Zimina Maria
Université de la Sorbonne Nouvelle - Paris 3
Equipe LEXICO (SYLED)

Alignement de textes bilingues par classification ascendante hiérarchique

Aligner deux textes consiste à mettre en relation des unités textuelles qui se correspondent :

1. The Declaration was adopted at once.

2. All the states pronounced unanimously for the abolition of the former treaty as it was judged out of date.

1. La déclaration fut adoptée sur le champ.

2. Tous les Etats se prononcèrent de manière unanime pour le retrait du Traité.

3. Il fut jugé qu'il ne répondait plus aux conjonctures actuelles.

Les unités en correspondance se positionnent à plusieurs niveaux : *paragraphes*, *phrases*, *syntagmes*, *mots*.

Domaines d'application : traduction, terminologie, lexicographie, industries de la langue, publication etc...

- L'alignement permet de réutiliser les ressources de traductions existantes.
- Il facilite la création des lexiques multilangues.
- Les techniques d'alignement aident à la maintenance de documentation technique dans des langues différentes.

Des méthodes analogues sont également utilisées dans l'appariement d'un texte et de sa traduction phonétique ou d'un texte et de son enregistrement.

L ’alignement permet d ’élaborer les *concordances multilangues*.

RECHERCHER ⇒ countervail

One of the strongest weapons the U.S. has been using to restrict entry of foreign goods onto its market has been **countervail duties**	Les **droits compensateurs** constituent l’une des principales armes que les Etats-Unis ont utilisée pour restreindre les importations de produits étrangers sur leur marché.

Un *bi-texte* est un ensemble de deux textes et des liens qui relient les unités en correspondance.

- Pour générer automatiquement un *bi-texte* on a recours aux *algorithmes d'alignement*.
- La plupart des algorithmes font appel aux méthodes statistiques.
- Au stade actuel, ces algorithmes ne sont pas encore capables de rendre explicites toutes les correspondances de traduction.

Méthodes d'alignement par *classification ascendante hiérarchique* de formes graphiques et segments répétés.

Corpus :

Convention de sauvegarde des Droits de l'homme et des libertés fondamentales

- Corpus anglais
  273 685 occurrences
- Corpus français
  285 961 occurrences

Appliquée au *tableau lexical* qui contient des formes et segments répétés de deux textes, la classification produit des regroupements d'éléments avec des *profils de répartitions* similaires.

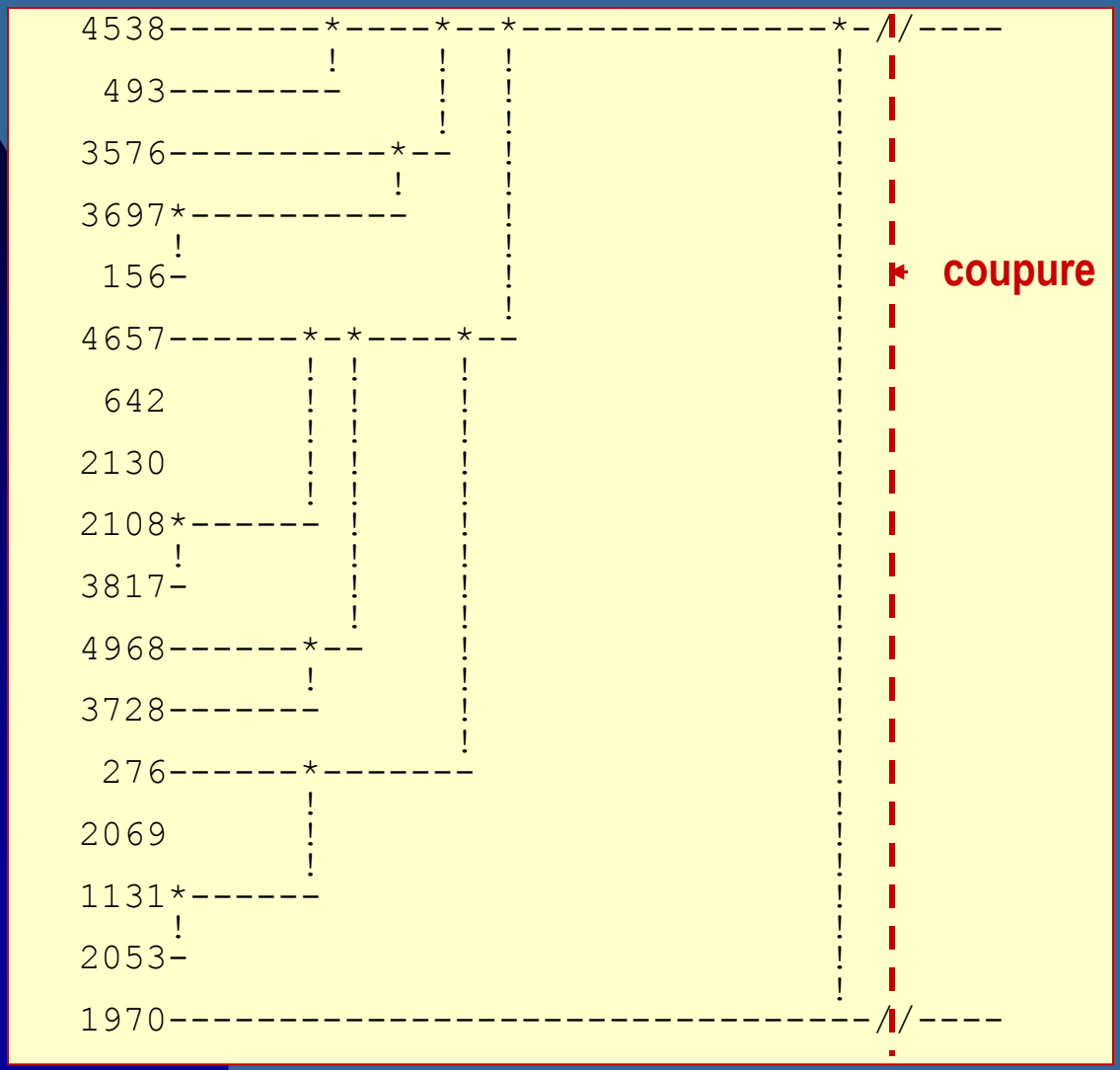
parties

fréquences

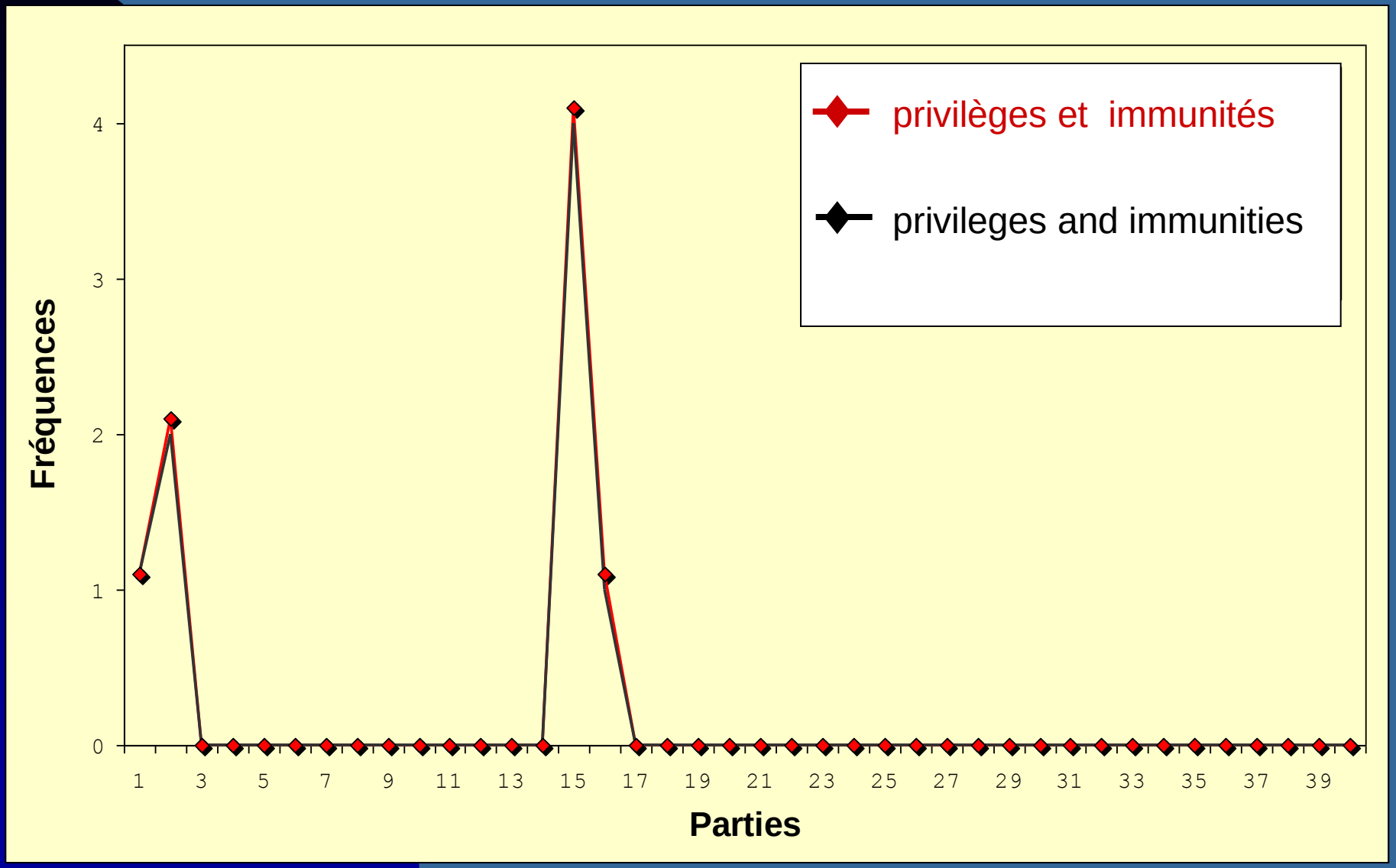
formes et segments répétés

	P=1	P=2	P=3	P=4	P=5	P=6	P=7	P=8	P=9	P=10	(...)
(...)											
taxe	0	2	0	0	0	1	0	0	1	0	(...)
tend	1	1	0	0	0	0	0	1	1	2	(...)
tentative	0	0	0	0	0	0	0	0	0	0	(...)
textes	2	0	0	0	1	0	1	1	0	0	(...)
théorie	0	0	2	1	0	1	0	0	0	0	(...)
tira	0	0	0	0	0	0	0	0	0	0	(...)
tirées	0	1	0	1	0	0	0	0	0	0	(...)
tolstoy	0	0	0	0	0	0	0	0	0	0	(...)
totalité	1	1	1	0	0	0	0	0	0	1	(...)
traitée	0	2	0	0	0	0	1	1	1	0	(...)
(...)											

La *coupure* du dendrogramme permet de considérer une classification résumée aux classes inférieures de la hiérarchie qui regroupent des éléments extrêmement proches.



Profils graphiques des segments répétés agrégés dans les mêmes classes.



La classification a dégagé un nombre important de classes qui regroupent des *individus - correspondances de traduction*.

CLASSE 2692

égalité des

equality of

CLASSE 2526

replies to

réponses à

CLASSE 2267

speculate as to

spéculer sur

CLASSE 2422

37 above

37 ci

CLASSE 2594

differences in

différences entre

CLASSE 2158

lecture des

reading out

■ = *français*

■ = *anglais*

CLASSE 2708

accepts that

admet que

CLASSE 2656

né en

born in

CLASSE 2523

côté de la

côté de

side of

side of the

CLASSE 2371

verdict d

verdict of

CLASSE 2669

contributed to

contribué à

CLASSE 2657

considerations of

considérations d

Filtrage des résultats

- apparier les formes et segments avec des *fréquences générales homogènes* :

■ = ***français***
■ = ***anglais***

```
CLASSE  2622
       248   puni de                F = 11
      1414   punishable by          F = 12
      1402   increased by           F =  5
```

```
CLASSE  2571
       772   résine de cannabis     F = 12
      1618   cannabis resin         F = 12
      1599   schedule 3             F = 15
      1895   ought to               F =  5
```

Filtrage des résultats

```
signature 1 4 1 6 5 0 1 0 0 0 2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0
signature   3 0   5   1   0   2   1   0   0   0   0   0   0   0   0   1   0   0   1   1   0

signature 1 3 1 6 5 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
signature   2 9   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   1

articles  1 5 2 1 1 0 0 1 0 0 0 1 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
articles    3 6   1   1   0   1   1   1   1   1   0   0   0   0   0   0   1   0   1   0   0

articles  1 3 2 1 1 0 0 1 0 0 0 0 1 0 0 1 0 1 2 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0
articles    3 4   1   1   0   1   1   2   1   0   0   1   0   0   0   0   1   0   0   0   1
```

- tenir compte de l'influence du *découpage* du corpus :

PARTIES = 20

ratification
signature
ratification
signature
articles
articles
approval

PARTIES = 40

ratification
ratification
signature
signature

articles
articles

approval

■ = ***français***
■ = ***anglais***

Pistes de développement

- utiliser les classes d'individus agrégés pour aligner les phrases correspondantes
- intégrer la classification aux systèmes d 'alignement des phrases pour *augmenter la résolution* jusqu'au niveau des mots ou des syntagmes

L'intérêt de cette méthode réside dans son degré de *flexibilité* : les correspondances de traduction sont recherchées librement dans l'ensemble de deux textes.

Cette recherche peut être automatisée.
Elle est peu coûteuse en calcul.