet

Transformer le corpus parallèle multilingue en corpus alingue

Alignements = {}

Répéter :

Tirer un sous-corpus

En extraire les séquences de mots de même distribution

Fin répéter

Calculer les scores des alignements

Algorithme complet

Transformer le corpus parallèle multilingue en corpus alingue

Alignements = {}

Répéter :

Tirer un sous-corpus de taille k avec $p(k) = \frac{-1}{k \log(1-k/N)}$

En extraire les séquences de mots de même distribution

Fin répéter

Calculer les scores des alignements

Algorithme complet

Transformer le corpus parallèle multilingue en corpus alingue

Alignements = {}

Répéter :

Tirer un sous-corpus de taille k avec $p(k) = \frac{-1}{k \log(1-k/N)}$

Calculer les vecteurs d'apparition des mots

Trier les mots par vecteurs pour former des *groupes*

Pour chaque *groupe* de mots :

Pour chaque *phrase* du sous-corpus où le groupe apparaît :

Alignements[groupe] ++

Alignements[phrase - groupe] ++

Fin répéter

Calculer les scores des alignements

Algorithme complet

Transformer le corpus parallèle multilingue en corpus alingue

Alignements = {}

Répéter :

Tirer un sous-corpus de taille k avec $p(k) = \frac{-1}{k \log(1-k/N)}$

Calculer les vecteurs d'apparition des mots

Trier les mots par vecteurs pour former des *groupes*

Pour chaque *groupe* de mots :

Pour chaque *phrase* du sous-corpus où le groupe apparaît :

Alignements[groupe] ++

Alignements[phrase - groupe] ++

Jusqu'à {certain temps écoulé, aucun nouvel alignement, . . . }

Calculer les scores des alignements

Algorithme complet

Transformer le corpus parallèle multilingue en corpus alingue

Alignements = {}

Répéter :

Tirer un sous-corpus de taille k avec $p(k) = \frac{-1}{k \log(1-k/N)}$

Calculer les vecteurs d'apparition des mots

Trier les mots par vecteurs pour former des *groupes*

Pour chaque *groupe* de mots :

Pour chaque *phrase* du sous-corpus où le groupe apparaît :

Alignements[groupe] ++

Alignements[phrase - groupe] ++

Jusqu'à {certain temps écoulé, aucun nouvel alignement, . . .}

Calculer les scores des alignements

ceb	dan	grc	eng	fin	fra	ind	lat	spa	swe	vie	zho
pedro	peter	πετρος	peter	pietari	pierre	petrus	petrus	pedro	petrus	phêrô	彼得
pablo	paulus	παυλος	paul	paavali	paul	paulus	paulus	pablo	paulus	phaolô	保羅
pilato	pilatus	πιλατος	pilate	pilatus	pilate	pilatus	pilatus	pilato	pilatus	philatô	彼
zabulon	sebulons	ζαβουλων	zabulon	sebulonin	zabulon	zebulon	zabulon	zabulón	sabulons	zabulôn	西布倫
alfeo	alfæus	αλφαιου	alphaeus	alfeuksen	alphée	alfeus	alphei	alfeo	alfeus	alphê	亞勒腓
moises	moses	μωσης	moses	mooses	moïse	musa	moses	moisés	moses	môsê	摩西
simon	simon	σιμων	simon	simon	simon	simon	simon	simón	simon	simôn	西門
pilato	pilatus	πιλατος	pilate	pilatus	pilate	pilatus	pilatus	pilato	pilatus	philatô	彼 拉多
arquipo	arkippus	αρχιππω	archippus	arkippukselle	archippe	arkhipus	archippo	arquipo	arkippus	arkhippô	亞基布
marta	martha	μαρθα	martha	martta	marthe	marta	martha	marta	marta	martha	馬大
juan	johannes	ιωαννης	john	johannes	jean	yohanes	iohannes	juan	johannes	yoan	約翰
elisabet	elisabeth	ελισαβετ	elisabeth	elisabet	élisabeth	elisabet	elisabeth	elisabet	elisabet	êlisabet	沙伯
herodes	herodes	ηρωδης	herod	herodes	hérode	herodes	herodes	herodes	herodes	hêrôđê	希
igsoon	brødre	αδελφοι	brethren	veljet	frères	saudara-sauda	fratres	hermanos	bröder	hỡi	弟兄們
escriba	skriftkloge	γραμματεις	scribes	kirjanoppineet	scribes	ahli-ahli	scribae	escribas	skriftlärde	ký lục	文 士
fariseo	farisæerne	φαρισαιοι	pharisees	fariseukset	pharisiens	farisi	pharisaei	fariseos	fariséerna	biệt phái	法利 賽
cordero	lammets	αρνιου	lamb	karitsan	agneau	domba	agni	cordero	lammets	chiên	羔羊
capernaum	kapernaum	καπερναουμ	capernaum	kapernaumiin	capernaüm	kapernaum	capharnaum	capernaúm	kapernaum	capharnaum	農
dios	guds	θεου	god	jumalan	dieu	allah	dei	dios	guds	chúa	神
damgo	drøm	οναρ	dream	unessa	songe	mimpi	somnis	sueños	drömmen	mộng	夢中
jesus	jesus	ιησους	jesus	jeesus	jésus	yesus	iesus	jesús	jesus	đức yêsu	耶穌
aleluya	halleluja	αλληλουια	alleluia	halleluja	alléluia	haleluya	alleluia	aleluya	halleluja	hallêluya	哈利
escriba	skriftkloge	γραμματεις	scribes	kirjanoppineet	scribes	ahli-ahli taurat	scribae	escribas	skriftlärde	ký lục	文 士
fariseo	farisæerne	φαρισαιοι	pharisees	fariseukset	pharisiens	farisi	pharisaei	fariseos	fariséerna	biệt	法利 賽
maria	maria	μαριαμ	mary	maria	marie	maria	maria	maría	maria	maria	馬利亞
tarso	tarsus	ταρσον	tarsus	tarsoon	tarse	tarsus	tarsum	tarso	tarsus	tarsô	大數
galilea	galilæa	γαλιλαιας	galilee	galilean	galilée	galilea	galilaeae	galilea	galileen	galilê	加利利
dios	guds	θεου	god	jumalan	dieu	allah	dei	dios	guds	thiên	神
juan	johannes	ιωαννην	john	johanneksen	jean	yohanes	iohannem	juan	johannes	yoan	約翰
cornelio	kornelius	κορνηλιος	cornelius	kornelius	corneille	kornelius	cornelius	cornelio	kornelius	cornêliô	哥尼 流
hari	konge	βασιλευς	king	kuningas	roi	raja	rex	rey	konung	vua	王
saulo	saulus	σαυλος	saul	saulus	saul	saulus	saulus	saulo	saulus	saulô	掃 羅
derbe	derbe	δερβην	derbe	derbeen	derbe	derbe	lystram	derbe	derbe	đerbê	特
dalmanuta	dalmanuthas	δαλμανουθα	dalmanutha	dalmanutan se	dalmanutha	dalmanuta	dalmanutha	dalmanuta	dalmanuta	đalmanutha	努
herodes	herodes	ηρωδης	herod	herodes	hérode	herodes	herodes	herodes	herodes	hêrôđê	希 律
benjamin	benjamins	βενιαμιν	benjamin	benjaminin	benjamin	benyamin	beniamin	benjamín	benjamins	benyamin	憫
judas	judas	ιουδας	judas	juudas	judas	yudas	iudas	judas	judas	yuđa	猶
siria	syrien	συριαν	syria	syyriaan	syrie	siria	syriam	siria	syrien	syri	敘利亞
agripa	agrippa	αγριππα	agrippa	agrippa	agrippa	agripa	agrippa	agripa	agrippa	agrippa	亞基帕
dragon	dragen	δρακων	dragon	lohikäärme	dragon	naga	draco	dragón	draken	rồng	龍
sambingay	lignelse	παραβολην	parable	vertauksen	parabole	perumpamaan	parabolam	parábola	liknelse	ví dụ	比喻
egipto	ægypten	αιγυπτω	egypt	egyptissä	égypte	mesir	aegypto	egipto	egypten	câp	埃及
elias	elias	ηλιας	elias	elias	élie	elia	helias	elías	elias	êlya	以利
santiago	jakob	ιακωβον	james	jaakobin	jacques	yakobus	iacobum	jacobo	jakob	yacôbê	雅
salome	salome	σαλωμη	salome	salome	salomé	salome	salome	salomé	salome	salômê	羅米
magdalena	magdalene	μαγδαληνη	magdalene	magdaleena	magdala	magdalena	magdalene	magdalena	magdala	magđala	大拉
babilonia	babylon	βαβυλων	babylon	babylon	babylone	babel	babylon	babilonia	babylon	babylon	巴比倫大
silas	silas	σιλας	silas	silas	silas	silas	silas	silas	silas	sila	西拉
cornelio	kornelius	κορνηλιος	cornelius	kornelius	corneille	kornelius	cornelius	cornelio	kornelius	cornêliô	哥尼

```
<?xml version="1.0"?>
<tmx version="1.4">
<header creationtool="anymalign" creationtoolversion="2.3 (July 20th 2009)"
 datatype="plaintext" segtype="phrase" adminlang="en-us" srclang="*all*"
 o-tmf="none" />
<body>
<tu>
 <prop type="freq">12</prop>
 <prop type="probas">1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
 </prop>
 <prop type="lexWeights">1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000</prop>
 <tuv xml:lang="bg"><seg>&#1095;&#1083;&#1077;&#1085;</seg></tuv>
 <tuv xml:lang="cs"><seg>&#269;l&#225;nek</seg></tuv>
 <tuv xml:lang="da"><seg>artikel</seg></tuv>
 <tuv xml:lang="de"><seg>artikel</seg></tuv>
 <tuv xml:lang="el"><seg>&#940;&#961;&#952;&#961;&#959;</seg></tuv>
 <tuv xml:lang="en"><seg>article</seg></tuv>
 <tuv xml:lang="es"><seg>art&#237;culo</seg></tuv>
 <tuv xml:lang="et"><seg>artikkel</seg></tuv>
 <tuv xml:lang="fi"><seg>artikla</seg></tuv>
 <tuv xml:lang="fr"><seg>article</seg></tuv>
 <tuv xml:lang="hu"><seg>cikk</seg></tuv>
 <tuv xml:lang="it"><seg>articolo</seg></tuv>
 <tuv xml:lang="lt"><seg>straipsnis</seg></tuv>
 <tuv xml:lang="lv"><seg>pants</seg></tuv>
 <tuv xml:lang="mt"><seg>artikolu</seg></tuv>
 <tuv xml:lang="nl"><seg>artikel</seg></tuv>
 <tuv xml:lang="pl"><seg>artyku&#322;</seg></tuv>
 <tuv xml:lang="pt"><seg>artigo</seg></tuv>
 <tuv xml:lang="ro"><seg>articolul</seg></tuv>
 <tuv xml:lang="sk"><seg>&#269;l&#225;nok</seg></tuv>
 <tuv xml:lang="sl"><seg>&#269;len</seg></tuv>
 <tuv xml:lang="sv"><seg>artikel</seg></tuv>
</tu>
```

No	Freq.	Translation probabilities	Lexical weights	es	en	ar	zh	ja
1	3279	0.71 0.72 0.71 0.72 0.94	1.00 1.00 1.00 1.00 0.64	.	.	.	。	。
2	457	0.10 0.10 0.10 0.10 0.93	1.00 1.00 1.00 1.00 0.41	.	.	.	。	です 。
3	161	0.72 0.65 0.61 0.85 0.65	0.99 0.95 0.94 0.75 0.82	Dónde	Where	أين	哪	どこ
4	143	0.03 0.03 0.03 0.03 0.86	1.00 1.00 1.00 1.00 0.35	.	.	.	。	ます 。
5	125	0.93 0.94 0.92 0.93 0.91	0.99 0.95 0.99 0.90 0.88	Japón	Japan	اليابان	日本	日本
6	81	0.95 0.95 0.98 0.98 0.95	0.99 0.98 1.00 1.00 1.00	Tokio	Tokyo	طوكيو	东京	東京
7	70	0.02 0.02 0.02 0.02 0.93	0.99 1.00 0.98 0.99 0.63	.	.	.	。	です
8	57	0.01 0.01 0.01 0.01 0.55	0.99 1.00 0.98 0.99 0.71	.	.	.	。	を
9	54	0.71 0.71 1.00 0.68 0.69	1.00 1.00 1.00 0.97 1.00	pasaporte	passport	جواز	护照	パスポート
10	48	0.01 0.01 0.01 0.01 0.79	0.99 1.00 0.98 0.99 0.66	.	.	.	。	の
11	44	0.01 0.01 0.01 0.01 0.90	0.99 1.00 0.98 0.99 0.84	.	.	.	。	が
12	43	0.98 0.98 0.98 0.98 1.00	0.88 1.00 0.98 0.89 0.97	aeropuerto	airport	المطار	机场	空港
13	34	1.00 1.00 1.00 1.00 1.00	1.00 0.98 1.00 1.00 1.00	Chicago	Chicago	شيكاغو	芝加哥	シカゴ
14	33	0.75 0.97 0.59 0.59 0.60	1.00 0.84 0.99 0.94 1.00	Nueva York	New York	نيويورك	纽约	ニューヨーク
15	33	0.01 0.01 0.01 0.01 0.59	0.99 1.00 0.98 0.99 0.69	.	.	.	。	に
16	32	0.01 0.01 0.01 0.46 0.01	1.00 1.00 1.00 1.00 0.64	.	.	.	了 。	。
17	29	0.94 0.94 0.94 1.00 0.94	0.96 1.00 1.00 0.98 1.00	Boston	Boston	بوسطن	波士顿	ボストン
18	27	0.96 0.96 1.00 0.96 0.96	1.00 0.97 1.00 1.00 1.00	Londres	London	لندن	伦敦	ロンドン
19	26	0.90 0.96 0.90 0.93 0.90	1.00 1.00 1.00 1.00 0.97	Tanaka	Tanaka	تاناكا	T a n a k a	タナカ
20	22	0.10 0.09 0.08 0.54 0.09	0.99 0.95 0.94 0.26 0.82	Dónde	Where	أين	在	どこ
21	21	1.00 1.00 1.00 1.00 1.00	1.00 1.00 1.00 1.00 1.00	Yamada	Yamada	يامادا	Y a m a d a	ヤマダ
22	21	0.00 0.00 0.00 0.51 0.01	1.00 1.00 1.00 1.00 0.64	.	.	.	的 。	。
23	20	1.00 0.91 0.87 0.95 0.87	0.96 0.99 0.73 0.72 0.86	hoy	today	اليوم	今天	今日
24	19	0.90 1.00 0.86 0.86 0.86	1.00 1.00 1.00 1.00 1.00	Miami	Miami	ميامي	迈阿密	マイアミ

Collocations	Décompte
(_)	957
monsieur _ président	612
aujourd’ hui	584
états membres	478
ne _ pas	331
union européenne	304
« _ »	291
chers collègues	101
nations unies	96
j’ ai	87
parlement européen	80
ad hoc	32
vifs applaudissements	29
⋮	⋮

1 *Contribution à l'étude des mots rares*

2 *Contribution à l'alignement de séquences de mots*

- Ébauche de la méthode : comment, quoi ?
- Extraction d'alignements à partir de corpus multilingues
- Calcul des scores
- Optimisation de la méthode

3 *Évaluation et perspectives*

Deux évaluations particulières

3 aligneurs :

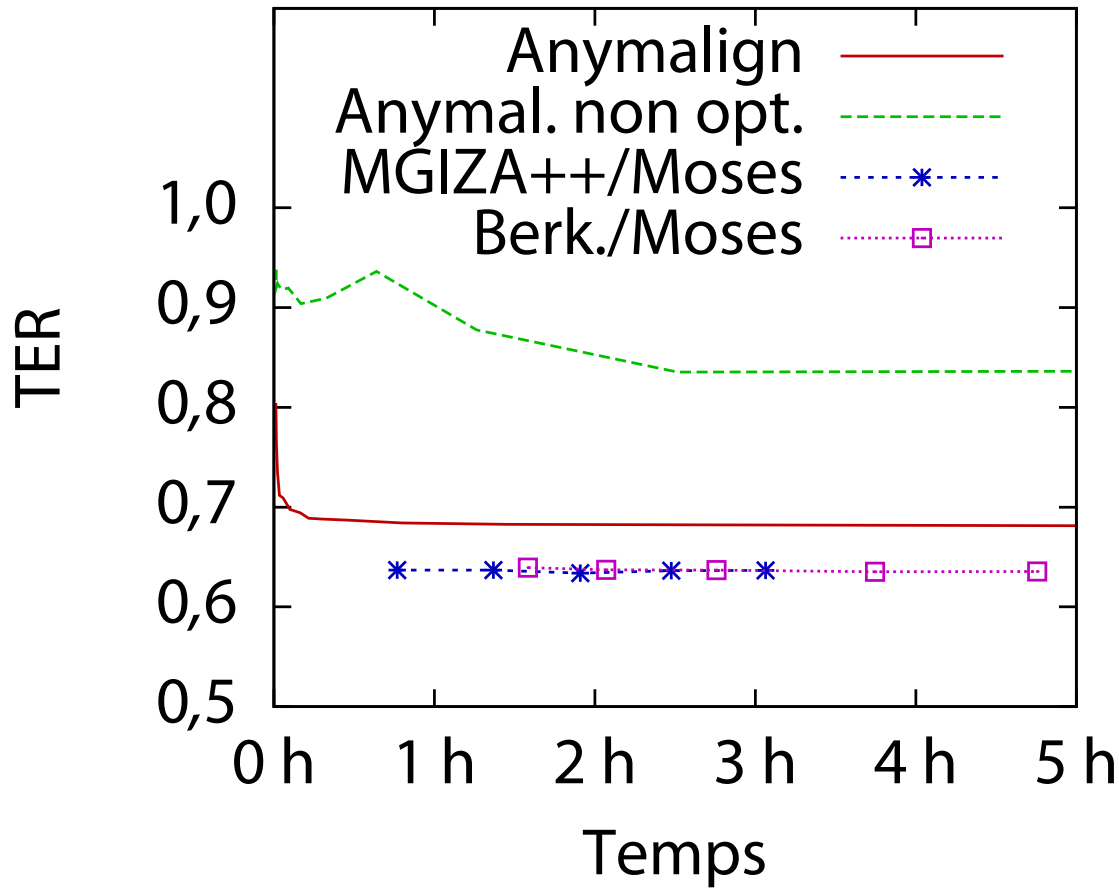
1. Anymalign
2. MGIZA++ (Gao et Vogel, 2008 ; Och et Ney, 2003 ; Al-Onaizan et coll., 1999 ; Brown et coll., 1993)
3. BerkeleyAligner (Liang et coll., 2006)

2 protocoles :

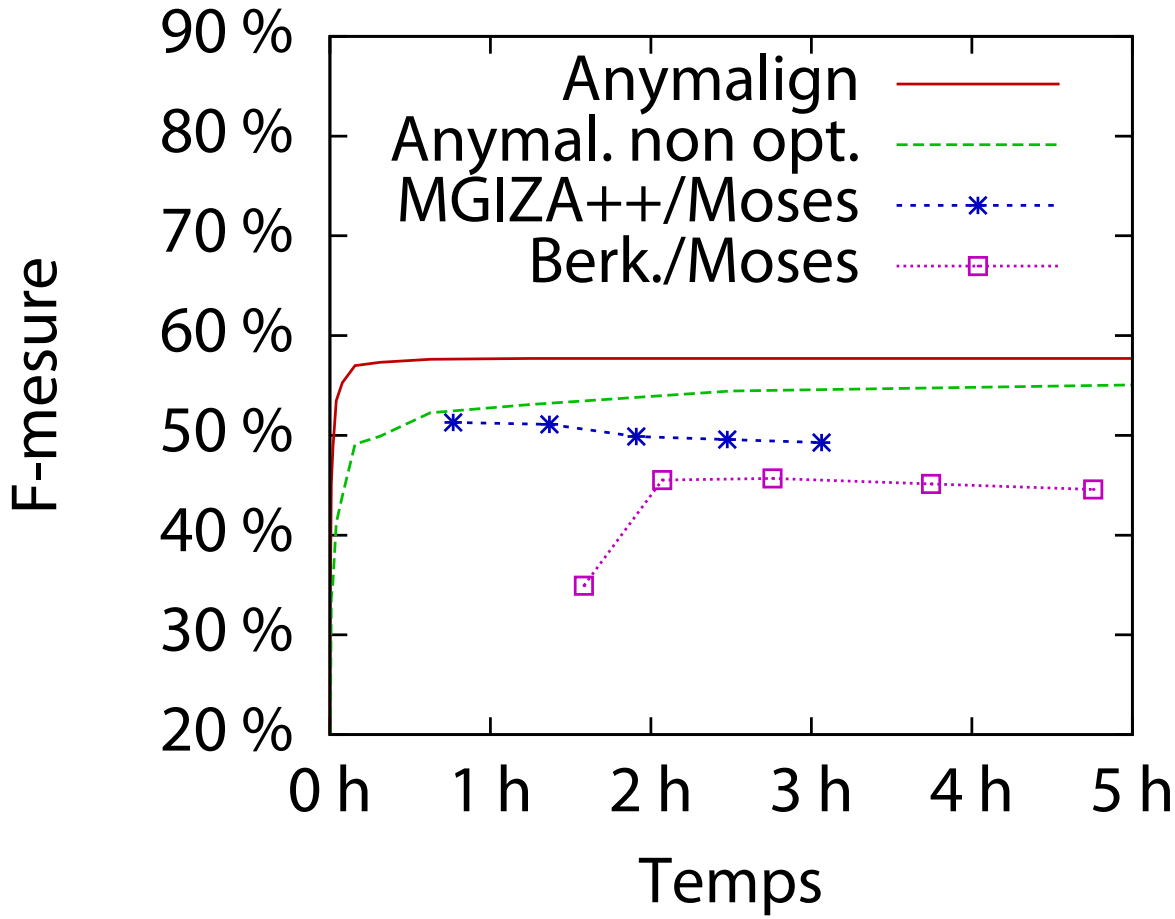
1. Traduction automatique statistique par segments
   + TER ou BLEU
2. Induction de lexiques bilingues
   + Rappel, précision, f-mesure

Entrée : 100 000 couples de phrases espagnol-français issus d'Europarl (≈ 30±18 mots/phrase)

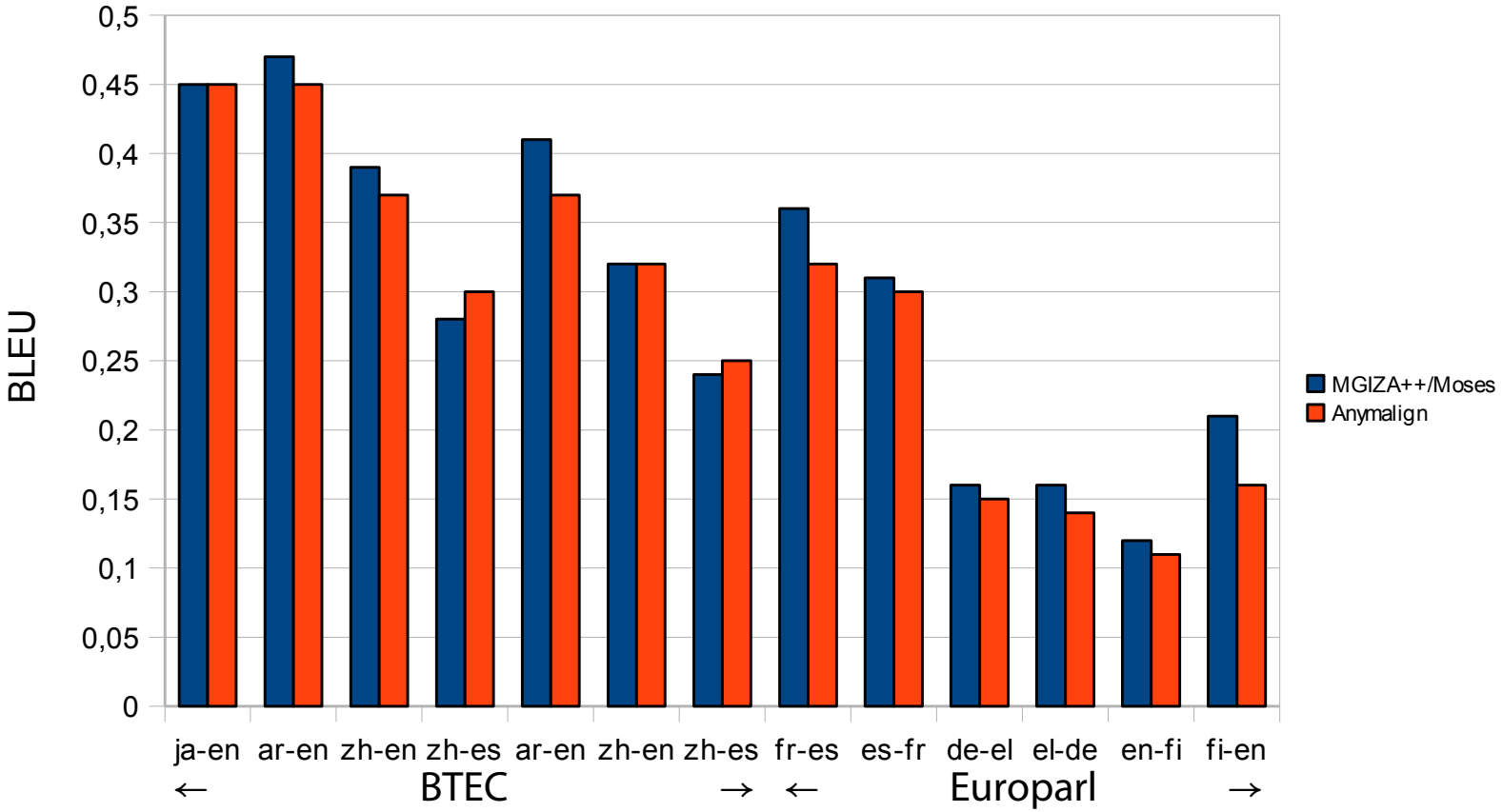
Résultats en traduction automatique



Résultats en induction de lexiques bilingues



Plus d'expériences de traduction automatique



Anymalign aboutit à des résultats égaux ou supérieurs dans 4 tâches sur 13 (en moyenne : −0,02 en BLEU)

Plus d'expériences d'induction de lexiques

Différence en F-mesure entre Anymalign et MGIZA++/Moses sur 42 couples de langues issus de la Bible (Resnik et coll., 1999)

	dan	eng	fin	fra	spa	swe	zho
dan		+ 3	−10	−15	+ 9	+8	− 4
eng	−15		−10	+13	+ 2	−6	− 5
fin	+ 2	+36		+70	+53	+7	+11
fra	−15	0	− 2		+ 1	−3	+ 5
spa	− 9	+15	+ 3	+13		−2	+15
swe	− 4	+ 7	−18	+19	+ 7		− 1
zho	−13	+16	0	+58	+31	+3	

Anymalign est meilleur dans 24 tâches sur 42 (en moyenne : + 7 %)

Des résultats a priori contradictoires

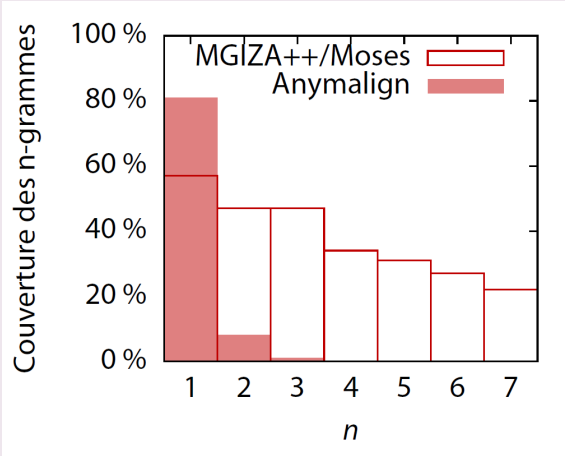
Les lexiques produits par Anymalign sont de qualité supérieure,
pourtant ils mènent à de moins bons scores
en traduction automatique statistique par segments

Des résultats a priori contradictoires

Les lexiques produits par Anymalign sont de qualité supérieure,
pourtant ils mènent à de moins bons scores
en traduction automatique statistique par segments

Des résultats qui étaient prévisibles

- Anymalign aligne principalement des mots de fréquences proches
- Il aligne donc moins de n-grammes
- Il aligne par contre davantage d'unigrammes



⇒ Problème de quantité plutôt que de qualité

Expérience en n-grammes

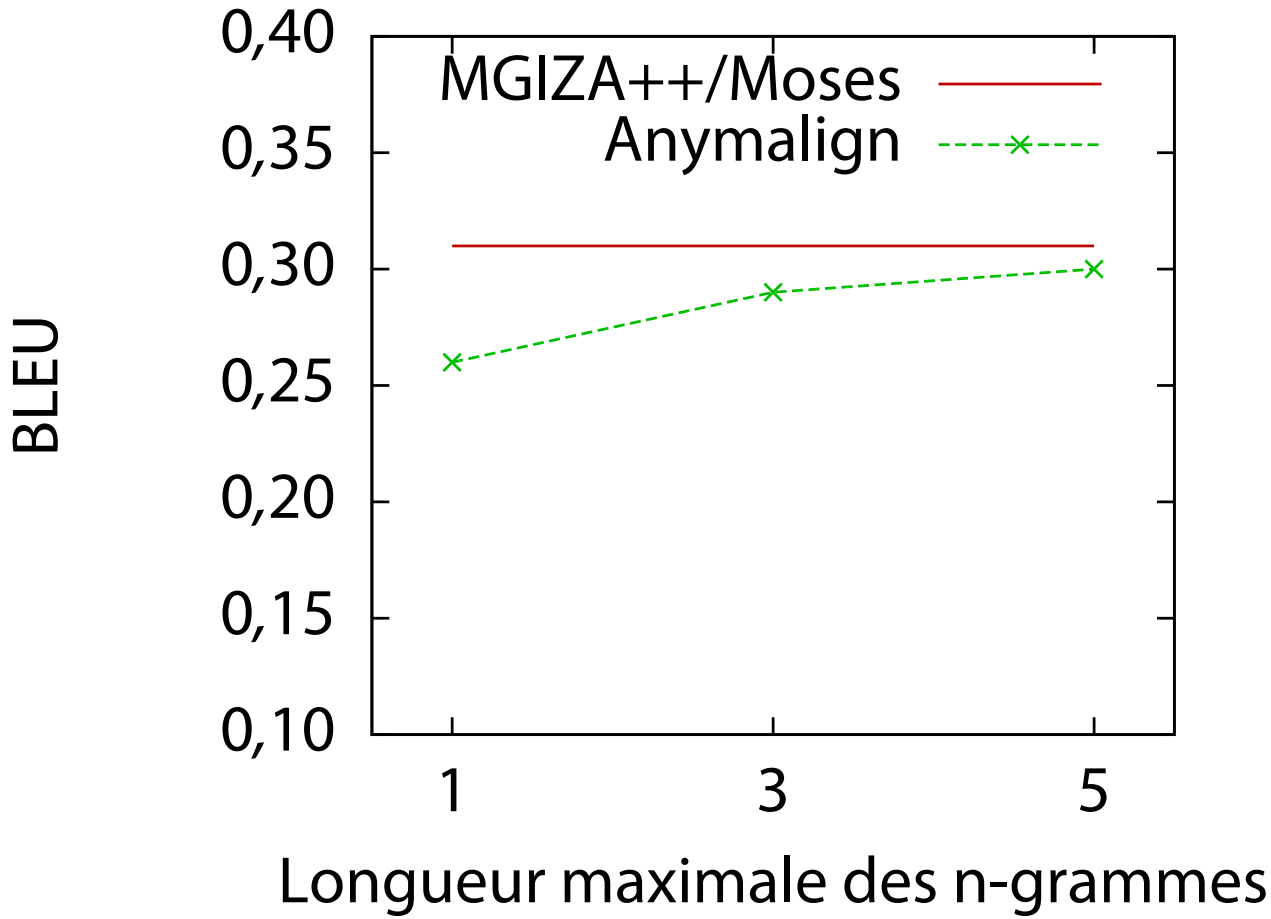
Nouvelle unité de traitement : n-gramme de mots

n	Phrase d'entrée
1	Un café , s'il vous plaît .
2	Un_café café_, ,_s'il s'il_vous vous_plaît plaît_.
3	Un_café_, café_,_s'il ,_s'il_vous s'il_vous_plaît vous_plaît_.
⋮	⋮

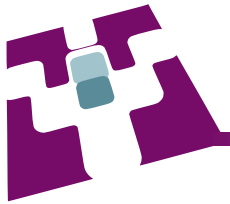
Puis alignement de toutes les combinaisons source-cible possibles jusqu'aux 5-grammes

$\Rightarrow$ 25 tables de traductions fusionnées en une seule

Résultats en n-grammes



Conclusion



Conclusion (1/2) : contributions

- Étude des mots rares en alignement
  (50 % d'hapax dans seulement 12 % des phrases)
- Nouvelle approche pour l'alignement sous-phrastique
  - multilingue : 60 langues simultanément ! (messages KDE)
  - simple
  - *anytime*
  - facilement parallélisable
- Logiciel libre sous licence GPL : Anymalign
  - téléchargé plus de 100 fois
  - PROMT, Lingua et Machina, Orange, Xerox, MITRE, Nintendo. . .
  - actuellement en cours de réimplémentation en Java à L&M

Conclusion (2/2) : principaux résultats

- À l'encontre des idées reçues
  - base de la méthode : utiliser les mots rares
  - technique : supprimer des données
- Des traitements alingues pour faire multilingue et bilingue (tables de traductions, lexiques multilingues) ou monolingue (co-occurrences et collocations)
- Résultats comparables à l'état de l'art
  avec une approche vraiment plus simple
  (−2 BP en traduction automatique, +7 % en lexique)

Publications

[1] Yves Lepage, Julien Gosme et **Adrien Lardilleux** : « The Structure of Unseen Trigrams and its Application to Language Models : a First Investigation ». In *Proceedings of IUCS 2010*, Pékin, octobre 2010. À paraître.

[2] Yves Lepage, Julien Gosme et **Adrien Lardilleux** : « Estimating the proximity between languages by their commonality in vocabulary structures ». In *LNAI*. Springer Heidelberg. À paraître.

[3] **Adrien Lardilleux**, Julien Gosme et Yves Lepage : « Bilingual Lexicon Induction : Effortless Evaluation of Word Alignment Tools and Production of Resources for Improbable Language Pairs ». In *Proceedings of LREC'10*, pages 252-256, La Valette, mai 2010.

[4] Yves Lepage, **Adrien Lardilleux** et Julien Gosme : « The GREYC Translation Memory for the IWSLT 2009 Evaluation Campaign : one step beyond translation memory ». In *Proceedings of IWSLT 2009*, pages 45-49, Tōkyō, décembre 2009.

[5] **Adrien Lardilleux** : « L'alignement sous-phrastique multilingue pour les nuls ». In *Actes de MajecSTIC 2009*, Avignon, novembre 2009.

[6] Julien Gosme, Yves Lepage et **Adrien Lardilleux** : « Translation of sublanguages by subgrammars ». In *Proceedings of EBMT3*, pages 77-84, Dublin, novembre 2009.

[7] **Adrien Lardilleux**, Jonathan Chevelu, Yves Lepage, Ghislain Putois et Julien Gosme : « Lexicons or phrase tables ? An investigation in sampling-based multilingual alignment ». In *Proceedings of EBMT3*, pages 45-52, Dublin, novembre 2009.

[8] Yves Lepage, **Adrien Lardilleux** et Julien Gosme : « Commonality across vocabulary structures as an estimate of the proximity between languages ». In *Proceedings of LTC'09*, pages 457-461, Poznań, octobre 2009.

[9] **Adrien Lardilleux** et Yves Lepage : « Sampling-based multilingual alignment ». In *Proceedings of RANLP 2009*, pages 214-218, Borovets, septembre 2009.

[10] **Adrien Lardilleux** et Yves Lepage : « Hapax Legomena : Their Contribution in Number and Efficiency to Word Alignment ». In Zygmunt Vetulani et Hans Uszkoreit, éditeurs : volume 5603 de *LNCS*, pages 440-450. Springer Heidelberg, août 2009.

[11] **Adrien Lardilleux** et Yves Lepage : « Anymalign : un outil d'alignement sous-phrastique libre pour les êtres humains ». In *Actes de TALN 2009*, Senlis, juin 2009.

[12] **Adrien Lardilleux** et Yves Lepage : « A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method ». In *Proceedings of AMTA 2008*, pages 125-132, Waikiki, octobre 2008.

[13] Yves Lepage, **Adrien Lardilleux**, Julien Gosme et Jean-Luc Manguin : « The GREYC Machine Translation System for the IWSLT 2008 Evaluation Campaign ». In *Proceedings of IWSLT 2008*, pages 39-45, Waikiki, octobre 2008.

[14] **Adrien Lardilleux** et Yves Lepage : « Multilingual Alignments by Monolingual String Differences ». In *Proceedings of Coling'08*, pages 55-58, Manchester, août 2008.

[15] Yves Lepage et **Adrien Lardilleux** : « The GREYC Machine Translation System for the IWSLT 2007 Evaluation Campaign ». In *Proceedings IWSLT 2007*, pages 49-54, Trente, octobre 2007.

[16] **Adrien Lardilleux** et Yves Lepage : « The contribution of the notion of hapax legomena to word alignment ». In *Proceedings of LTC'07*, pages 458-462, Poznań, octobre 2007.