# Equivalences traductionnelles

## Bi-texte juridique

## [Equivalences]

Maria Zimina
CLILLAC-ARP, Université Paris Diderot – Paris 7
mzimina@eila.univ-paris-diderot.fr

**Résumé:** ***CONVENTION*** est un bi-texte juridique français/anglais aligné jusqu'au niveau de la phrase à l'aide du logiciel textométrique *mkAlign* (§1). Les *Types* bilingues français/anglais ***administr+/administ+*** sont appariés en raison de leur parenté sémantique dans le corpus. Dans le bi-texte découpé en sections, leurs distributions présentent des divergences. Une suite d'opérations textométriques permet de cerner les causes de ces discordances. On découvre deux phénomènes sensiblement différents : 1) les asymétries sont dues au décalage dans l'alignement des sections ; 2) il existe des contextes originaux où les mots français commençant par ***administr+*** (*administration*, *administrer*, etc.) ne sont pas en équivalence avec des mots anglais commençant par ***administ+*** *(administration*, *administering*, etc.) et réciproquement (§2). On en déduit deux méthodes de travail sur corpus parallèles : 1) une méthode de synchronisation d'alignement phrastique à l'aide de la carte des sections bi-textuelle ; 2) une méthode d'exploration bi-textuelle permettant le repérage de passages originaux où sont attestées des équivalences traductionnelles peu communes (§3).

**Mots-clés :** alignement de corpus, textométrie, traduction

**Abstract:** ***CONVENTION*** is a legal bi-text aligned up to the sentence level with a textometric tool *mkAlign* (§1). The bilingual French/English *Types* ***administr+/administ+*** are put into correspondence following their semantic proximity in the corpus. Their distributions differ within the bi-text divided into sections. A series of textometric operations allow to detect the origins of these variations. Two substantially different phenomena are discovered: 1) Asymmetry of distributions is sometimes due to the section alignment shift; 2) In some contexts, French words starting with ***administered+*** (*administration*, *administration*, etc.) do not correspond to the English words starting with ***administ+*** (*administration*, *administering*, etc.) and vice versa (§ 2). We then suggest two working methods to process parallel text corpora: 1) Sentence alignment synchronization with a bi-text map; 2) Bi-textual exploration to reveal original translation passages with uncommon equivalents (§3).

**Keywords:** Corpus alignment, textometry, translation

La recherche dans le domaine de l'alignement de corpus a montré que le repérage automatique des correspondances est relativement simple dans le cas d'unités textuelles de taille importante, telles que chapitres, paragraphes, sections, etc. En revanche, la recherche des correspondances plus fines est complexe lorsqu'elle implique le découpage des « unités de sens » qui rentrent dans la construction de l'espace sémantique de la traduction au sein des phrases équivalentes (Tiedemann 2011).

Dans cette perspective, les comparaisons isolées des unités textuelles recensées dans les différents volets d'un corpus sont généralement insuffisantes pour explorer les correspondances lexico-grammaticales dont la structure est complexe. Sur le plan fréquentiel, cette complexité se manifeste par des écarts des fréquences générales d'unités de décompte qui ne constituent des équivalences traductionnelles que dans certains contextes.

Au-delà des extractions automatiques d'appariements présentés sous forme de listes, la textométrie a permis de développer une série de méthodes de navigation en corpus multilingues, modulables en fonction de besoins particuliers. Dans le contexte de cette approche, l'expertise humaine de textes

est appuyée par de nombreux outils de lecture et de visualisation qui offrent plusieurs possibilités d'investigation de l'espace intertextuel (Fleury et Zimina, 2008).

Notre travail tentera de montrer que l'utilisation de la « cartographie de correspondances », fondée sur la représentation topographique de corpus multilingues se prête particulièrement bien à l'exploration du tissu d'équivalences traductionnelles. Une étude ciblée d'un bi-texte juridique français-anglais que nous présentons ici a pour but d'illustrer cette démarche méthodologique.

## 1 Bi-texte juridique *CONVENTION*

Le corpus ***CONVENTION*** mobilisé pour cette étude est constitué de textes juridiques français/anglais de la *Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales*, de ses protocoles intégraux, et d'une série d'arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995. Deux versions de chaque document existent parallèlement ; il est difficile de distinguer une langue source et une langue cible. Ce corpus a été réuni dans le cadre d'une étude plus large qui avait pour objectif la construction d'un lexique bilingue des droits de l'homme à base de corpus parallèles alignés (Bourigault *et al.*, 1999). Au cours du projet, le corpus a été aligné semi automatiquement jusqu'au niveau du paragraphe. On estime le taux de précision du découpage en phrases à 90 % environ.

Le corpus compte 12 913 formes pour 296 396 occurrences dans le volet français et 9 530 formes pour 284 958 occurrences dans le volet anglais. La partition naturelle du corpus en 3 parties dont chacune correspond à un ensemble de documents juridiques d'un certain type amène les résultats que l'on peut voir au tableau 1. Les arrêts de la Cour européenne constituent la principale partie du corpus.

**Tableau 1 :**

Structure du corpus ***CONVENTION***

| Corpus *Convention* | volet français<br>296 396 occ. | volet anglais<br>284 958 occ. |
|---|---|---|
| *Convention européenne des Droits de l'Homme* | 5 953 occ. | 5 710 occ. |
| Protocoles intégraux de la Convention | 8 984 occ. | 8 773 occ. |
| Arrêts de la Cour Européenne des Droits de l'Homme | 281 459 occ. | 274 475 occ. |

La figure 2 montre un extrait du texte des arrêts en français et en anglais de ce même corpus après l'appariement des phrases réalisé à l'aide du logiciel textométrique *mkAlign* (Fleury et Zimina, 2005, 2012). Les fonctionnalités de cet outil permettent de construire ou de corriger un alignement de deux textes. Cet alignement en cours peut ensuite être modifié via un éditeur à double entrée intégré dans l'interface du logiciel, puis exporté au format TMX[1].

[1] TMX (Translation Memory eXchange) est une norme d'échange de TM (Translation Memory, en français : Mémoire de Traduction) compatible avec un très grand nombre de logiciels d'aide à la traduction.

Dans notre exemple, le caractère **§** sert de délimiteur de sections appariées sous *mkAlign*. Le texte a été allégé des mises en forme (gras, italiques, etc.) et muni de balises qui permettent d'associer à chaque section alignée :

– la clé facultative <texte> qui distingue deux langues (français : "fr" et anglais : "en") ;

– le caractère § qui matérialise l'alignement des phrases ;

– le caractère * qui permet d'identifier des lettres (à l'origine) en majuscules.

| MKA-ed | SOURCE | CIBLE |
|---|---|---|
| | | |
| 1 | <texte="fr"> du cote gibraltarien de la frontiere, les fonctionnaires des douanes et de la police en service normal ne furent ni informes ni associes a la surveillance, au motif que cela impliquerait que l'information soit communiquee a un trop grand nombre de personnes. § | <texte="en"> on the *gibraltar side of the border, the customs officers and police normally on duty were not informed or involved in the surveillance on the basis that this would involve information being provided to an excessive number of people. § |
| 2 | <texte="fr"> aucune mesure ne fut prise pour ralentir la file de voitures lors de leur entree, ou pour examiner tous les passeports, car on craignait que cela puisse alerter les suspects. § | <texte="en"> no steps were taken to slow down the line of cars as they entered or to scrutinise all passports since it was felt that this might put the suspects on guard. § |
| 3 | <texte="fr"> une equipe de surveillance distincte se trouvait cependant a la frontiere et un groupe prepose a l'arrestation etait poste dans le secteur de l'aeroport voisin. § | <texte="en"> there was, however, a separate surveillance team at the border and, in the area of the airfield nearby, an arrest group. § |
| 4 | <texte="fr"> le temoin *m, qui dirigeait une equipe de surveillance postee a la frontiere, exprima sa deception au vu du manque apparent de cooperation entre les divers groupes impliques a *gibraltar, mais il comprit que les choses etaient ainsi organisees pour des questions de securite. | <texte="en"> witness *m who led a surveillance team at the frontier expressed disappointment at the apparent lack of co-operation between the various groups involved in *gibraltar but he understood that matters were arranged that way as a matter of security. |

**Figure 2 :**

***Convention*** *:* Arrêts de la Cour européenne des Droits de l'Homme (bi-texte aligné sous *mkAlign*)

## 2 Asymétries distributionnelles des *Types* bilingues appariés

A l'issue des premiers dépouillements textométriques, la confrontation des dictionnaires de formes graphiques constitués à partir de chacun des volets du corpus nous amène à nous interroger sur les particularités d'un ensemble de vocabulaire associé dans les deux langues du bi-texte à la notion d'*administration* (en anglais : *administration*). Nous allons constituer un *Type* particulier, que nous appellerons ***administr+*** à partir de toutes les formes graphiques commençant par cette chaîne de caractères dans le volet français du corpus. Puis, de la même façon, nous allons construire un deuxième type à partir de toutes les formes graphiques commençant par la chaîne ***administ+*** dans le volet anglais du corpus.
*A priori*, on peut s'attendre à ce que ces entités soient liées sur le plan de la traduction.

**Tableau 3 :**

*Types* sélectionnés pour une exploration parallèle dans les deux langues

| volet français<br>***adminustr**+* [478 occ.] | volet anglais<br>***adminust**+* [482 occ.] |
|---|---|
| administrative [192 occ.] | administrative [441occ.] |
| administration [103 occ.] | administration [32 occ.] |
| administratif [90 occ.] | administered [4 occ.] |
| administratives [58 occ.] | administer [2 occ.] |
| administratifs [21 occ.] | administering [1 occ.] |
| administrateur [6 occ.] | administrations [1 occ.] |
| administre [2 occ.] | administrator [1 occ.] |
| administrer [2 occ.] | |
| administrant [1 occ.] | |
| administrateurs [1 occ.] | |
| administrations [1 occ.] | |
| administrée [1 occ.] | |

Le tableau 3 montre chacun des *Types* ***administr+*** [478 occ.] et ***administ+*** [482 occ.] (français/anglais) constitué par l'ensemble d'occurrences des formes graphiques regroupées en raison de leur parenté sémantique dans le corpus textométrique.[2]

### *2.1 Carte bi-textuelle*

Les fonctionnalités développées au sein des logiciels textométriques (*Lexico3*, *mkAlign*, *Trameur*) sont conçues pour visualiser la ventilation des unités textuelles à l'aide de la « carte des sections ». Cette représentation topographique a pour objectif de monter la position de n'importe quelle unité textuelle située dans une section donnée du corpus, qu'elle soit de la taille d'une partie, d'un paragraphe, d'une phrase ou d'un fragment textuel déterminé en fonction des besoins de l'étude (Lamalle *et al.*, 2001 ; Fleury et Zimina 2005, 2012 ; Fleury 2007, 2012).

L'utilisateur dispose d'un ensemble d'outils permettant de choisir à partir d'un dictionnaire des *formes*, *lemmes*, *catégories*, *segments* (ou d'un groupes d'*unités lexico-grammaticales*)[3] sélectionnées à l'aide des *expressions régulières*, etc., un *Type* d'unité textuelle sur lequel portera son exploration (Lamalle et Salem, 2002). Après avoir sélectionné le *Type*, il est possible d'étudier sa ventilation dans le corpus divisé en *sections* (phrases, paragraphes, fragments textuels). L'activation de l'unité choisie déclenche sa recherche et l'affichage des résultats de cette recherche sur la carte. La ventilation de l'unité est alors représentée sur la carte par des carrés « coloriés ». Le fragment textuel lié à la section activée sur la carte s'affiche dans la fenêtre du bas de l'éditeur.

---

[2] Dans la famille des logiciels textométriques (*Lexico3*, *mkAlign*, *Trameur*) le langage des « expressions régulières » permet à l'utilisateur de constituer des groupes d'unités textuelles correspondant au *Type* de son choix et d'enregistrer la liste de ces unités pour une exploration ultérieure.

[3] Actuellement, les fonctionnalités du logiciel textométirique *Trameur* se prêtent particulièrement bien à l'analyse lexico-grammaticale grâce à l'intégration explicite d'un système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation et la gestion des annotations multiples sur les unités du texte.

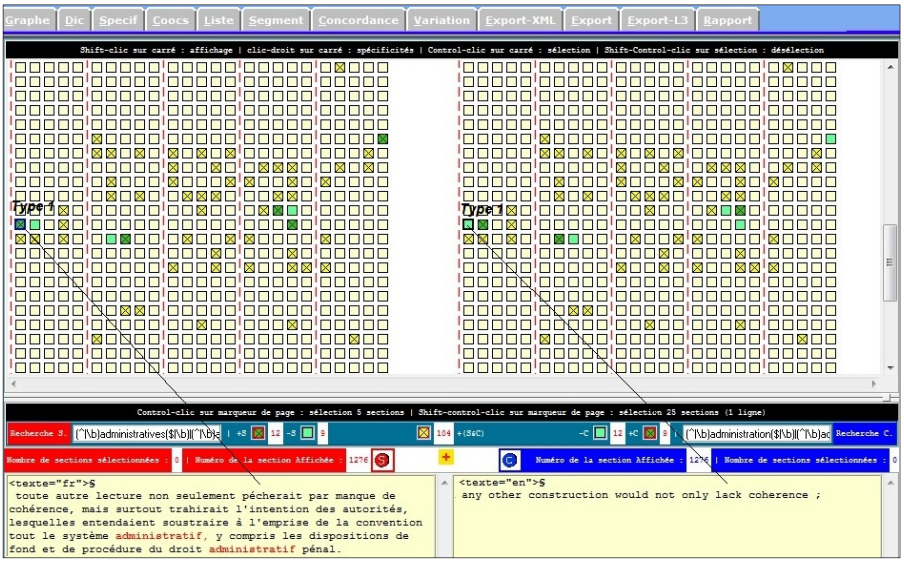



**Figure 4 :**

***Types*** français/anglais ***administr+ / administ+*** ventilés dans le corpus aligné au niveau de la phrase

Afin de poursuivre notre exploration, nous allons créer une carte bi-textuelle en s'appuyant sur l'alignement des phrases réalisé sous *mkAlign*. Sur la figure 4, l'alignement des sections (phrases) du bi-texte juridique est matérialisé par des carrés formant une carte de sections parallèles. Le coloriage des carrés de la carte indique la présence des *Types* étudiés dans les sections concernées du bi-texte juridique. Ainsi, dans notre exemple, les carrés cochés de couleur jaune ☒ signalent les sections bi-textuelles où les mots français commençant par la chaîne ***administr+*** (*administration*, *administrer,* etc.) sont traduits par des mots anglais commençant par la chaîne ***administ+*** (*administration*, *administering*, etc.). Les carrés de couleur vert clair (sans coche ☐) et vert foncé (avec coche ☒) correspondent aux sections du bi-texte où le *Type* français ***administr+*** et le type anglais ***administ+*** ne se correspondent pas. En cliquant sur l'un de ces carrés, le texte correspondant à la section où les deux types ne sont pas liés s'affiche en bas de la carte.

## *2.2 Constats*

La figure 4 montre la ventilation des types ***administr+*** / ***administ+*** dans les sections appariées du corpus. Une conclusion s'impose : dans le corpus ***CONVENTION***, même si l'on peut constater des similitudes importantes qui concernent des sections équivalentes, les distributions de ces *Types* présentent des divergences. Ce constat amène une question : quelles sont les particularités des contextes où les mots français commençant par la chaîne ***administr+*** ne sont pas en correspondance avec des mots anglais commençant par la chaîne ***administ+*** et vice versa ?

La réponse à cette question peut être recherchée dans deux directions distinctes (sans que l'on puisse exclure, *a priori*, que le phénomène soit dû à une combinaison de ces deux possibilités) :

– ***Type 1*** : il existe des décalages dans l'alignement des sections parallèles du corpus, ce qui expliquerait la présence de sections bi-textuelles où les *Types* ***administr+*** et ***administ+*** ne sont pas en correspondance.

– ***Type 2*** : le *Type* ***administr+*** n'est pas toujours en équivalence avec ***administ+*** et il existe des contextes originaux, où sont attestées des équivalences traductionnelles de nature différente, susceptibles d'intéresser le chercheur.

## 3 Résolution du problème par l'exploration textométrique successive

Les jeux de couleurs utilisées pour représenter les distributions des *Types* ***administr+*** et ***administ+***, combinés avec les accès contextuels, permettent de trier entre les cas ***Type 1*** [décalages dans l'alignement des sections] et ***Type 2*** [équivalences traductionnelles peu communes].

### *3.1 Type 1*

Comme indiqué sur la figure 5, le jaune est utilisé pour matérialiser la distribution symétrique. En cliquant sur un carré jaune ⊠ il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types sont liés. Le vert est utilisé pour matérialiser l'asymétrie distributionnelle. En cliquant sur l'un des carrés verts □ ou ⊠, il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types ne sont pas liés :

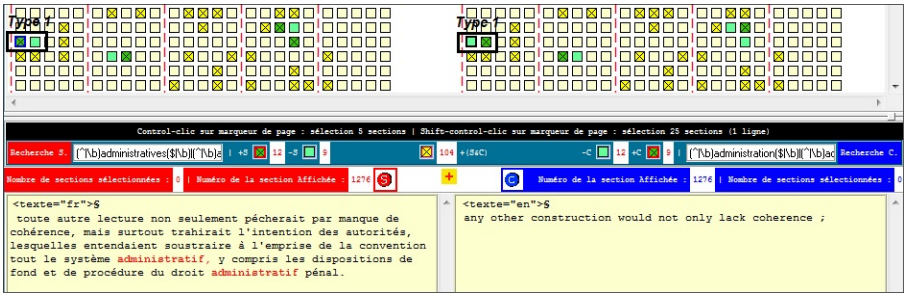
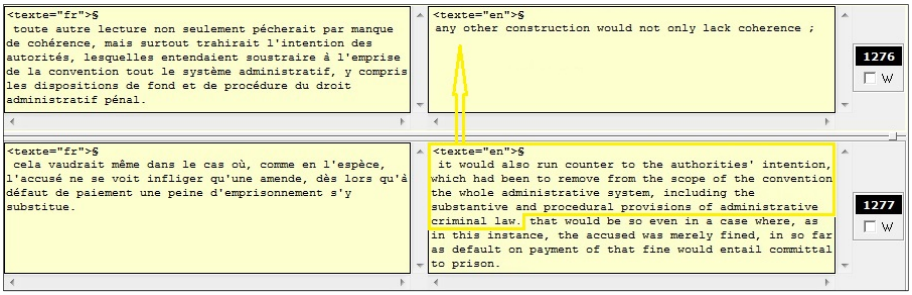



**Figure 5 :**

***Type 1*** [décalages dans l'alignement des sections]

Lorsque deux sections coloriées en vert clair et en vert foncé se succèdent sur les deux volets de la carte :

□⊠ | ⊠□ (dans l'ordre inversé), on peut généralement constater les décalages dans l'appariement des sections. Dans ce cas, les erreurs de l'alignement initial peuvent être corrigées dans l'éditeur dédié de *mkAlign* (cf. figure 6).[4]




[4] Sous *mkAlign*, l'éditeur d'alignement est disponible dans l'onglet *Align* (modes SPLIT/MERGE).

**Figure 6 :**

Edition d'erreurs de l'alignement initial

### *3.2 Type 2*

La symétrie distributionnelle matérialisée par les carrés jaunes ☒ est interrompue par un éparpillement irrégulier de sections coloriées en vert clair ☐ ou en vert foncé ☒ sur les deux volets de la carte (cf. figure 7) :

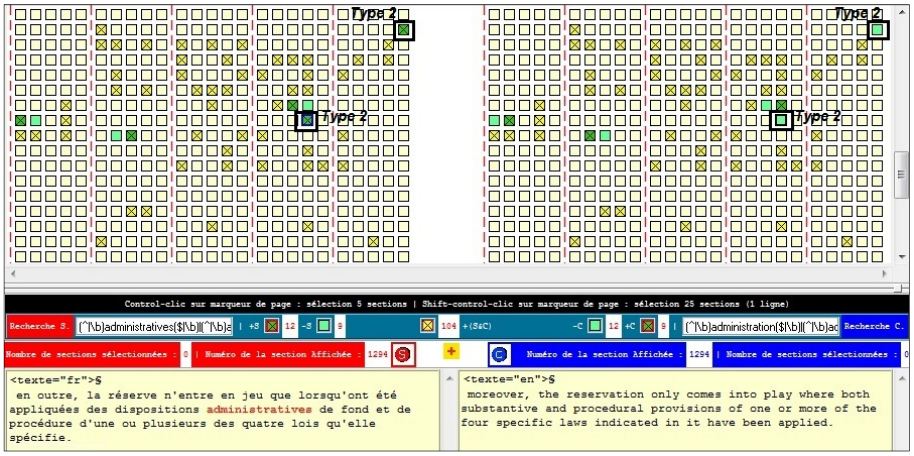



**Figure 7 :**

***Type 2*** [équivalences traductionnelles peu communes]

Ce type de rupture distributionnelle révèle des contextes originaux où les mots français commençant par la chaîne ***administr+*** (*administration*, *administratif*, etc.) ne sont en équivalence avec des mots anglais commençant par la chaîne ***administ+*** (*administration*, *administrative*, etc.) et réciproquement (cf. tableau 8).

La matérialisation de ces sections sur une carte représentant le corpus parallèle permet de dresser une véritable topographie bi-textuelle. Il devient possible d'explorer des contextes singuliers où sont attestées des équivalences lexicales originales, susceptibles d'intéresser l'expert humain pour la construction de ressources textuelles.

**Tableau 8 :**

***Type 2*** [équivalences traductionnelles peu communes]

| *français* | *anglais* |
|---|---|
| le recours administratif | the non-contentious application |
| l'administration des douanes | the customs |
| bonne administration | good governance |
| dépositions administratives | provisions |
| l'administration du district | district authority |
| l'administration des eaux | water-rights authority |
| procédures antérieures | earlier administrative proceedings |

Comme le montre la figure 9, le repérage visuel à l'aide de la carte est renfoncé par la possibilité d'export automatique des contextes correspondants dans un rapport d'exploration textométrique.

| MKA-ed | SOURCE | CIBLE |
|---|---|---|
| | SOURCE Seulement | |
| 133 | <texte="fr">§<br>en outre, la réserve n'entre en jeu que lorsqu'ont été appliquées des dispositions **administratives** de fond et de procédure d'une ou plusieurs des quatre lois qu'elle spécifie. | <texte="en">§<br>moreover, the reservation only comes into play where both substantive and procedural provisions of one or more of the four specific laws indicated in it have been applied. |
| 1150 | <texte="fr">§<br>par une '' décision pénale '' ( straferkenntnis ) du même jour, l'**administration** du district<br>( bezirkshauptmannschaft ) de bregenz le condamna au paiement d'une amende de 10 000 schillings autrichiens ( ats ), assortie d'une peine de 480 heures d'emprisonnement à défaut de paiement, pour non-respect de l' b ) combiné avec l' du code de la route (straßenverkehrsordnung - paragraphes 11 et 12 ci-dessous ). | <texte="en">§<br>in a '' sentence order '' ( straferkenntnis ) of the same day the bregenz district authority<br>( bezirkshauptmannschaft ) ordered him to pay a fine of 10, 000 austrian schillings ( ats ) with 480 hours' imprisonment in default, for an offence under section 99 ( 1 ) ( b ) taken together with section 5 ( 2 ) of the road traffic act ( straßenverkehrsordnung - see paragraphs 11 and 12 below ). |
| 1294 | <texte="fr">§<br>en outre, la réserve n'entre en jeu que lorsqu'ont été appliquées des dispositions **administratives** de fond et de procédure d'une ou plusieurs des quatre lois qu'elle spécifie. | <texte="en">§<br>moreover, the reservation only comes into play where both substantive and procedural provisions of one or more of the four specific laws indicated in it have been applied. |

**Figure 9 :**

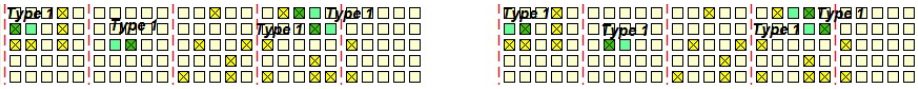
Contextes originaux repérés à l'aide de la topographie bi-textuelle

## 2.4 Méthodes de travail sur corpus parallèles

### *3.2 Synchronisation de l'alignement*

On pose l'équivalence de *Types* bilingues issus de chaque volet du corpus parallèle aligné au niveau du paragraphe ou de la phrase. Le rapprochement des *Types* peut être effectué en prenant en considération leur proximité sémantique ou thématique dans le corpus. On matérialise les distributions des *Types* sur une carte des sections bi-textuelle. Si les distributions sont toujours parallèles mais très légèrement décalées dans certaines parties du corpus, les ruptures du parallélisme signalent le décalage dans l'alignement des sections.

Les paires de ***sections voisines*** □⊠ ou ⊠□ signalent généralement les passages où il existe des erreurs. Voici un diagramme sommaire réalisé à partir d'une telle ventilation :




### *3.2 Repérage de passages originaux dans la traduction*

On matérialise les distributions des types bilingues appariés sur une carte des sections bi-textuelle. Si les distributions se ressemblent, à quelques asymétries près, la ***présence isolée de sections*** □ ou ⊠ montre le plus souvent des passages originaux dans la traduction où sont attestées des équivalences lexicales susceptibles d'intéresser le chercheur. Le diagramme d'une telle ventilation se présente de la façon suivante :



## 4 Conclusion et perspectives

La démarche proposée permet de comprendre les raisons d'asymétries dans les distributions parallèles du vocabulaire bilingue correspondant aux *Types* appariés. En suivante cette approche, l'appariement des *Types* peut aussi être réalisé sur la base d'annotations multiples d'unités textuelles, avec la prise en compte de l'étiquetage morphosyntaxique.[5]

La suite des opérations textométriques convoquées pour localiser les ruptures de parallélisme sur un diagramme représentant le bi-texte aligné constitue une méthode largement applicable à d'autres corpus pluritextuels dans des couples de langues différentes.

A la phase de repérage direct, appuyée sur la topographie bi-textuelle, succède une phase de remise en contexte des particularités distributionnelles constatées. Cette dernière phase débouche sur une édition contrastée des erreurs d'alignement phrastique et de contextes originaux, où sont attestées des équivalences traductionnelles peu communes, difficiles à postuler *a priori*.

## 5 Références

Bourigault D., Chodkiewicz Ch., 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 troisième conférence 'Terminologie et Intelligence Artificielle'*, Nantes, 1999

Fleury S. (2007, mise à jour 2012) « Le Trameur, Manuel d'utilisation ». EA2290 SYLED/CLA2T, http://www.tal.univ-paris3.fr/trameur/leMetierLexicometrique.pdf

Fleury S., Zimina M. (2005, mise à jour 2012) « *mkAlign*, Manuel d'utilisation ». EA2290 SYLED/CLA2T, http://www.tal.univ-paris3.fr/mkAlign/mkAlignDOC.pdf

Fleury S., Zimina M. (2008) « Utilisation de *mkAlign* pour la traduction philologique. » *Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles*, Ecole normale supérieure Lettres et sciences humaines (ENS), Lyon, 2008

Lamalle C., Martinez W., Fleury S, Salem A., Kuncova A., Maisondieu A. (2001) « Dix premiers pas avec *Lexico3*. Manuel d'utilisation abrégé ». EA2290 SYLED/CLA2T, http://www.tal.univ-paris3.fr/lexico/lex3-10pas/Lexico3-10premierspas.pdf

Lamalle C., Salem A. (2002) « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. » *Actes des 6emes journées d'analyse statistique des données textuelles*, Inria, St Malo, 2002

Tiedemann J. (2011) *Bitext Alignment. Synthesis Lectures on Human Language Technologies*. London, Morgan and Claypool Publishers

[5] Le logiciel textométrique *Trameur* est développé dans cette perspective.