



**HAL**  
open science

## **Anymalign : un outil d'alignement sous-phrastique libre pour les êtres humains**

Adrien Lardilleux, Yves Lepage

### ► **To cite this version:**

Adrien Lardilleux, Yves Lepage. Anymalign : un outil d'alignement sous-phrastique libre pour les êtres humains. TALN 2009, Jun 2009, Senlis, France. hal-00488772

**HAL Id: hal-00488772**

**<https://hal.science/hal-00488772>**

Submitted on 2 Jun 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## ***anymalign* : un outil d’alignement sous-phrastique libre pour les êtres humains**

Adrien Lardilleux   Yves Lepage  
GREYC, Université de Caen Basse-Normandie  
prenom.nom@info.unicaen.fr

**Résumé.** Nous présentons *anymalign*, un aligneur sous-phrastique grand public. Ses résultats ont une qualité qui rivalise avec le meilleur outil du domaine, GIZA++. Il est rapide et simple d’utilisation, et permet de produire dictionnaires et autres tables de traduction en une seule commande. À notre connaissance, c’est le seul outil au monde permettant d’aligner un nombre quelconque de langues simultanément. Il s’agit donc du *premier aligneur sous-phrastique réellement multilingue*.

**Abstract.** We present *anymalign*, a sub-sentential aligner oriented towards end users. It produces results that are competitive with the best known tool in the domain, GIZA++. It is fast and easy to use, and allows dictionaries or translation tables to be produced in a single command. To our knowledge, it is the only tool in the world capable of aligning any number of languages simultaneously. It is therefore *the first truly multilingual sub-sentential aligner*.

**Mots-clés :** alignement sous-phrastique, multilinguisme, table de traduction.

**Keywords:** sub-sentential alignment, multilinguism, translation table.

### **1 Vue d’ensemble**

L’alignement sous-phrastique consiste à extraire des traductions d’unités textuelles de grain inférieur à la phrase à partir de textes multilingues dont les phrases ont préalablement été mises en correspondance. Le plus connu des outils disponibles est sans conteste GIZA++ (Och & Ney, 2003). Il implémente les modèles probabilistes IBM. Bien que produisant des résultats d’une grande qualité, beaucoup de ses utilisateurs sont loin d’en saisir toutes les subtilités, car de tels modèles sont d’une grande complexité. Cette complexité se répercute sur l’interface d’utilisation de l’outil.

Pour répondre à toutes ces difficultés, nous mettons à disposition de tous un aligneur sous-phrastique qui se distingue de GIZA++ selon plusieurs points. Il est :

**réellement multilingue :** il permet d’aligner un nombre quelconque de langues simultanément. Contrairement aux modèles d’alignement traditionnels, la méthode sous-jacente n’est pas nécessairement bilingue. Elle repose sur un traitement *alingue*, ce qui la rend en pratique multilingue ;

**simple d’utilisation :** une *unique* commande suffit dans la plupart des cas. Il n’y a pas de paramètre superflu ;

**rapide :** la qualité des alignements produits ne dépend pas du temps d’exécution. C’est la couverture du texte d’origine par les alignements qui augmente en fonction du temps. L’utilisateur peut à tout moment interrompre le traitement et récolter tous les alignements



*anymalign* : un outil d'alignement sous-phrastique libre pour les êtres humains

- (a) la séquence ordonnée constituée des mots du groupe ;
- (b) la séquence complémentaire de cette séquence ordonnée également.

Par exemple, le groupe 2 permet d'extraire la séquence ordonnée « café<sub>1</sub> coffee<sub>2</sub> قهوة<sub>3</sub> » et son complémentaire « Un<sub>1</sub> \_ ,1 s'il<sub>1</sub> vous<sub>1</sub> plaît<sub>1</sub> .1 One<sub>2</sub> \_ ,2 please<sub>2</sub> .2 3. 3 فضلك<sub>3</sub> من<sub>3</sub> » de la ligne 1. Il permet aussi d'extraire « café<sub>1</sub> coffee<sub>2</sub> قهوة<sub>3</sub> » et son nouveau complémentaire « Ce<sub>1</sub> \_ est<sub>1</sub> excellent<sub>1</sub> .1 This<sub>2</sub> \_ is<sub>2</sub> excellent<sub>2</sub> .2 3. 3 ممتازة<sub>3</sub> \_ هذه<sub>3</sub> » de la ligne 3. Chaque groupe est donc susceptible de produire deux alignements par ligne ;

5. répéter les étapes 2 à 4 tant que de nouveaux alignements sont produits, ou jusqu'à ce que l'utilisateur interrompe le processus. Le nombre de fois que chaque alignement a été obtenu permet de déduire les probabilités de traduction associées. Des poids lexicaux peuvent également être calculés à partir des cooccurrences de mots dans le corpus initial.

### 3 En pratique

Le programme implémentant la méthode décrite ci-dessus est téléchargeable à l'adresse :

<http://users.info.unicaen.fr/~alardill/anymalign/>

Il permet de traiter des quantités de données considérables tout en minimisant les ressources nécessaires. Étant donné un corpus multilingue disponible dans des fichiers séparés avec une phrase par ligne, la simple commande :

```
python anymalign.py langue1.txt langue2.txt [...]
```

suffit pour obtenir une table de traduction. La figure 1 donne un exemple de résultat en HTML.

No	Freq.	Translation probabilities	Lexical weightings	ja	en	ar
1	50759	0.98 0.74 0.75	0.95 0.97 0.98	か	?	؟
2	47610	0.83 0.71 0.72	0.64 1.00 1.00	。	.	.
3	8161	0.95 0.97 0.93	0.83 0.96 0.94	どこ	Where	أين
4	6446	0.92 0.10 0.10	0.41 1.00 1.00	です。	.	.
5	5372	0.92 0.08 0.08	0.40 1.00 1.00	か。	?	؟
6	3344	0.99 0.99 0.99	1.00 0.99 1.00	東京	Tokyo	طوكيو
7	3134	0.96 0.97 0.96	0.73 0.95 0.99	日本	Japan	اليابان

FIG. 1 – Exemple d'un début de sortie d'*anymalign* sur un corpus japonais-anglais-arabe. Les alignements sont triés par fréquence. Les meilleurs alignements sont mis en évidence par la couleur.

### Références

LARDILLEUX A. & LEPAGE Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. Dans *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, p. 125–132, Waikiki, Hawai'i, États-Unis.

LARDILLEUX A. & LEPAGE Y. (2009). Sampling-based multilingual alignment. Dans *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgarie. À paraître.

OCH F. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**, 19–51.