Université de Strasbourg

Traduction automatique

TA statistique : alignement phrastique

Enzo Doyen

enzo.doyen@unistra.fr

2025-09-19

🕮 Plan

I. Tâche d’alignement phrastique

Corpus parallèles, mais pas nécessairement alignés

- Même si un corpus est la traduction d'un autre, il est nécessaire de connaitre quelles phrases dans la source correspondent à quelles phrases dans la cible.
- Il n'y a pas forcément une correspondance 1-1 entre les phrases source et cible :
  - une phrase source peut être traduite par plusieurs phrases cibles (p. ex., phrase longue divisée en plus courtes) ;
  - une phrase source peut être omise dans la cible ;
  - …

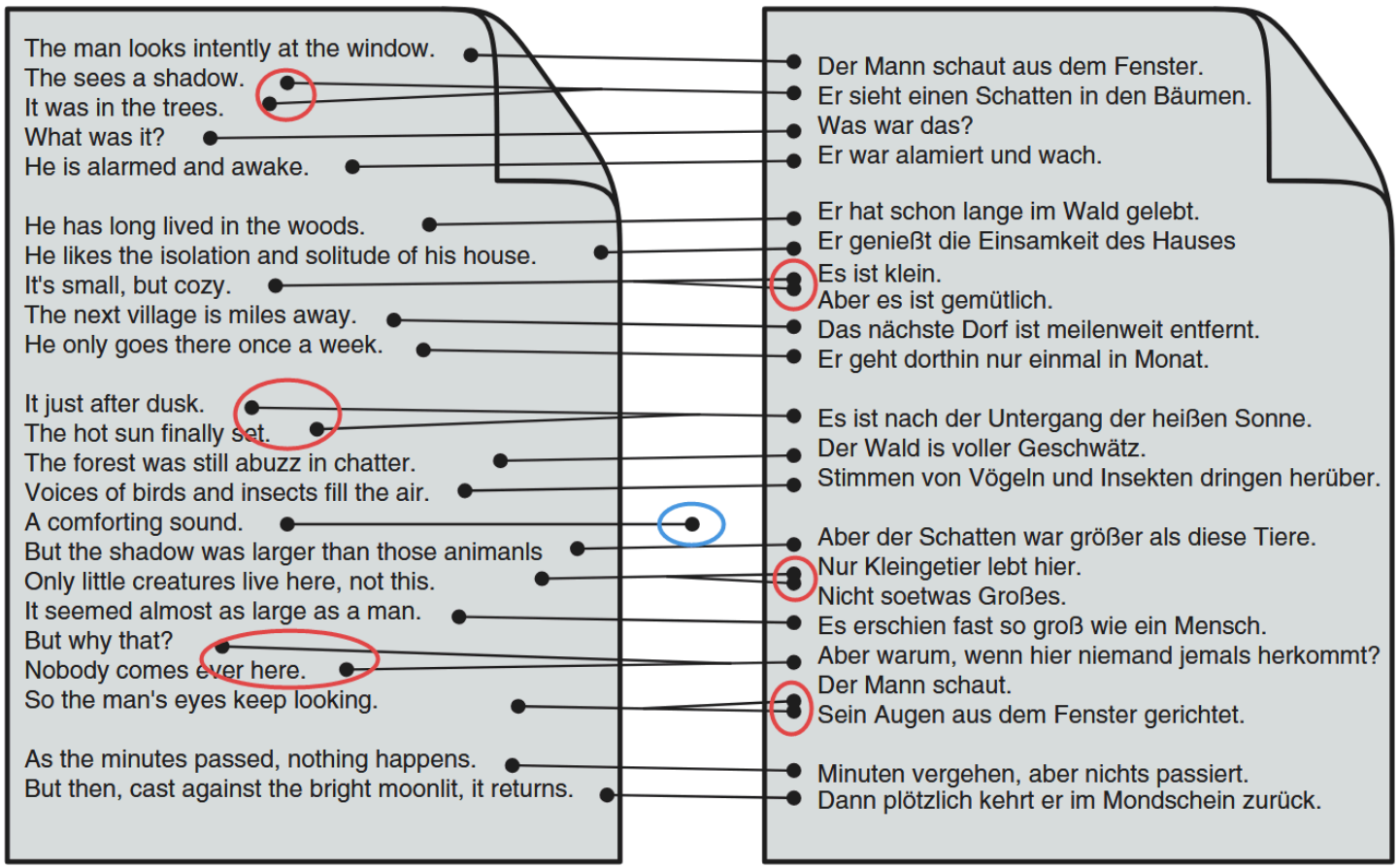
Corpus parallèles, mais pas nécessairement alignés

Il est nécessaire d'**aligner** les corpus pour pouvoir les exploiter statistiquement.

Cela implique des **correspondances** au niveau des phrases : ce qu'on appelle l'**alignement phrastique**.

À un plus petit niveau, cela se fait sous forme d'**alignement sous-phrastique**.

Exemple d'alignement



Source : **Koehn (2009)**

Tâche d'alignement

E1
The crisis our farmers are in right now will affect all of us at a certain point in time.

E2
We are all consumers and we all need a strong and healthy agricultural sector.

E3
I am glad that the Hon. Member for Algoma (Mr. Foster) mentioned figures in his remarks.

E4
Otherwise, the Government might have eluded the problem once again.

F1
La crise que vivent en ce moment nos agriculteurs se répercutera sur tous et chacun de nous à un certain moment.

F2
Nous sommes des consommateurs.

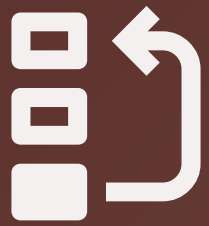
F3
Nous avons tous besoin d'une agriculture saine et forte.

F4
Heureusement que le député d'Algoma (M Foster) a mentionné des chiffres dans ses remarques, sans cela ce gouvernements'en serait sorti en douce encore une fois.

Tâche d'alignement : indices (anglais-maltais)

Where it is necessary to publish the Official Journal when the information system of the Publications Office is not operational pursuant to a disruption as referred to in paragraph 1, only the printed edition of the Official Journal shall be authentic and shall produce legal effects.

Fejn ikun meħtieġ li jiġi ppubblikat il-Ġurnal Uffiċjali meta s-sistema ta' informazzjoni fl-Uffiċċju tal-Pubblikazzjonijiet ma tkunx operattiva minħabba interruzzjoni kif imsemmi fil-paragrafu 1, l-edizzjoni stampata biss tal-Ġurnal Uffiċjali għandha tipproduċi effetti legali.



II. Algorithmes d'alignement et outils

Algorithmes d'alignement : Gale et Church (1993)

- Critère de base : **longueur de phrases**.
- On cherche à minimiser la différence de longueur entre les phrases source et cible (généralement, les phrases source courtes sont traduites par des phrases cibles courtes, idem pour les phrases longues).

Type d'alignement	Description
1:1	Une phrase pour une phrase
1:0	Phrase supprimée dans la cible
0:1	Phrase ajoutée dans la cible
2:1	2 phrases source fusionnées en 1 phrase cible
1:2	1 phrase source divisée en 2 phrases cibles

Algorithmes d'alignement : Gale et Church (1993)

Gale et Church (1993) utilisent le nombre de **caractères** pour évaluer la longueur des phrases.

Étant donné que, d'une langue à l'autre, le nombre de caractères peut varier, on utilise un **facteur de normalisation** pour chaque paire de langues.

L'algorithme évalue ensuite la probabilité de la différence de longueur entre les candidats d'alignement pour toutes les configurations possibles (1-1, 1-0, 0-1, etc.), et choisit la solution la plus probable.

Algorithmes d'alignement : Kraif (2001) ; Simard et al. (1992)

Amélioration par rapport à **Gale et Church (1993)** : utilisation de **cognats** et d'**ancres lexicales** pour améliorer la précision de l'alignement.

Cognats : mots qui ont la même origine dans deux langues différentes, et proches orthographiquement (en français-anglais : « liste »/« list », « fonction »/« function », etc.).

Ancres lexicales : éléments linguistiques identiques dans les deux langues, qui permettent de faire le lien entre les phrases source et cible (chiffres, noms propres, signes de ponctuation, abréviation, signes mathématiques…).

Algorithmes d'alignement : ancres lexicales

EN	AR
Meetings normally should be held during regular meeting hours, namely, from 10 a.m. to 1 p.m. and from 3 p.m. to 6 p.m., on working days	تُعقد الاجتماعات عادة خلال ساعات الاجتماعات العادية، أي من الساعة 00/10 إلى الساعة 00/13 ومن الساعة 00/15 إلى الساعة 00/18، أيام العمل؛

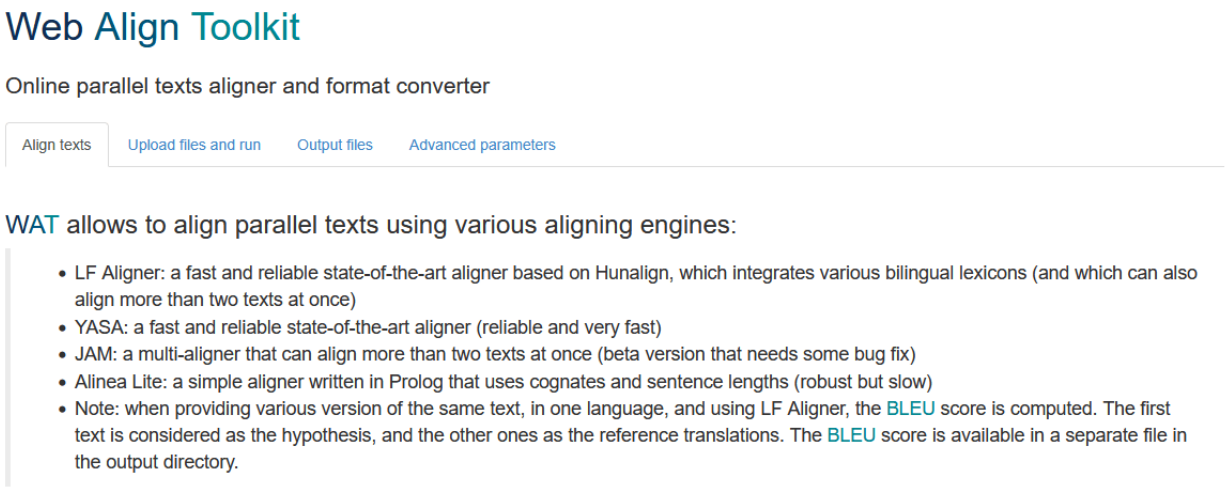
Algorithmes d'alignement : Hunalign (Varga et al., 2007)

- Permet d'exploiter un dictionnaire bilingue si fourni, ce qui améliore la précision de l'alignement.
- En plus des correspondances provenant du dictionnaire, évalue les cognats/ancres lexicales, et la différence en longueur.

GitHub : https://github.com/danielvarga/hunalign

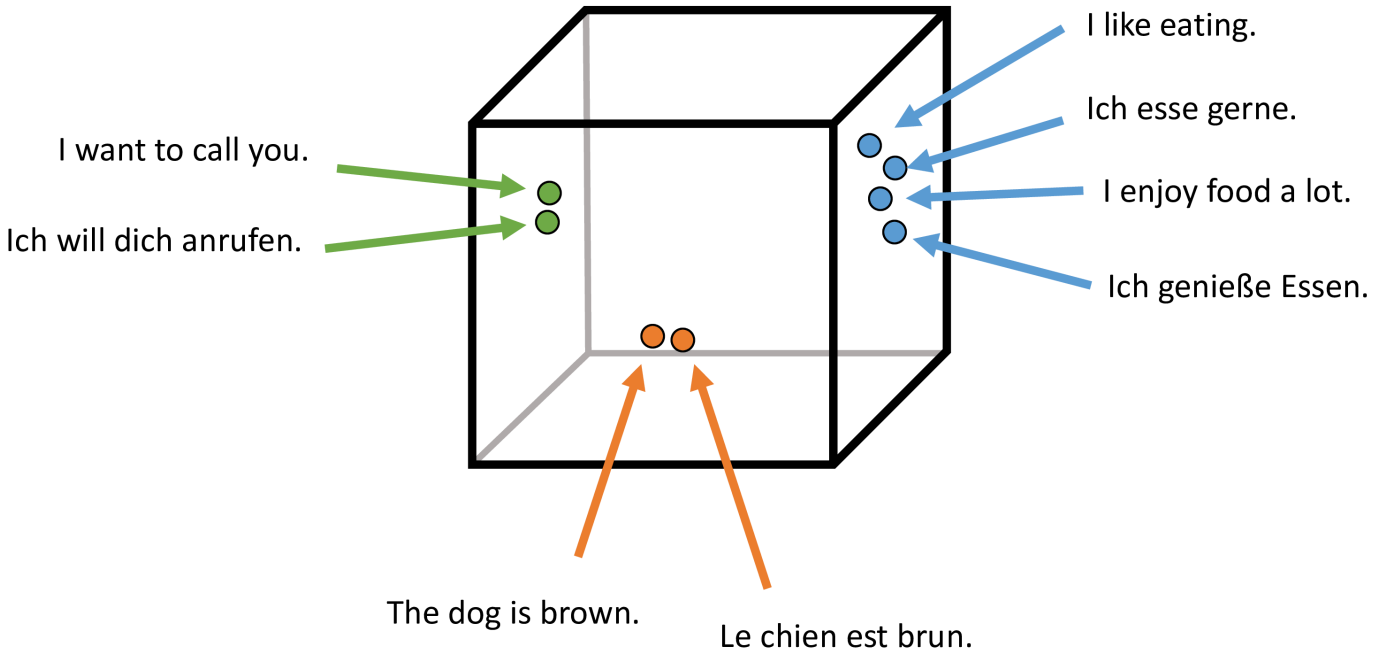
WAT : Web Align Toolkit

- Interface en ligne donnant accès à plusieurs aligneurs, développée à l'Université de Grenoble.
- Offre une implémentation de Hunalign (LF Aligner) et d'autres algorithmes d'alignement.



http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit/

Alignement neuronal



Source : https://github.com/thompsonb/vecalign

Alignement neuronal : Vecalign (Thompson et Koehn, 2019)

- Compare des représentations vectorielles des phrases, encodées avec un réseau récurrent (BiLSTM).
- Mesure la distance entre les représentations de candidats d'alignement dans un espace vectoriel.
- Favorise les alignements entre phrases dont la représentation est proche.
- Fonctionne dans près de 100 langues.

GitHub : https://github.com/thompsonb/vecalign

Alignement neuronal : BERTAlign

- Comme Vecalign, mais utilise des représentations de phrases encodées avec BERT, à l'aide de la bibliothèque sentence-transformers **(Reimers et Gurevych, 2019)**.

GitHub : https://github.com/bfsujason/bertalign

Notebook d'exemple : https://colab.research.google.com/drive/123GhXwgwmQp1F5SVZ74_uIgyxo6hLRq0?usp=sharing

III. Mise en pratique

Mise en pratique

Sur Moodle, vous trouverez un corpus parallèle anglais-français de *Madame Bovary*.

Utilisez ce corpus pour tester les différents outils d'alignement présentés précédemment :

- WAT (Web Align Toolkit) et les différents algorithmes ;
- BERTAlign (utilisez le notebook d'exemple).

Examinez les erreurs d'alignement : quels patrons d'erreur peut-on discerner ? Quelles sont les causes possibles de ces erreurs ?

Bibliographie

Gale, W. A., et Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75-102.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press. **10.1017/CBO9780511815829**

Kraif, O. (2001). Exploitation Des Cognats Dans Les Systèmes d'alignement Bi-Textuel : Architecture et Évaluation. *Revue TAL : traitement automatique des langues*, *42*(3), 833-867.

Reimers, N., et Gurevych, I. (août 2019). *Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks* (Numéro arXiv:1908.10084). arXiv. **10.48550/arXiv.1908.10084**

Simard, M., Foster, G. F., et Isabelle, P. (juin 1992). Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

Thompson, B., et Koehn, P. (novembre 2019). Vecalign: Improved Sentence Alignment in Linear Time and Space. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1342-1348. **10.18653/v1/D19-1136**

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., et Trón, V. (décembre 2007). Parallel Corpora for Medium Density Languages. *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*.

♡ Remerciements

- Pablo Ruiz Fabo pour le contenu de certaines diapositives.