ZIMINA-POIROT Maria (*SYLED* - ILPGA, Sorbonne nouvelle – Paris 3)
zimina@msh-paris.fr

Thème : Corpus multilingues et alignements textométriques

TP « Analyse de corpus de textes juridiques (français/anglais) à l'aide d'outils de statistique textuelle – ***Lexico3*** ».

Ressources : le corpus ***TRAD_JUR*** (français/anglais) peut être téléchargé à l'adresse suivante :
http://www.cavi.univ-paris3.fr/ilpga/ed/student/stmz/ED268-PagePersoMZ_fichiers/stmz/page2.htm/

Sur cette page, les répertoires compressés (*.zip) comportent trois fichiers :

trad_jur**FR**.txt	(le volet français du corpus)
trad_jur**EN**.txt	(le volet anglais du corpus)
trad_jur**ALL**.txt	(les deux volets bilingues alignés au niveau de la phrase)

Le corpus ***TRAD_JUR*** est composé d'un échantillon de textes juridiques (*français/anglais*) issus des arrêts rendus par la *Cour Européenne des Droits de l'Homme* de Strasbourg. Deux versions de chaque document faisant partie du corpus existent parallèlement ; sans que l'on puisse distinguer une langue source et une langue cible. Pour plus d'informations sur les arrêts, on consultera le site officiel de la Cour Européenne des Droits de l'Homme : http://www.echr.coe.int.

Outils : le logiciel ***Lexico3***

Note : Pour plus d'informations sur le logiciel, on consultera la page d'accueil de ***Lexico*** :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

Les fonctionnalités de ***navigation topographique*** de ***Lexico3*** permettent une exploration parallèle de corpus dans des langues différentes. Ces pratiques peuvent aider l'utilisateur des données bi-textuelles (traducteur, lexicographe, terminologue etc.) dans l'extraction de ressources traductionnelles à base de corpus multilingues.

ZIMINA-POIROT Maria (*SYLED* - ILPGA, Sorbonne nouvelle – Paris 3)
zimina@msh-paris.fr

Quelques exemples d'explorations à faire sous *Lexico3* :

Première partie :

Pour réaliser les explorations décrites dans cette partie, nous utiliserons les fichiers
trad_jur**FR**.txt
trad_jur**EN**.txt

1) Calculer les segments répétés dans les volets français et anglais du corpus ***TRAD_JUR***. Pour cette première exploration, on fixera le seuil de section des segments répétés à 5.

2) Enregistrer la liste des segments du volet français comportant le terme *commission*, puis la liste des segments du volet anglais comportant le terme équivalent *commission*. Comparer les deux listes.

3) Comparer les *inventaires distributionnels* des segments répétés recensés autour des formes équivalentes *commission / commission*.

4) Créer une carte des sections pour chaque volet du corpus (d'après le *caractère-délimiteur* des phrases **§**). Sur chaque carte, afficher la partition selon la clé « droit ». Les deux volets bilingues du corpus ***TRAD_JUR*** sont alors découpés en cinq parties correspondant aux différentes législations citées dans le texte.

5) Comparer les ventilations des segments répétés recensés autour du pôle bilingue *commission / commission* dans les volets français et anglais du corpus.

   Note : L'analyse des profils de ventilation des unités textuelles bilingues dans les corpus parallèles fragmentés en parties permet d'envisager l'alignement automatique des phrases.

Deuxième partie :

Pour réaliser les explorations décrites dans cette partie, nous utiliserons le fichier
trad_jur**ALL**.txt
(le corpus français/anglais ***TRAD_JUR*** aligné au niveau de la phrase)

Note : Pour accéder à des lectures nuancées de l'information bi-textuelle au niveau lexical, on peut utiliser la carte des sections parallèles bilingues (cf. *Figure 1*).

ZIMINA-POIROT Maria (*SYLED* - ILPGA, Sorbonne nouvelle – Paris 3)
zimina@msh-paris.fr

6) Identifier la principale traduction du mot anglais *court* dans le volet français du corpus ***TRAD_JUR***. Quelles sont les autres traductions de ce mot ? Utiliser la carte des sections bi-textuelles pour accéder à ces contextes (cf. *Figure 1*).

7) Faire des explorations similaires pour le mot français *requête* et le mot anglais *applicant*.

8) Rechercher les contextes 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.). Sauvegarder ces sections dans un rapport.

Voici une liste (non exhaustive) de ce genre de correspondances lexicales :

français	**anglais**
l'**administr**ation des douanes bonne **administr**ation dépositions **administr**atives le recours **administr**atif	the customs good governance procedural provisions the non-contentious application

Notons que les résultats de cette exploration soulèvent également un certain nombre de problèmes liés à la détection de simplifications dans la traduction. Voici, par exemple, un couple de phrases en correspondance de traduction tirées du corpus :

français	**anglais**
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.	any other construction would not only lack coherence.

Cet exemple montre que l'on peut envisager l'utilisation de la topographie bi-textuelle pour la vérification de l'*homogénéité traductionnelle*.

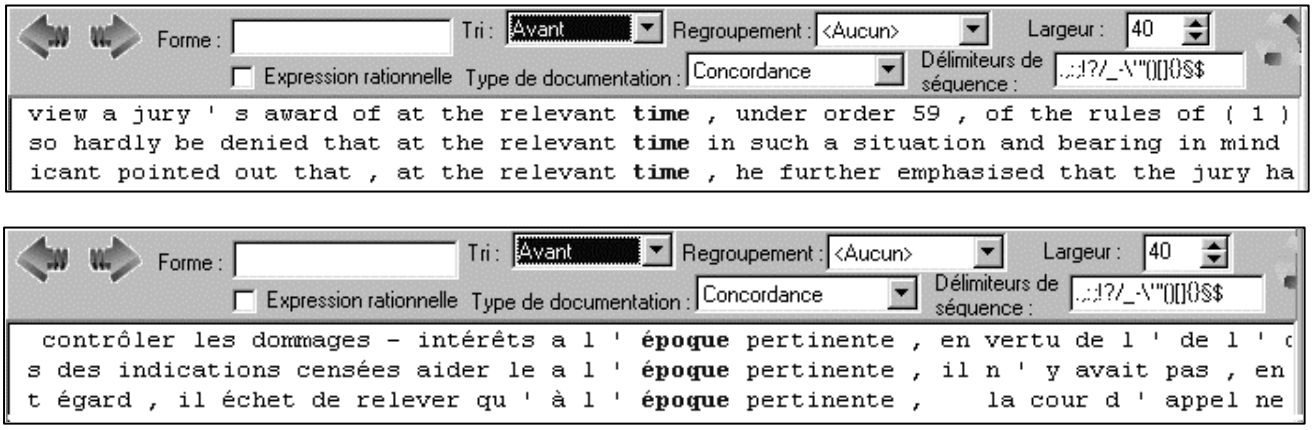
9) Identifier le vocabulaire caractéristique des phrases contenant le pôle *démocrat+/democra+*.

10) Comment est traduit en anglais le segment répété *fonction publique* (F=55) ?

ZIMINA-POIROT Maria (*SYLED* - ILPGA, Sorbonne nouvelle – Paris 3)
zimina@msh-paris.fr

11) Faire des explorations similaires pour le segment répété anglais *constitutional system* et le segment français l'*époque des faits*.

Note : En plus de la *carte des sections*, il est possible d'afficher les *concordances* pour chaque couple d'unités équivalentes :



12) Afficher le graphique correspondant à la ventilation du mot anglais *applicant* dans les cinq parties du corpus. Créer une unité complexe (*type*) correspondant à la liste de ses traductions dans le corpus, par exemple : *requérant(e)*, *intéressé(e)* etc. (cf. question 7 ci-dessus). Sous ***Lexico3***, il est possible d'enregistrer cette liste d'unités lexicales pour une exploration ultérieure.

Comparer le profil de ventilation de ce nouveau type avec celui du mot *applicant*. Enregistrer les résultats de cette comparaison dans le rapport. Les différences constatées dans les ventilations de ces unités signalent les parties du corpus dans lesquelles *applicant* reçoit d'autres traductions que *requérant(e)* et *intéressé(e)*. Lesquelles ? Faire une recherche à l'aide de la *carte des sections* découpée en parties.

ZIMINA-POIROT Maria (*SYLED* - ILPGA, Sorbonne nouvelle – Paris 3)
zimina@msh-paris.fr

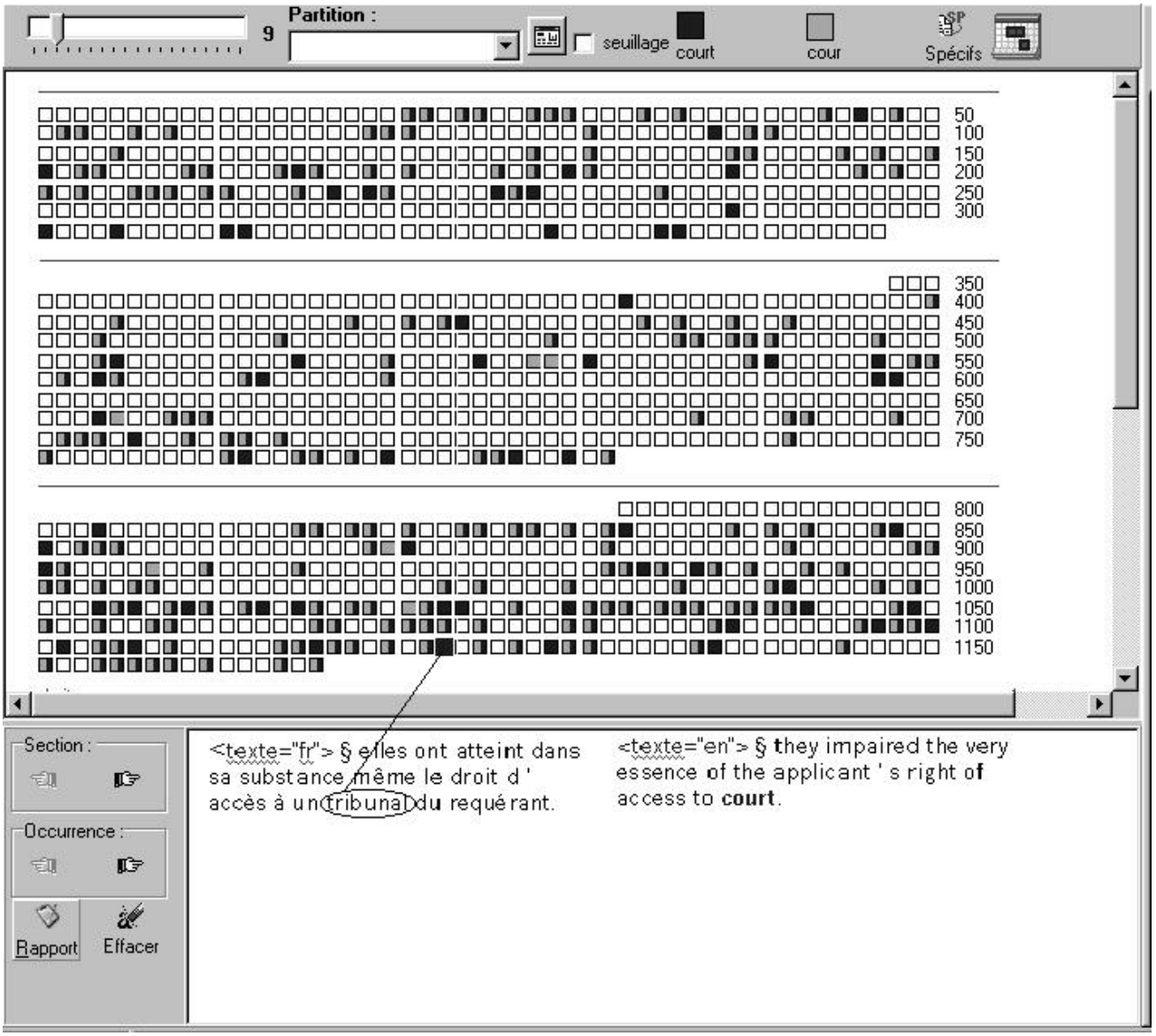


Figure 1 : Le repérage topographique des traductions du terme anglais *court*

Guide de lecture : Les sections bicolores de la carte correspondent aux sections du bi-texte où le terme anglais *court* est traduit par *cour* en français. La présence de sections monochromes sur la carte indique que le mot anglais reçoit d'autres traductions en français. Il s'agit notamment du mot français *tribunal*.

Références

Sur les méthodes de la ***textométrie multilingue***, on consultera le site du groupe ***GADT*** :

http://www.cavi.univ-paris3.fr/lexicometrica/jadt/textometrie-multilingue/