

Equivalences traductionnelles

[Equivalences]

Maria Zimina

zimina@msh-paris.fr

Résumé : Les *Types* bilingues français/anglais *administr+ / administ+* sont appariés en raison de leur parenté sémantique dans le corpus parallèle. Dans le bi-texte découpé en sections, leurs distributions respectives 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 la chaîne *administr+* (*administration, administrer* etc.) ne sont pas traduits par des mots anglais commençant par la chaîne *administ+* (*administration, administering* etc.) et réciproquement. 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 lexicales peu communes.

1 Contexte de la recherche

Le corpus *Convention* 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 au niveau de la phrase (Bourigault *et al.*, 1999). Au cours du projet, le corpus *Convention* 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.

Tableau 1
Structure du corpus *Convention*

Corpus <i>Convention</i>	volet français 296 396 occ.	volet anglais 284 958 occ.
<i>Convention européenne des Droits de l'Homme</i>	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.

Les arrêts de la Cour européenne constituent la principale partie du corpus *Convention*. On trouve un extrait du texte des arrêts en français et en anglais au tableau 2 ci-dessous.

Tableau 2
Convention : Arrêts de la Cour européenne des Droits de l'Homme (extraits)

volet français	volet anglais
<p><texte="fr"> § du côté gibraltarien de la frontière, les fonctionnaires des douanes et de la police en service normal ne furent ni informés ni associés à la surveillance, au motif que cela impliquerait que l'information soit communiquée à un trop grand nombre de personnes.</p>	<p><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.</p>
<p><texte="fr"> § aucune mesure ne fut prise pour ralentir la file de voitures lors de leur entrée, ou pour examiner tous les passeports, car on craignait que cela puisse alerter les suspects.</p>	<p><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.</p>
<p><texte="fr"> § une équipe de surveillance distincte se trouvait cependant à la frontière et un groupe préposé à l'arrestation était posté dans le secteur de l'aéroport voisin.</p>	<p><texte="en"> § there was, however, a separate surveillance team at the border and, in the area of the airfield nearby, an arrest group.</p>
<p><texte="fr"> § le témoin *m, qui dirigeait une équipe de surveillance postée à la frontière, exprima sa déception au vu du manque apparent de coopération entre les divers groupes impliqués à *gibraltar, mais il comprit que les choses étaient ainsi organisées pour des questions de sécurité.</p>	<p><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.</p>

Guide de lecture du tableau 2 :

Dans cet extrait du corpus parallèle *Convention*, plusieurs types de codage sont mis en évidence :

- la clé <texte> texte qui distingue deux langues (français : "fr" , anglais : "en") ;
- le caractère § qui matérialise l'alignement des phrases ;
- le caractère * qui permet d'identifier des lettres (à l'origine) en majuscules.

2 Asymétries distributionnelles des *Types* bilingues appariés

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 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.¹ 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

Convention : transformation pour une exploration parallèle sous *Lexico3*

```
§
<texte="fr"> aucune mesure ne fut prise pour ralentir la file de voitures
lors de leur entrée, 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.
§
```

Sur la figure 4, chacun des types *administr+* [478 occ.] et *administ+* [482 occ.] (français/anglais) est constitué par l'ensemble d'occurrences des formes graphiques regroupées en raison de leur parenté sémantique dans le corpus transformé pour une exploration parallèle sous *Lexico3* (voir l'extrait présenté au tableau 3) :²

¹ Sous *Lexico3*, le langage des « expressions régulières » permet à l'utilisateur de constituer des groupes de mots correspondant au type de son choix et d'enregistrer la liste de ces unités pour une exploration ultérieure.

² Dans l'état actuel, les fonctionnalités de *Lexico3* ne permettent pas encore de charger séparément les dictionnaires de formes correspondant à chaque volet d'un corpus bi-textuel. Pour contourner cette difficulté, nous avons différencié les deux langues en introduisant le caractère « _ » (*underscore*) devant chaque forme graphique du volet anglais. Automatisée par une opération Rechercher/Remplacer, l'insertion de cette marque a permis d'éviter toute confusion entre les vocabulaires correspondant à chaque volet du corpus.

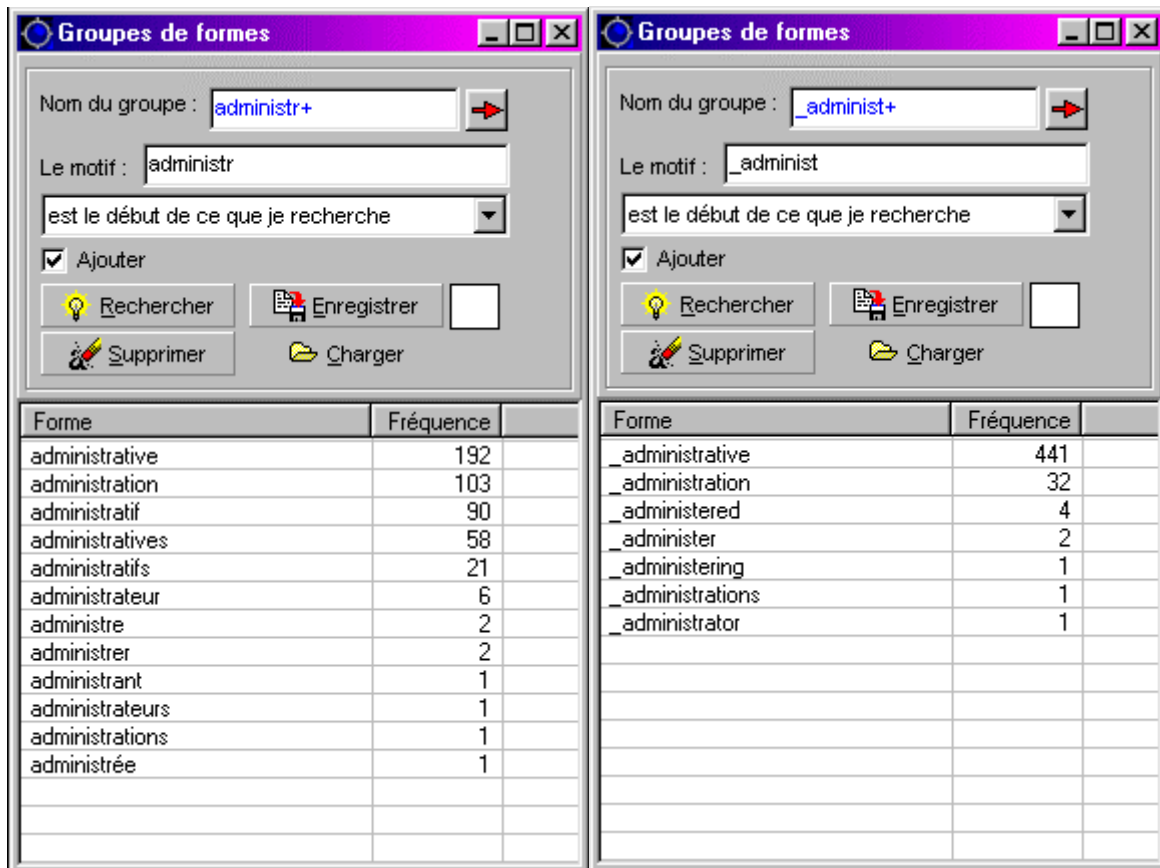


Figure 4
Sélection des *Types* bilingues pour une exploration parallèle

Afin de poursuivre notre exploration, nous allons créer une carte bi-textuelle en s'appuyant sur l'alignement des sections parallèles.³

³ La mise en correspondance des parties équivalentes du corpus parallèle a été réalisée l'aide du logiciel *mkAlign* qui permet de construire ou de corriger un alignement de deux textes. L'outil permet de visualiser l'alignement en cours et de le modifier via un éditeur à double entrée (dans notre exemple, le caractère § sert de délimiteur de sections appariées). *mkAlign* donne la possibilité d'exporter l'alignement au format *Lexico3*. Pour plus d'informations sur les fonctionnalités de cet outil, on consultera la documentation à l'adresse suivante :

<http://tal.univ-paris3.fr/mkAlign/mkAlignDOC/mkAlignDOC.htm>

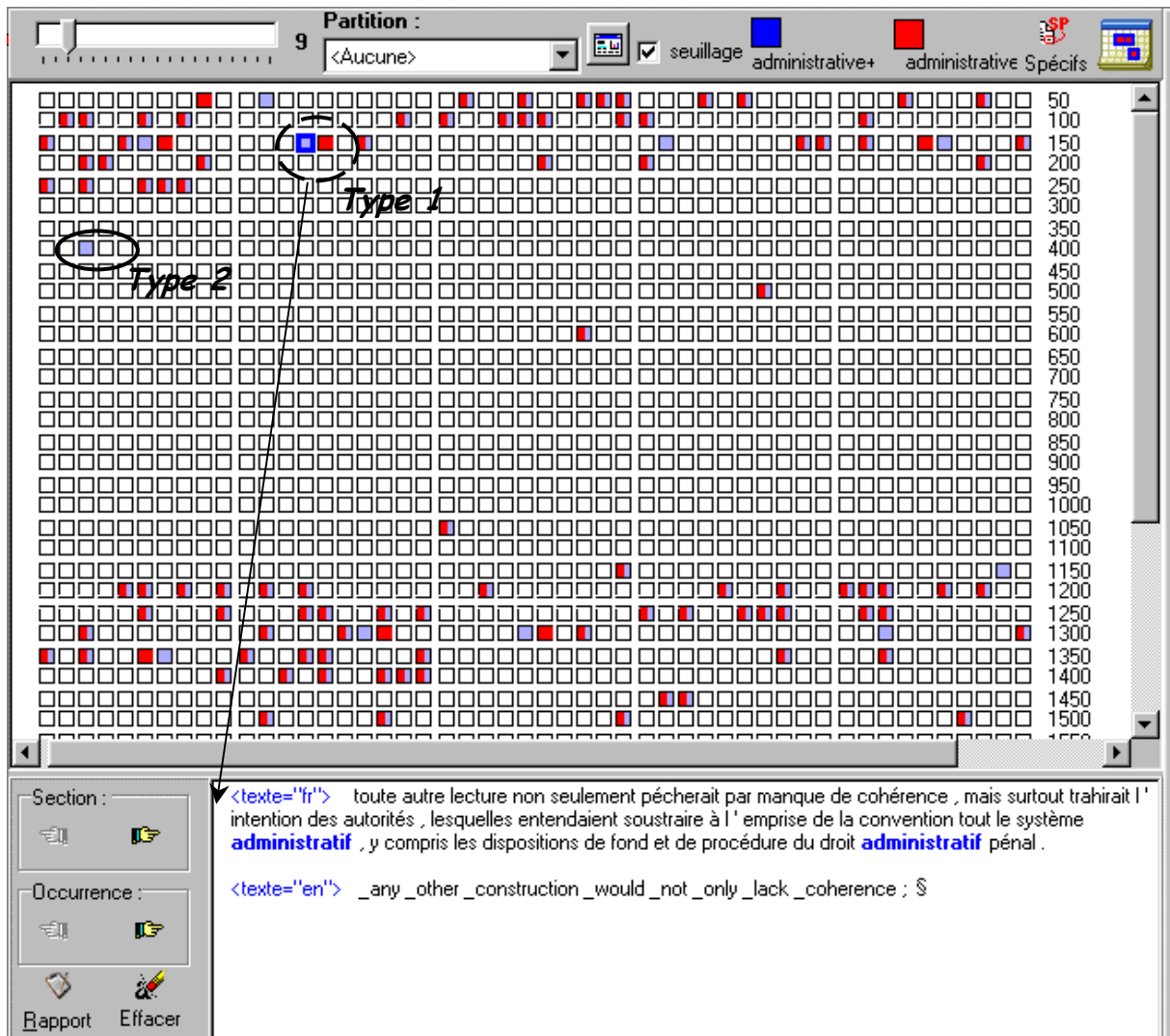


Figure 5

Ventilations des *Types* français/anglais **administr+** / **administ+** dans le corpus aligné au niveau de la phrase : recherche d'asymétries distributionnelles

Guide de lecture de la figure 5 :

L'alignement des *sections* (phrases) du bi-texte est matérialisé par des carrés. Le coloriage des carrés indique la présence des *types* étudiés dans les sections concernées :

■ – les *carrés bicolores* de la carte 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 monochromes* correspondent aux sections du bi-texte où le type français **administr+** et le type anglais **administ+** ne se correspondent pas dans le corpus. En cliquant sur un *carré monochrome* (bleu ou rouge), il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types ne sont pas liés. On peut ensuite étudier les particularités de ces contextes et trier entre les cas qui correspondent aux *décalages dans l'alignement* des sections parallèles du corpus (**Type 1**) et les autres (susceptibles de révéler des équivalences lexicales peu communes – **Type 2**).

Rappel sur les fonctionnalités de la carte des sections bi-textuelle (cf. figure 5) :

Pour étudier la ventilation des *types* sur la carte des sections, on procède de la façon suivante :

On sélectionne le *Tgen* (à partir du dictionnaire, du *Garde-mots*, de la liste des segments répétés, etc.) et on le fait glisser sur la carte (bouton gauche maintenu enfoncé).

On sélectionne la section à visualiser dans la fenêtre du bas en cliquant sur le carré qui la représente dans la carte des sections.

La case *seuillage* permet de régler deux seuils en probabilités qui entraîneront un coloriage (plus ou moins sombre) des sections.

Pour une représentation simultanée de deux *Tgen(s)*, ce processus doit être réitéré (en prenant soin de changer la couleur dans la boîte correspondante). Il faut maintenir la touche Control en position basse lors du second glisser/déposer.

La figure 5 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 parties équivalentes, les distributions des 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+** ?*

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 deux types ne sont pas en correspondance.

Type 2 : le type *administr+* n'est pas toujours traduit par le type *administ+* et il existe des contextes originaux, où sont attestées des équivalences lexicales peu communes, susceptibles d'intéresser le chercheur.

La figure 5 permet de trier entre les cas qui correspondent à la première hypothèse et les autres.

3 Résolution du problème

Les fonctionnalités de la carte des sections rendent possible une visualisation simultanée de la présence/absence des types bilingues. Comme indiqué sur la figure 5, la couleur bleu est utilisée pour matérialiser le type français *administr+* et le rouge pour le type anglais *administ+*. En cliquant sur un *carré bicolore*, il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types sont liés. L'analyse de ces sections signale l'équivalence lexicale des types appariés :

volet français	volet anglais
<p><texte="fr"> les extraits du dossier administratif que cite l'appelant à l'appui de sa thèse ne confortent toutefois pas cette affirmation.</p>	<p><texte="en"> the passages from the administrative file which the appellant cites in evidence in this connection do not, however, support that assertion.</p>


La présence de sections monochromes sur la carte montre qu'il existe des cas de non-correspondance entre les types. En cliquant sur un *carré monochrome* (bleu ou rouge), il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types ne sont pas liés :

volet français	volet anglais
<p><texte="fr"> toute autre lecture non seulement pécherait par manque de cohérence, mais surtout trahirait l'intention des autorités, lesquelles entendaient soustraire à l'emprise de la *convention tout le système administratif, y compris les dispositions de fond et de procédure du droit administratif pénal.</p>	<p><texte="en"> any other construction would not only lack coherence;</p>

Type 1

Lorsque deux sections monochromes coloriées en bleue et rouge se succèdent sur la carte, on peut généralement constater les décalages dans l'appariement des sections. Par exemple :

volet français	volet anglais
<p><texte="fr"> toute autre lecture non seulement pécherait par manque de cohérence, mais surtout trahirait l'intention des autorités, lesquelles entendaient soustraire à l'emprise de la *convention tout le système administratif, y compris les dispositions de fond et de procédure du droit administratif pénal.</p>	<p><texte="en"> any other construction would not only lack coherence;</p>
<p><texte="fr"> cela vaudrait même dans le cas où, comme en l'espèce, l'accusé ne se voit infliger qu'une amende, dès lors qu'à défaut de paiement une peine d'emprisonnement s'y substitue.</p>	<p><texte="en"> it would also run counter to the authorities' intention, which had been to remove from the scope of the *convention the whole administrative system, including the substantive and procedural provisions of administrative criminal law. that would be so even in a case where, as in this instance, the accused was merely fined, in so far as default on payment of that fine would entail committal to prison.</p>

Les erreurs de l’alignement initial peuvent être corrigées si l’on prend soin de sauvegarder les sections concernées dans un rapport. Pour ajouter une section au rapport, il suffit de cliquer sur l’icône *Rapport*  située en bas de la fenêtre de la carte des sections (cf. *Figure 5*).⁴

Type 2

La présence isolée de sections monochromes colorisées en bleu ou en rouge révèle des contextes originaux où les mots français commençant par la chaîne *administr+* (*administration, administratif, etc.*) ne sont pas traduits par des mots anglais commençant par la chaîne *administ+* (*administration, administrative, etc.*) et réciproquement.

La matérialisation de ces sections sur une carte représentant le corpus parallèle permet de dresser une véritable topographie bi-textuelle. Il devient possible d’isoler des contextes singuliers où sont attestées des équivalences lexicales originales, susceptibles d’intéresser l’expert humain pour la construction de ressources textuelle (cf. *Tableau 6*) :

- 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

Tableau 6

Convention : Contextes originaux repérés à l’aide de la topographie bi-textuelle

volet français	volet anglais
<pre><texte="fr"> 1. [le recours administratif] /.../</pre>	<pre><texte="en"> 1. [the non-contentious application] /.../</pre>
<pre><texte="fr"> il prononça la confiscation des marchandises saisies et infligea aux prévenus une amende, assortie de la contrainte par corps, à payer à [l'administration des douanes], partie poursuivante jointe et qui s'était constituée partie civile à l'audience.</pre>	<pre><texte="en"> the court also ordered confiscation of the goods seized and sentenced the defendants to pay a fine, with imprisonment in default, to [the customs], which was a co- prosecutor and had also joined the proceedings as a civil party.</pre>
<pre><texte="fr"> en pareil cas, le tiers peut aussi chercher à démontrer que le directeur a agi en violation d'un principe général de [bonne administration] (algemeen beginsel van behoorlijk bestuur).</pre>	<pre><texte="en"> in so doing, the third party may also base his claim of unlawfulness on the allegation that the *commissioner has acted in breach of a general principle of [good governance] (algemeen beginsel van behoorlijk bestuur).</pre>

⁴ Les erreurs recensées dans l’alignement des sections bi-textuelles peuvent être corrigées à l’aide du programme *mkAlign* (Fleury, 2005).

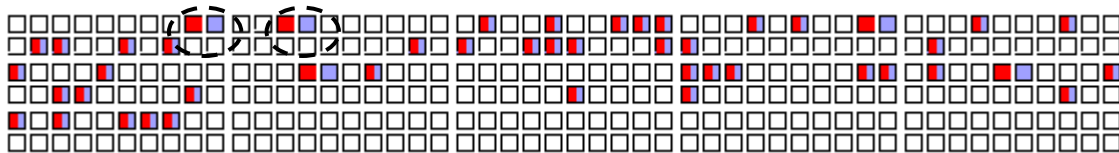
Tableau 6 (suite)

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

volet français	volet anglais
<p><texte="fr"> en outre, la réserve n'entre en jeu que lorsqu'ont été appliquées des <u>dispositions administratives</u> de fond et de procédure d'une ou plusieurs des quatre lois qu'elle spécifie.</p>	<p><texte="en"> moreover, the reservation only comes into play where both substantive and procedural <u>provisions</u> of one or more of the four specific laws indicated in it have been applied.</p>
<p><texte="fr"> il ressort des mémoires soumis par les parties à la procédure devant elle et des dossiers des <u>procédures antérieures</u> qu'une audience ne contribuera sans doute pas à éclaircir l'affaire.</p>	<p><texte="en"> it is apparent to the *court from the pleadings of the parties to the proceedings before it and from the files relating to the <u>earlier administrative proceedings</u> that an oral hearing is not likely to clarify the case further.</p>

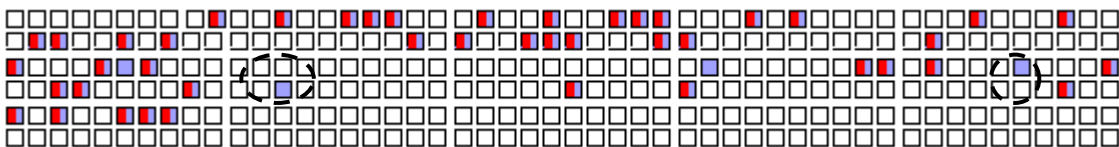
4 Une méthode de synchronisation de l'alignement

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 monochromes* voisines signalent généralement les passages où il existe des erreurs. Voici un diagramme sommaire réalisé à partir d'une telle ventilation :



5 Une méthode de 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 monochromes* 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 :



6 Conclusion

La démarche proposée permet de comprendre les raisons d'asymétries dans les distributions parallèles du vocabulaire bilingue correspondant aux *Types* appariés. 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.

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 lexicales peu communes, difficiles à postuler *a priori*.

7 Références

- Bourigault D., Chodkiewicz Ch., Humbley J. « Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné. », in *actes de la troisième conférence 'Terminologie et Intelligence Artificielle'*, Nantes, 1999.
- Fleury S. « MKAlign », *documentation*. Paris : Université de la Sorbonne nouvelle – Paris 3, (Travaux du SYLED-CLA²T, 2005), <http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm>
- Lamalle C., Salem A., « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in *actes des 6emes journées d'analyse statistique des données textuelles*, Inria, St Malo, 2002.
- Zimina M. « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. », in *actes des 7emes journées d'analyse statistique des données textuelles*, Presses universitaires de Louvain, Louvain-la-neuve, 2004
- Zimina M. *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Thèse de Doctorat en Sciences du langage. Université de la Sorbonne nouvelle – Paris 3, 2004.
- Zimina M. « Exploration textométrique de corpus de traduction », in *actes du colloque « Pour une traductologie proactive » – META '50*, Presses de l'Université de Montréal, Montréal, 2005 (à paraître).

8 Fonctionnalités Lexico3 utilisées dans cette navigation

N°	Fonctionnalité	Résultat
8	Sélection d'un Type (occurrences de formes graphiques commençant par une chaîne de caractères définie)	Figure 4
7	Carte des sections (sections bi-textuelles, présence/absence des Types bilingues français/anglais <i>administr+</i> / <i>administ+</i>)	Figure 5