



HAL
open science

Alignement de textes bilingues par classification ascendante hiérarchique

Maria Zimina

► **To cite this version:**

Maria Zimina. Alignement de textes bilingues par classification ascendante hiérarchique. JADT 2000, Ecole Polytechnique Fédérale de Lausanne, Mar 2000, Lausanne, Suisse. pp.171-178. hal-01224603

HAL Id: hal-01224603

<https://u-paris.hal.science/hal-01224603>

Submitted on 10 Mar 2022

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.

Alignement de textes bilingues par classification ascendante hiérarchique

Maria ZIMINA

LEXICO (SYLED) – EA 2290 Université de la Sorbonne Nouvelle – Paris 3 (France)

Abstract

Existing translations contain a wealth of ready-made solutions that can be reused to generate new high-quality translations. For this reason, translation resources are frequently stored in electronic databases providing certain information retrieval facilities. The concept of bilingual text alignment enables a more efficient use of the translation resources, by reconstructing the links maintaining translation equivalence between the corresponding segments of the text and its translations in different languages.

Current text alignment algorithms perform quite successfully on a sentence level. However, there is a need to continue research in finer-grained text alignment. In this regard, we propose to identify translation correspondences on the basis of hierarchical cluster analysis of graphical forms and repeated segments of bilingual texts. The principles of this technique enable to yield, through progressive agglomeration, clusters of textual units with similar (or identical) distributional profiles. The results obtained following this technique suggest that hierarchical cluster analysis can be applied for a wide range of purposes in bilingual text alignment.

Résumé

Le stockage électronique conjoint de textes originaux avec leurs traductions existantes facilite le travail quotidien du traducteur en mettant à sa disposition des solutions toutes faites aux nombreux problèmes de traduction. La conversion d'un ensemble de documents en une base de données bi-textuelles exige l'élaboration de techniques d'alignement. Il faut, donc, introduire une dimension interactive en reconstituant automatiquement des liens entre un grand nombre d'éléments du texte original et sa traduction.

Les algorithmes développés pour calculer automatiquement une représentation bi-textuelle ne sont pas encore capables de rendre explicites toutes les correspondances de traduction dans un couple de textes donnés. Notre travail est orienté vers une étude de nouvelles méthodes statistiques d'alignement à base de classification hiérarchique ascendante des formes graphiques et des segments répétés. Les procédures de classification permettent d'agréger successivement formes et segments en fonction de leur répartition dans l'ensemble d'un corpus. Ce genre de regroupement est potentiellement utilisable pour la mise en correspondance de textes bilingues.

Mots-clés : corpus bilingues, bi-texte, correspondances de traduction, alignement, concordances bilingues, classification hiérarchique ascendante, formes graphiques, segments répétés, profils de répartition.

1. Corpus bilingues et traduction

Avec la croissance du marché de la traduction, les agents économiques, les organisations internationales s'intéressent de plus en plus à l'archivage électronique conjoint de textes et de leurs traductions dans différentes langues. Ces documents représentent le noyau de la communication multilingue et rendent possible l'échange d'information entre communautés. L'information qu'ils contiennent revêt une importance capitale dans plusieurs domaines socio-économiques. C'est pourquoi de vastes corpus de textes sont systématiquement archivés dans les textothèques et bases de données informatiques. Ces banques textuelles sont ensuite consultées pour récupérer des informations sur des références terminologiques ou bien pour comparer plusieurs versions d'un même document. Le problème est alors de disposer d'un accès rapide et efficace à l'information contenue dans ces documents. L'archivage électronique des données textuelles ainsi que la création de systèmes de recherche documentaire (information retrieval) fournissent une solution partielle à ce problème. Néanmoins, pour les

rendre facilement consultables et pour pouvoir exploiter complètement les ressources présentes dans ces documents, il est nécessaire d'établir un système de mise en relation entre segments correspondants dans des couples de textes. Pour résoudre ce problème, les aides informatisées sont indispensables.

2. Mise en correspondance de textes bilingues

On appelle *corpus bilingues* des corpus constitués de paires de textes dont l'un est une traduction de l'autre. Il s'agit, en général, de textes sources et de traductions (effectuées par des traducteurs humains) présentés sous forme électronique. Ce type de corpus est souvent appelé *bi-texte* (Harris, 1988). La conversion d'un ensemble de documents en une base de données bi-textuelles exige l'élaboration de techniques d'*alignement*. Il faut, donc, introduire une dimension interactive en ajoutant des liens entre un grand nombre d'éléments d'un texte original et sa traduction. Une fois ces liens établis, on peut créer une variété d'outils d'analyse et mettre en évidence certaines régularités de traduction. Les textes bilingues alignés deviennent plus facilement utilisables. L'exploitation des données de traduction contenues dans les corpus bilingues permet d'automatiser certaines étapes de la traduction et notamment de développer des méthodes permettant la reconstitution automatique des *correspondances de traduction*.

2.1. Les outils de réutilisation des ressources de traduction

Pour assurer une mise en correspondance interactive entre les segments de textes bilingues, une nouvelle génération d'aides à la traduction informatisée à *base de corpus* a été conçue. Il s'agit, notamment, de programmes de *concordances bilingues* (Isabelle, 1992). Ces programmes permettent d'extraire à partir du gisement des traductions existantes de l'information et des solutions utilisables pour la production de nouvelles traductions. La création de ces outils a été rendue possible grâce à l'intégration de modèles statistiques et linguistiques capables d'*aligner* les segments correspondants (paragraphe, phrases, syntagmes, et parfois mots) de deux textes avec un taux de précision relativement élevé. Les algorithmes développés pour calculer automatiquement une représentation bi-textuelle à partir d'un texte et de sa traduction, permettent de construire des bases de données de textes alignés. Cependant, au stade actuel, ces algorithmes ne sont pas encore capables de rendre explicites *toutes* les correspondances de traduction dans un couple de textes donnés. Malgré les progrès récents dans le domaine d'alignement, la mise en correspondance des matériaux textuels, de textes bilingues, ou corpus *bi-textuels*, reste relativement complexe et exige que les recherches soient poursuivies dans ce domaine.

3. L'alignement à base de classification ascendante hiérarchique

Notre travail est orienté vers une étude de nouvelles méthodes statistiques d'*alignement à base de classification hiérarchique ascendante automatisée des formes graphiques et des segments répétés* (Lebart et Salem, 1994). Les procédures de classification permettent d'agréger successivement les formes et segments en fonction de leur *répartition* dans l'ensemble d'un corpus. Appliquée au *tableau lexical entier* (TLE), qui range les décomptes des occurrences de l'ensemble de formes et segments dans chacune des parties du corpus, la classification hiérarchique ascendante produit des regroupements d'éléments caractérisés par des profils de répartition similaires (ou identique). Globalement, l'étude statistique de leur comportement dans les deux parties bilingues d'un corpus pourrait aider à identifier les formes et segments représentant des traductions mutuelles (cf. figures 1-2).

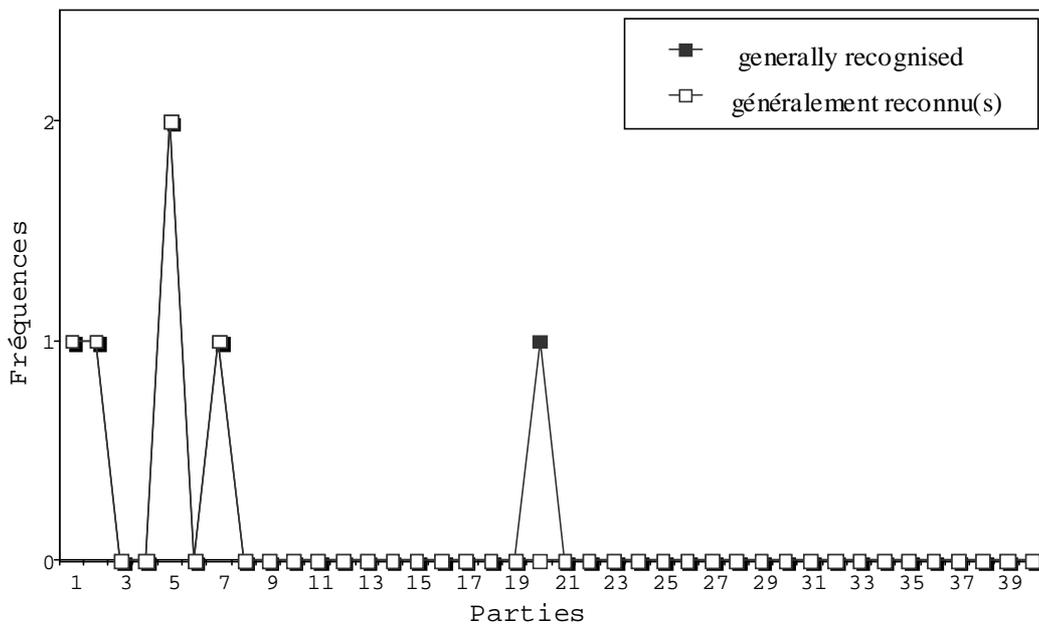
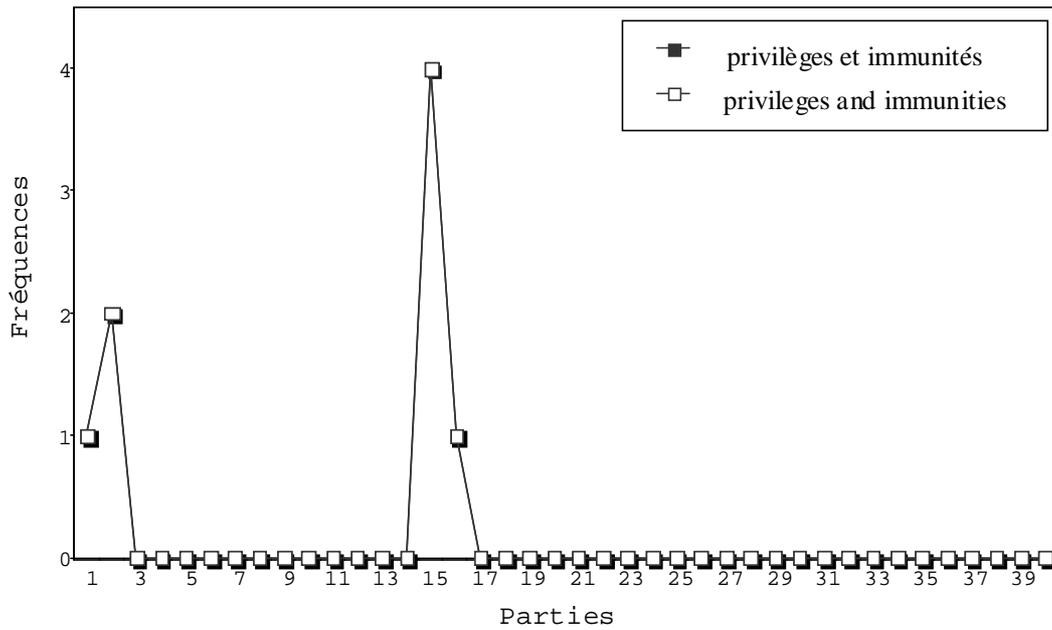


Figure 1. Profils graphiques des segments répétés agrégés dans les mêmes classes

1./ont à la charge du conseil de l' europe. *privilèges et immunités* des juges les j/
 /shall be borne by the council of europe. *privileges and immunities* of judges the/
 /dant l' exercice de leurs fonctions, des *privilèges et immunités* prévus à l' art/
 / the exercise of their functions, to the *privileges and immunities* provided for /

2./ is quite distinct from the authorities' *generally recognised* discretion to make/
 /ent distincte du pouvoir discrétionnaire *généralement reconnu* à l' administratio/
 /ut also whether they had duly observed'' *generally recognised* legal and administ/
 /s principes juridiques et administratifs *généralement reconnus*''(45). une derni/

Figure 2. Retours au contexte

3.1. Description du corpus

Les expérimentations présentées portent sur le corpus textuel bilingue (français/anglais) constitué à partir de la *Convention de sauvegarde des Droits de l'Homme et des libertés fondamentales*, ainsi que d'une douzaine de protocoles, et de 36 arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995. La Convention a été signée à Rome le 4 novembre 1950. Élaborée au sein du Conseil de l'Europe, elle définit un certain nombre de droits fondamentaux et institue un mécanisme de contrôle et de sanction propre à assurer le respect de ces droits par les Etats signataires. Il existe deux versions officielles des textes mentionnés ci-dessus : l'une en français, l'autre en anglais. Les deux versions de chaque document existent parallèlement, et il est impossible de distinguer une langue source et une langue cible. Les corpus anglais (273 685 occurrences) et français (285 961 occurrences) sont découpés en 12 131 phrases (une phrase correspond à la séquence comprise entre deux retours à la ligne) (Bourigault et al., 1999).

3.2 Mécanisme de la classification

Il existe une grande variété des méthodes de classification hiérarchique, puisqu'il y a plusieurs façons de calculer le poids et les distances par rapport aux éléments à classer. Nous avons utilisé une des variantes, basée sur la *méthode des voisins réciproques* proposée par J. Juan (1982).

Pour travailler sur l'ensemble des formes et segments répétés du corpus bilingue, on a fusionné en un seul tableau lexical les données générées séparément pour les corpus anglais et français. Les nombres à l'intersection des lignes et des colonnes de ce tableau correspondent aux sous-fréquences des formes et segments répétés dans chacune de parties du corpus. La classification automatique projetée sur les lignes du tableau (formes et segments) décrit leurs proximités en les regroupant en classes. Les regroupements effectués à chaque pas de l'algorithme de classification hiérarchique rassemblent des éléments qui sont plus au moins proches entre eux. Une *classe* est un ensemble d'éléments terminaux rassemblés dans un noeud. La classification produit une hiérarchie indicée de classes partiellement emboîtées les unes dans les autres.

On détermine a priori le nombre des classes dans lesquelles on désire répartir l'ensemble des éléments à classer, ou le nombre de noeuds retenus pour la classification. En fixant ce nombre de classes à la moitié du nombre total d'individus de deux corpus dans le tableau on tente de se rapprocher au maximum d'un résultat souhaitable : produire des petites classes de deux éléments dont les profils sont très similaires (voir identiques) dans les corpus anglais et français.

3.3. Influence de la variation du découpage

Dans le cadre de notre expérimentation, nous avons effectué une série de partition du corpus afin d'obtenir des parties de taille équivalente. On a obtenu des fragments de texte consécutifs n'ayant pas d'intersection. On peut constater que la variation du découpage influe de manière importante sur la qualité des résultats (cf. figure 3). L'augmentation du nombre de parties permet de préciser les profils des individus à classer. En conséquence, les individus agrégés dans les mêmes classes sont plus proches entre eux. Plus la partition est fine, plus les résultats sont fiables.

1. CLASSE 2526 <i>replies to</i> <i>réponses à</i>	3. CLASSE 2965 <i>whereas the commission accepted it</i> <i>whereas the commission</i> <i>tandis que la commission</i> <i>tandis que la commission y souscrit</i>
2. CLASSE 2026 <i>official gazette</i> <i>journal officiel</i>	4. CLASSE 2273 <i>trente jours</i> <i>thirty days</i>

-
1. /heard addresses by,, mr jäckel, and, and **replies to** a question put by it. partic/
/ir basil hall, lord lester and, and also **replies to** questions put by one of its /
/ions,, me jäckel, et, ainsi qu' en leurs **réponses à** sa question. les circonstanc/
/asil hall, lord lester et, ainsi que des **réponses à** des questions posées par un /
 2. /rative procedure, bgbl[federal **official gazette**] no. 172/ 1950, subject to revi/
/rative procedure, bgbl[federal **official gazette**] no. 172/ 1950, subject to revi
/lois de procédure administrative, bgbl.[**journal officiel** fédéral], concernant l/
/lois de procédure administrative, bgbl.[**journal officiel** fédéral], concernant l/
 3. /ion. the government contested this view, **whereas the commission accepted it.** the/
/ion. le gouvernement combat cette thèse, **tandis que la commission y souscrit** en /
 4. /at the vendor exercised his right within **thirty days** of delivery to the purchase/
/e le vendeur exerçât ses droits dans les **trente jours** de la livraison à l' achet/
-

Figure 8. *Retours aux contexte*

Références

- Bourigault D., Chodkiewicz C. and Humbley J. (1999). Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné. *Actes de la 3ème conférence Terminologie et Intelligence Artificielle (TIA'99)*.
- Harris B. (1988). Bi-Text, a New Concept in Translation Theory. *Language Monthly*, vol.(54): 8-10.
- Isabelle P. (1992). La bi-textualité: vers une nouvelle génération d' aides à la traduction et la terminologie. *META*, vol.(37), no 4: 721-737.
- Isabelle P. and Warwick-Armstrong S. (1993). Les corpus bilingues : une nouvelle ressource pour le traducteur. In Bouillon, P. and Clas, A., editors, *La Traductive*. Montréal : Les Presses de l' Université de Montréal, 288-306.
- Juan J. (1982). Classification ascendante hiérarchique selon les voisins réciproques. *Cahiers de l' analyse des données*,vol.(7), no 2.
- Lamalle C., Martinez W. and Salem A. (1998). Lexico2 : outils de statistique textuelle. Université de la Sorbonne nouvelle - Paris 3. LEXICO (SYLED).
- Lebart L. and Salem A. (1994). *Statistique textuelle*. Paris: Dunod.