

## Base textométrique de textes alignés

10/03/2015 21:56:54

Serge Fleury

(document de travail)

Ce document présente la mise à jour mise en œuvre à partir de la **version 12.00** du *Trameur* permettant de traiter une base intégrant 2 (ou plusieurs) textes alignés. Cette mise à jour reprend une fonctionnalité déjà disponible dans *mkAlign*<sup>1</sup> permettant notamment de mettre à jour des équivalents traductionnels dans des textes alignés par un calcul de cooccurrence réalisé en parallèle sur 2 volets de l'alignement traité.

Cette mise à jour complète les fonctionnalités déjà disponibles dans le *Trameur* pour le traitement des fichiers multilingue au format TMX (cf documentation du *Trameur*).

### Sommaire

|   |    |
|---|----|
| Base textométrique de textes alignés.....                                     | 1  |
| 1. Base textométrique : 2 volets alignés.....                                 | 2  |
| 1.1 Chargement d'une nouvelle base intégrant 2 textes bilingues alignés ..... | 2  |
| 1.2 Chargement d'une nouvelle base intégrant 2 textes alignés.....            | 7  |
| 2. Base textométrique : alignement au format TMX.....                         | 10 |
| 3. Importer une base « d'alignements » multi-annotée .....                    | 17 |
| 4. Calculs contrastifs sur une base textométrique de textes alignés .....     | 19 |

---

<sup>1</sup> <http://www.tal.univ-paris3.fr/mkAlign/>

## 1. Base textométrique : 2 volets alignés

### 1.1 Chargement d'une nouvelle base intégrant 2 textes bilingues alignés

Pour illustrer cette fonctionnalité, nous commençons par choisir un exemple de texte bilingue à charger par la suite dans le [Trameur](#). Nous travaillerons ici sur un texte de convention européenne disponible en français et en anglais, ces 2 volets étant alignés (ces 2 fichiers sont disponibles dans le programme *mkAlign*).

```
<partie = conv>
convention de sauvegarde des droits de l'homme et des libertés fondamentales
telle qu'amendée par le protocole n° 11
rome, 4.xi.1950
les gouvernements signataires, membres du conseil de l'europe,
considérant la déclaration universelle des droits de l'homme, proclamée par l'assemblée générale des nations unies le 10 décembre 1948;
considérant que cette déclaration tend à assurer la reconnaissance et l'application universelles et effectives des droits qui y sont énoncés;
considérant que le but du conseil de l'europe est de réaliser une union plus étroite entre ses membres, et que l'un des moyens d'atteindre ce but est la sauvegarde et le développement
des droits de l'homme et des libertés fondamentales;
réaffirmant leur profond attachement à ces libertés fondamentales qui constituent les assises mêmes de la justice et de la paix dans le monde et dont le maintien repose essentiellement
sur un régime politique véritablement démocratique, d'une part, et, d'autre part, sur une conception commune et un commun respect des droits de l'homme dont ils se réclament;
résolus, en tant que gouvernements d'etats européens animés d'un même esprit et possédant un patrimoine commun d'idéal et de traditions politiques, de respect de la liberté et de prééminence du droit, à prendre les premières mesures propres à assurer la garantie collective de certains des droits énoncés dans la déclaration universelle,
sont convenus de ce qui suit:
#
article 1 - obligation de respecter les droits de l'homme 1
les hautes parties contractantes reconnaissent à toute personne relevant de leur juridiction les droits et libertés définis au titre 1 de la présente convention:
titre 1 - droits et libertés 1
#
article 2 - droit à la vie 1
le droit de toute personne à la vie est protégé par la loi. la mort ne peut être infligée à quiconque intentionnellement, sauf en exécution d'une sentence capitale prononcée par un tribunal
lunual au cas où le délit est puni de cette peine par la loi.
la mort n'est pas considérée comme infligée en violation de cet article dans les cas où elle résulterait d'un recours à la force rendu absolument nécessaire:
pour assurer la défense de toute personne contre la violence illégale;
pour effectuer une arrestation régulière ou pour empêcher l'évasion d'une personne régulièrement détenue;
pour réprimer, conformément à la loi, une émeute ou une insurrection.
#
article 3 - interdiction de la torture 1
nul ne peut être soumis à la torture ni à des peines ou traitements inhumains ou dégradants.
#
article 4 - interdiction de l'esclavage et du travail forcé 1
nul ne peut être tenu en esclavage ni en servitude.
nul ne peut être astreint à accomplir un travail forcé ou obligatoire.
n'est pas considéré comme "travail forcé ou obligatoire" au sens du présent article:
tout travail requis normalement d'une personne soumise à la détention dans les conditions prévues par l'article 5 de la présente convention, ou durant sa mise en liberté conditionnelle;
tout service de caractère militaire ou, dans le cas d'objecteurs de conscience dans les pays où l'objection de conscience est reconnue comme légitime, à un autre service à la place du
service militaire obligatoire;
tout service requis dans le cas de crises ou de calamités qui menacent la vie ou le bien-être de la communauté;
tout travail ou service formant partie des obligations civiles normales.
#
```

Figure 1 : fichier 1, volet français (extrait)

```
<partie = conv>
convention for the protection of human rights and fundamental freedoms
as amended by protocol no. 11
rome, 4.xi.1950
the governments signatory hereto, being members of the council of europe,
considering the universal declaration of human rights proclaimed by the general assembly of the united nations on 10th december 1948;
considering that this declaration aims at securing the universal and effective recognition and observance of the rights therein declared;
considering that the aim of the council of europe is the achievement of greater unity between its members and that one of the methods by which that aim is to be pursued is the maintenance
and further realisation of human rights and fundamental freedoms;
reaffirming their profound belief in those fundamental freedoms which are the foundation of justice and peace in the world and are best maintained on the one hand by an effective political
democracy and on the other by a common understanding and observance of the human rights upon which they depend;
being resolved, as the governments of european countries which are like-minded and have a common heritage of political traditions, ideals, freedom and the rule of law, to take the first
steps for the collective enforcement of certain of the rights stated in the universal declaration,
have agreed as follows:
#
article 1 - obligation to respect human rights 1
the high contracting parties shall secure to everyone within their jurisdiction the rights and freedoms defined in section 1 of this convention.
section 1 - rights and freedoms 1
#
article 2 - right to life 1
everyone's right to life shall be protected by law. no one shall be deprived of his life intentionally save in the execution of a sentence of a court following his conviction of a crime
for which this penalty is provided by law.
deprivation of life shall not be regarded as inflicted in contravention of this article when it results from the use of force which is no more than absolutely necessary:
in defence of any person from unlawful violence;
in order to effect a lawful arrest or to prevent the escape of a person lawfully detained;
in action lawfully taken for the purpose of quelling a riot or insurrection.
#
article 3 - prohibition of torture 1
no one shall be subjected to torture or to inhuman or degrading treatment or punishment.
#
article 4 - prohibition of slavery and forced labour 1
no one shall be held in slavery or servitude.
no one shall be required to perform forced or compulsory labour.
for the purpose of this article the term "forced or compulsory labour" shall not include:
any work required to be done in the ordinary course of detention imposed according to the provisions of article 5 of this convention or during conditional release from such detention;
any service of a military character or, in case of conscientious objectors in countries where they are recognised, service exacted instead of compulsory military service;
any work or service exacted in case of an emergency or calamity threatening the life or well-being of the community;
any work or service which forms part of normal civic obligations.
#
```

Figure 2 : fichier 1, volet anglais (extrait)

Les 2 textes sont structurés en partie (*via* un système de balisage). L'alignement est matérialisé dans les 2 volets par le caractère # qui induit les alignements respectifs des différents paragraphes.

Dans *mkAlign*, les 2 fichiers sont chargés un à un (le volet FR et le volet EN) dans les 2 volets de l'alignement.

# Le Trameur : documentation complémentaire



Figure 3 : Edition d'un alignement dans mkAlign

Pour le *Trameur*, la base textométrique est construite ici par concaténation des 2 fichiers en un seul fichier, en intégrant en outre un marqueur de partie supplémentaire pour contraster les 2 volets ( <volet="FR"> et <volet="EN"> au début de chacun des 2 volets).

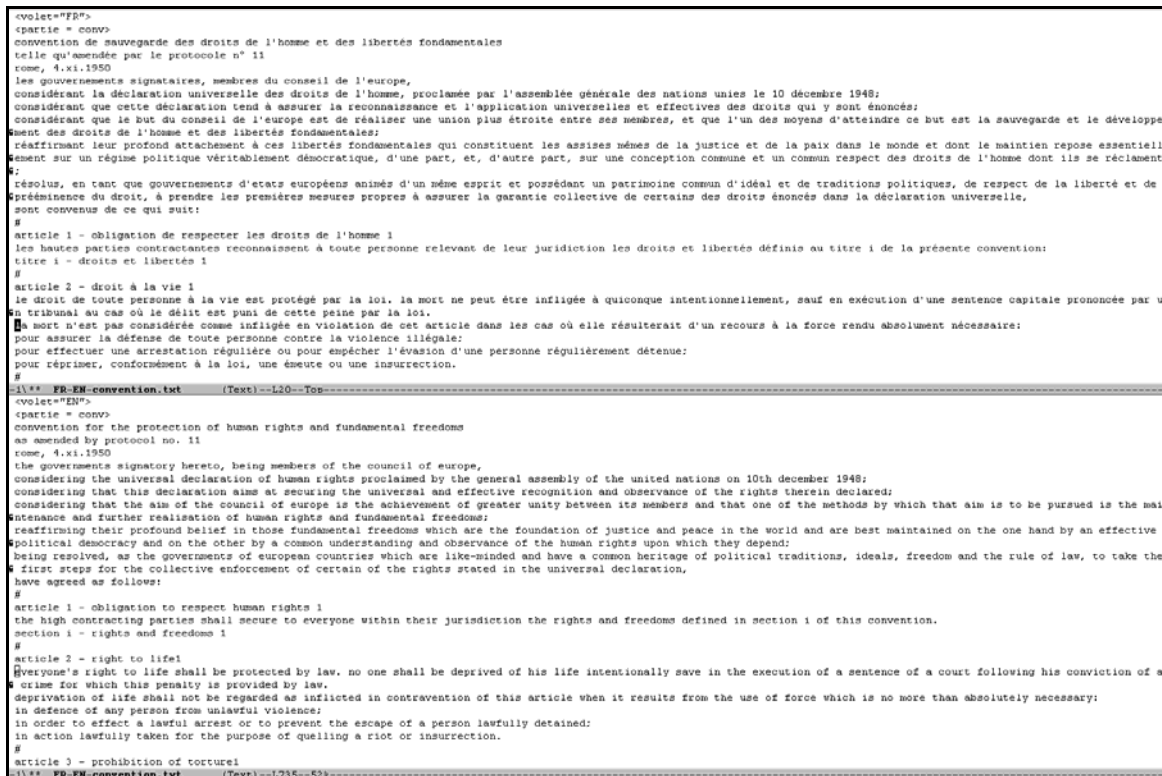


Figure 4 : Préparation d'une base textométrique de textes bilingues alignés

La figure précédente donne à voir le fichier construit en mettant au jour les 2 volets (2 zones différentes du même fichier : le volet FR suivi du volet EN). Une fois chargé dans le *Trameur* (le caractère # ayant été ajouté à la liste des délimiteurs), la *Carte des sections* a l'allure suivante (avec le caractère # comme délimiteur de sections):

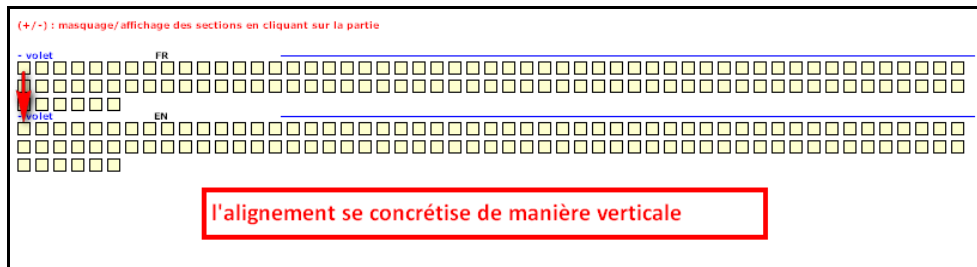


Figure 5 : Carte des sections d'un alignement dans le Trameur

On choisit d'afficher la partition VOLET pour contraster les 2 volets et on aligne à gauche les sections.

Dans *mkAlign*, la *Carte des sections* de l'alignement a l'allure suivante :

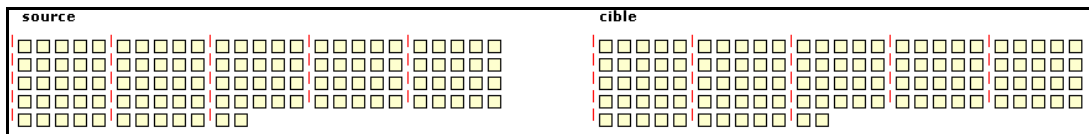


Figure 6 : Carte des sections d'un alignement dans mkAlign

L'alignement se lit de manière horizontale. Contrairement à *mkAlign*, l'alignement produit *via* la *Carte des sections* dans le *Trameur* se lit de manière verticale ; dans la figure 5, le premier carré FR a son "miroir" au « bout de la flèche »...

Si on projette une forme sur la carte, on obtient ceci :

Dans *mkAlign* :

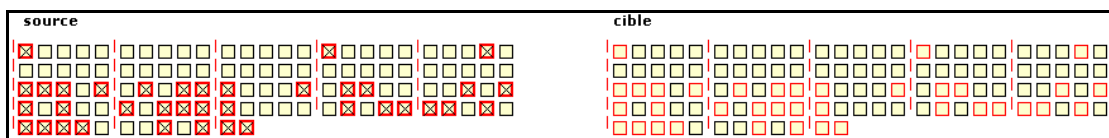


Figure 7 : Projection d'une forme sur carte des sections dans mkAlign

Le mot « conseil » est projeté sur le volet FR (à gauche), la zone miroir est matérialisée dans le volet EN par des carrés cerclés de rouge.

Dans le *Trameur* :



Figure 8 : Projection d'une forme sur la carte des sections dans Le Trameur

Le mot « conseil » est projeté sur la carte, seul le volet FR (en haut) met au jour la présence de la forme visée dans les sections cochées, la zone miroir n'est pas directement matérialisée dans le volet EN.

Dans le *Trameur*, le calcul des spécificités des sections contenant une forme projetée tient compte du vocabulaire de l'ensemble du corpus (ici les 2 volets). Alors que dans *mkAlign*, ce calcul se fait sur le volet FR et en parallèle sur le volet EN (les 2 volets étant dissociés).

La nouvelle fonctionnalité introduite dans le *Trameur* consiste à "dissocier" les 2 volets de l'alignement et à reproduire un calcul en parallèle sur les 2 volets (*i.e* comme dans *mkAlign*) :

- Calcul des cooccurrences sur les sections contenant la forme projetée sur le volet FR
- En parallèle, calcul des cooccurrences sur les sections miroirs des précédentes

Pour cela, on dispose d'un nouveau menu dans l'onglet "Section" :

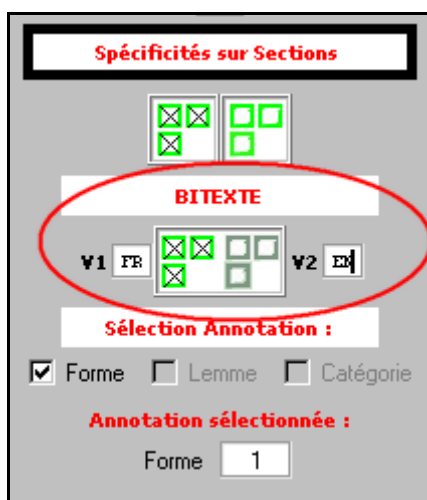


Figure 9 : Paramétrage d'un bitexte

Dans la zone BITEXTE : il est possible de considérer la *Carte des sections* comme une concaténation de 2 volets alignés.

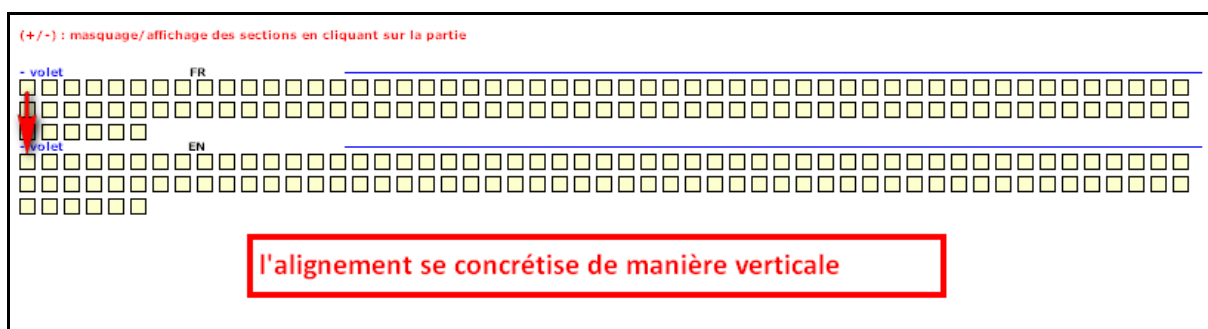


Figure 10 : Mise au jour du bitexte via la carte des sections

C'est le cas ci-dessus, pour la partition *VOLET*, on dispose bien d'une partie *VOLET=FR* et d'une partie *VOLET=EN*, les sections de chacun de ces volets étant alignées verticalement.

Pour amorcer le calcul, on paramètre donc les 2 volets à contraster : le volet V1 est celui contenant la forme projetée (ici FR est renseignée dans la zone de saisie V1) et le volet V2 est le miroir (ici EN dans la zone de saisie V2).

Le bouton du milieu permet de lancer le calcul qui produit ceci : (2 extraits : volet SOURCE et volet CIBLE)

**Spécificités du vocabulaire (Annotation Forme) sur les sections SOURCE contenant le motif : conseil**

Nombre d'occurrences du texte global = 11099  
 Nombre d'occurrences dans la partie visée = 5314  
 Seuil : 10

----- SOURCE -----  
 (cible infra)

(Spécificités positives en haut de liste, négatives en bas)  
 Le fichier construit : C://SERGEF-1/SF-WORK/07-DEVEL/ANR-TE-1/MKGRAP-1/textes/CONVEN-1/FR-EN/export/resultspecif-source-1414344563.txt

| Forme        | Ind-Specif | Eg-Totale | Eg-Partie |
|--------------|------------|-----------|-----------|
| conseil      | 26.4       | 79        | 79        |
| europe       | 26.1       | 78        | 78        |
| du           | 19.6       | 192       | 152       |
| protocole    | 18.1       | 95        | 85        |
| secrétaire   | 16.1       | 47        | 47        |
| déclaration  | 15.0       | 49        | 48        |
| ratification | 13.4       | 44        | 43        |
| le           | 12.9       | 231       | 163       |
| général      | 12.6       | 50        | 47        |
| membres      | 9.7        | 32        | 31        |
| présent      | 8.9        | 83        | 65        |
| vigueur      | 8.5        | 36        | 33        |
| suit         | 8.4        | 23        | 23        |
| acceptation  | 8.0        | 22        | 22        |
| approbation  | 7.7        | 21        | 21        |
| signature    | 7.6        | 25        | 24        |
| mois         | 7.4        | 32        | 29        |
| qui          | 7.3        | 91        | 67        |
| territoires  | 7.1        | 19        | 19        |
| premier      | 7.1        | 19        | 19        |
| entrera      | 6.8        | 18        | 18        |
| n°           | 6.8        | 18        | 18        |
| jour         | 6.4        | 17        | 17        |
| tout         | 6.1        | 67        | 50        |
| notification | 6.1        | 16        | 16        |
| date         | 6.0        | 27        | 24        |

Figure 11 : Spécificités volet SOURCE de l'alignement

----- CIBLE -----

Spécificités du vocabulaire (Annotation Forme) sur les sections CIBLE miroir de celles contenant le motif : conseil

Nombre d'occurrences du texte global = 10453  
 Nombre d'occurrences dans la partie visée = 5047  
 Seuil : 10  
 (Spécificités positives en haut de liste, négatives en bas)

| Forme        | Ind-Specif | Eg-Totale | Eg-Partie |
|--------------|------------|-----------|-----------|
| council      | 25.2       | 76        | 76        |
| europe       | 24.8       | 75        | 75        |
| secretary    | 15.9       | 47        | 47        |
| protocol     | 15.7       | 87        | 77        |
| declaration  | 15.6       | 46        | 46        |
| ratification | 13.3       | 44        | 43        |
| general      | 10.7       | 55        | 49        |
| member       | 9.3        | 31        | 30        |
| day          | 8.6        | 24        | 24        |
| signature    | 8.6        | 24        | 24        |
| into         | 8.4        | 28        | 27        |
| of           | 8.1        | 873       | 496       |
| acceptance   | 8.0        | 22        | 22        |
| approval     | 7.6        | 21        | 21        |
| date         | 7.3        | 32        | 29        |
| at           | 6.7        | 54        | 43        |
| first        | 6.3        | 21        | 20        |
| signed       | 6.1        | 16        | 16        |
| month        | 6.1        | 16        | 16        |
| notification | 6.1        | 16        | 16        |
| force        | 6.1        | 31        | 27        |
| territories  | 6.1        | 16        | 16        |
| being        | 5.7        | 26        | 23        |
| deposit      | 5.7        | 15        | 15        |
| territory    | 5.1        | 33        | 27        |
| enter        | 5.1        | 17        | 16        |
| after        | 4.9        | 26        | 22        |
| expiration   | 4.8        | 12        | 12        |
| signatory    | 4.8        | 12        | 12        |
| instrument   | 4.8        | 16        | 15        |
| deposited    | 4.5        | 15        | 14        |
| fundamental  | 4.4        | 18        | 16        |
| rome         | 4.2        | 10        | 10        |
| instruments  | 4.2        | 10        | 10        |
| receipt      | 4.2        | 10        | 10        |
| terms        | 4.2        | 10        | 10        |
| addressed    | 4.2        | 10        | 10        |
| this         | 4.0        | 118       | 74        |

Figure 12 : Spécificités volet CIBLE de l'alignement

Le calcul a produit le calcul des cooccurrences sur les sections du volet FR contenant la forme projetée (le mot « conseil »), *i.e* le volet SOURCE ; en parallèle ce calcul a aussi été réalisé sur le volet CIBLE : les sections miroir des précédentes dans lesquelles on trouve par exemple l'équivalent traductionnel de « conseil » mis au jour par le calcul. C'est à dire la même chose que *via mkAlign* :

| Spécificités du vocabulaire sur les sections SOURCE contenant le motif : <conseil>  |            |           |         | Spécificités du vocabulaire sur les sections CIBLE associées aux sections SOURCE contenant  |            |           |         |
|---|------------|-----------|---------|---|------------|-----------|---------|
| Nombre d'occurrences du texte global = 11099<br>Nombre d'occurrences dans la partie visée = 5314<br>Seuil : 10<br>(Spécificités positives en haut de liste, négatives en bas)<br>Le fichier construit : C:/SERGEF-1/SF-WORK/07-DEVEL/MKALIGN-1/corpus/CONVEN-1/CONVEN-1/FR/export |            |           |         | Nombre d'occurrences du texte global = 10453<br>Nombre d'occurrences dans la partie visée = 5047<br>Seuil : 10<br>(Spécificités positives en haut de liste, négatives en bas)<br>Le fichier construit : C:/SERGEF-1/SF-WORK/07-DEVEL/MKALIGN-1/corpus/CONVEN-1/CONVEN-1/FR/export |            |           |         |
| Forme   | Ind-Specif | Eq-Totale | Eq-Part | Forme   | Ind-Specif | Eq-Totale | Eq-Part |
| conseil   | 26.4       | 79        | 79      | council   | 25.2       | 76        | 76      |
| europe  | 26.1       | 78        | 78      | europe  | 24.8       | 75        | 75      |
| du  | 19.6       | 192       | 152     | secretary   | 15.9       | 47        | 47      |
| protocole   | 16.1       | 95        | 85      | protocol  | 15.7       | 87        | 77      |
| secrétaire  | 16.1       | 47        | 47      | déclaration   | 15.6       | 46        | 46      |
| déclaration   | 15.0       | 49        | 48      | ratification  | 13.3       | 44        | 43      |
| ratification  | 13.4       | 44        | 43      | general   | 10.7       | 55        | 49      |
| le  | 12.9       | 291       | 163     | member  | 9.9        | 31        | 30      |
| général   | 12.6       | 50        | 47      | day   | 8.6        | 24        | 24      |
| membres   | 9.7        | 32        | 31      | signature   | 8.6        | 24        | 24      |
| présent   | 8.9        | 83        | 65      | into  | 8.4        | 28        | 27      |
| vigueur   | 8.5        | 36        | 33      | of  | 8.1        | 873       | 496     |
| suit  | 8.4        | 23        | 23      | acceptance  | 8.0        | 22        | 22      |
| acceptation   | 8.0        | 22        | 22      | approval  | 7.6        | 21        | 21      |
| approbation   | 7.7        | 21        | 21      | date  | 7.3        | 32        | 29      |
| signature   | 7.6        | 25        | 24      | at  | 6.7        | 54        | 43      |
| mois  | 7.4        | 32        | 29      | first   | 6.3        | 21        | 20      |
| qui   | 7.3        | 91        | 67      | force   | 6.1        | 31        | 27      |
| territoires   | 7.1        | 19        | 19      | signed  | 6.1        | 16        | 16      |
| premier   | 7.1        | 19        | 19      | month   | 6.1        | 16        | 16      |
| entrera   | 6.8        | 18        | 18      | notification  | 6.1        | 16        | 16      |
| et  | 6.8        | 18        | 18      | territoires   | 6.1        | 16        | 16      |
| jour  | 6.4        | 17        | 17      | deposit   | 5.7        | 15        | 15      |
| tout  | 6.1        | 67        | 50      | being   | 5.7        | 26        | 23      |

Figure 13 : Spécificités des volets SOURCE et CIBLE via mkAlign

## 1.2 Chargement d'une nouvelle base intégrant 2 textes alignés

Le même type de calcul est possible sur une base intégrant 2 textes alignés d'une même langue, par exemple deux traductions alignées d'un même texte. Nous considérons ici 2 traductions<sup>2</sup> de l'ouvrage d'Hemingway « Le vieil homme et la mer » : celle de J. Dutourd (1952) et celle de F. Bon (2012).

| source="TRADUCTION1"   | target="TRADUCTION2"  |
|--|---|
| Il était une fois un vieil homme, tout seul dans son bateau qui pêchait au milieu du Gulf Stream. En quatre-vingt-quatre jours, il n'avait pas pris un poisson. Les quarante premiers jours, un jeune garçon l'accompagna ; mais au bout de ce temps, les parents du jeune garçon déclarèrent que le vieux était décidément et sans remède salado ce qui veut dire aussi guignard qu'on peut l'être. On embarqua donc le gamin sur un autre bateau, lequel, en une semaine, ramena trois poissons superbes. Chaque soir le gamin avait la tristesse de voir le vieux rentrer avec sa barque vide. Il ne manquait pas d'aller à sa rencontre et l'aidait à porter les lignes serrées en spirales, la gaffe, le harpon, ou la voile roulée autour du mât. La voile était rapiécée avec de vieux sacs de farine ; ainsi repliée, elle figurait le drapeau en berne de la défaite. # | Le vieil homme était maigre et sec, avec des rides comme des coups de couteau sur la nuque. Les taches brunes de cet inoffensif cancer de la peau que cause la réverbération du soleil sur la mer des Tropiques marquaient ses joues ; elles couvraient presque entièrement les deux côtés de son visage ; ses mains portaient les entailles profondes que font les filins au bout desquels se débattaient les lourds poissons. Mais aucune de ces entailles n'était récente elles étaient vieilles comme les érosions d'un désert sans poissons. # |
| Le vieil homme pêchait seul dans le Gulf Stream sur son canot depuis quatre-vingt-quatre jours sans avoir pris un poisson. Les quarante premiers jours, le garçon était venu avec lui. Mais après ces quarante jours, les parents du garçon lui avaient dit que le vieil homme était finalement et définitivement salado, ce qui est la pire forme pour dire pas de chance, et selon leurs ordres, le garçon était parti sur un autre bateau, lequel avait pris trois gros poissons la première semaine. Cela le rendait triste, le garçon, de voir le vieil homme revenir chaque soir le canot vide, et toujours il le rejoignait pour l'aider à porter les lignes enroulées, la gaffe, le harpon, et la voile ferlée autour du mât. Une voile rapiécée avec des sacs de farine qui pendait ainsi comme le drapeau d'une permanente défaite. #                                  | Le vieil homme était maigre et hâve, avec de profondes rides dans l'arrière du cou. Sur ses joues, les taches brunes d'un cancer de la peau bénin à cause de la réflexion du soleil sur la mer des tropiques. Les taches lui tombaient de chaque côté du visage et ses mains gardaient les cicatrices profondément plissées des poissons hâlés lourdement sur la corde. Mais aucune de ces cicatrices n'était récente. Elles étaient aussi vieilles que l'érosion dans le désert sans poisson. #  |
| Tout en lui était vieux, sauf les yeux - et ils étaient de la même couleur que la mer, joyeux et invincibles. #  | Tout en lui était vieux, sauf les yeux - et ils étaient de la même couleur que la mer, joyeux et invincibles. #   |

Figure 14 : Base textométrique de 2 traductions alignées

<sup>2</sup> <http://www.tal.univ-paris3.fr/mkAlign/#p7b>



Comme précédemment, une base regroupant les 2 traductions est obtenue en concaténant les 2 fichiers. La figure précédente donne à voir le début des 2 volets dans la base ainsi constituée. Une fois le fichier chargé, la *Carte des sections* a cette allure :

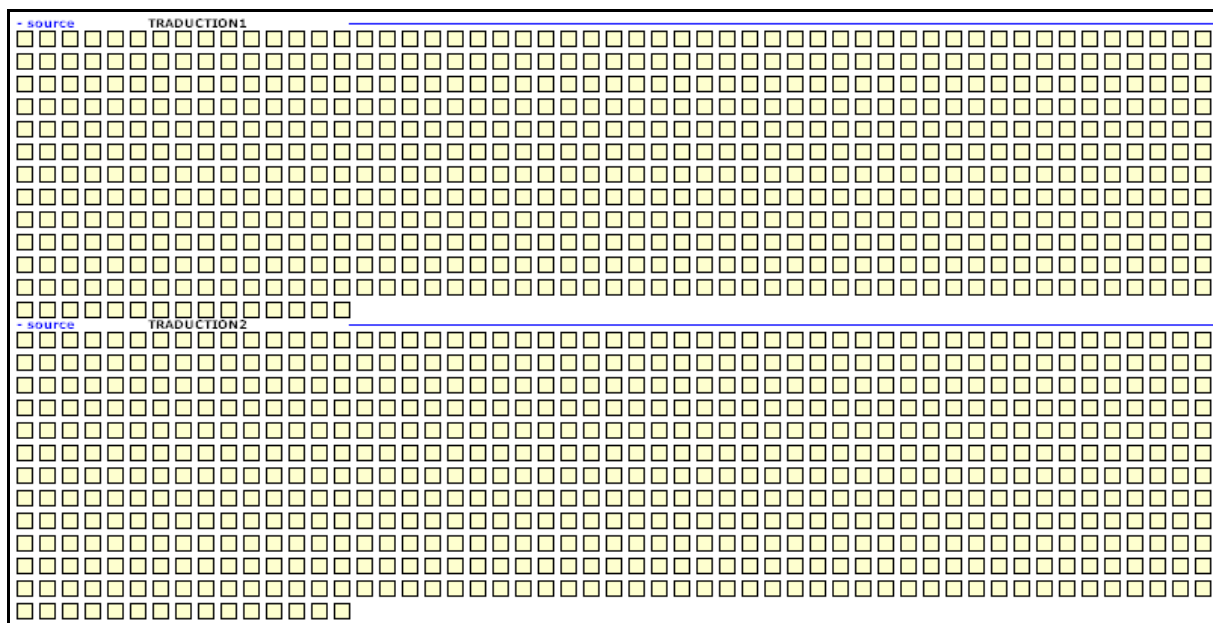


Figure 15: Carte des sections d'un alignement de traductions dans le Trameur

Le volet « TRADUCTION1 » correspond à la traduction de Dutourd, le volet « TRADUCTION2 » à celle de Bon. La projection de la forme « vieil » sur la carte produit cette ventilation :

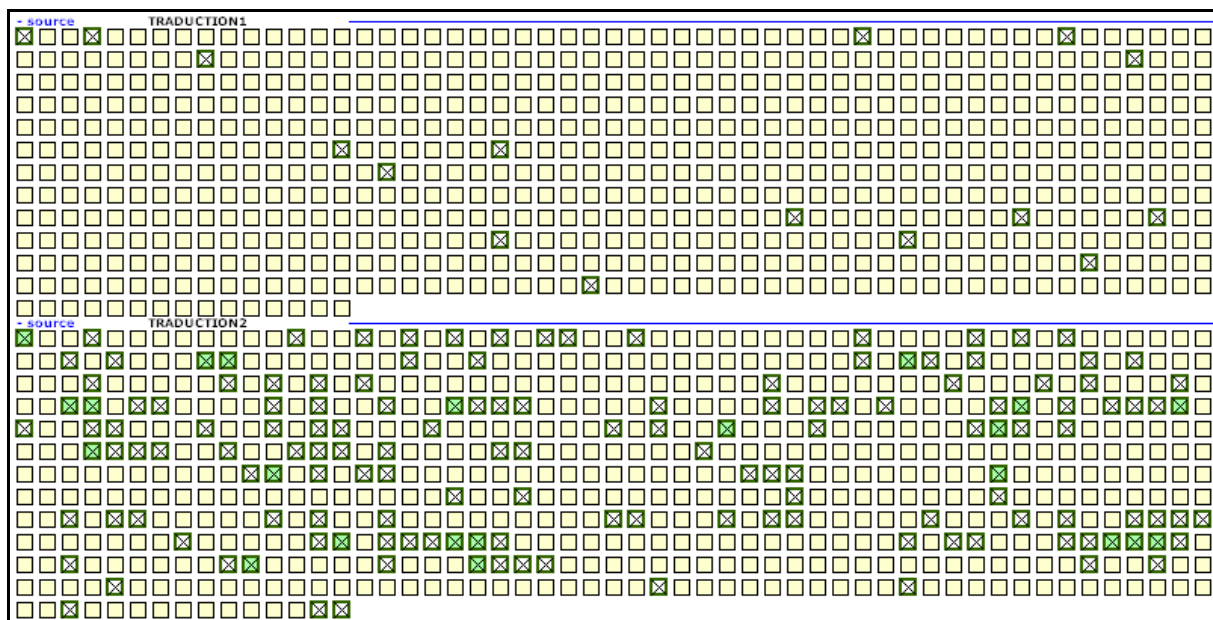


Figure 16 : Projection d'une forme sur les 2 traductions alignées

Cette forme est présente dans les 2 volets et massivement dans le second. Le calcul des cooccurrents présenté *supra* se fait en considérant les 2 volets comme 2 parties disjointes. Ici, la partie TRADUCTION2 est considérée comme le volet SOURCE et la partie TRADUCTION1 comme le volet CIBLE. Le résultat produit est le suivant :



**Spécificités du vocabulaire (Annotation Forme) sur les sections SOURCE contenant le motif : vieil**

Nombre d'occurrences du texte global = 29257  
 Nombre d'occurrences dans la partie visée = 10191  
 Seuil : 10

----- SOURCE -----  
 ----- (cible infra) -----  
 (Spécificités positives en haut de liste, négatives en bas)  
 Le fichier construit :  
 C:/DOCUME~1/FLEURY~1/MESDOC~1/HYDROP~1/WIP/WIP-TR~1/ALIGNE~1/ALIGNE~1/export/resultspecif-source-1414521180.txt

| Forme      | Ind-Specif | Fq-Totale | Fq-Partie |
|------------|------------|-----------|-----------|
| vieil      | **         | 185       | 185       |
| homme      | **         | 211       | 189       |
| Le         | 11.6       | 125       | 80        |
| le         | 8.8        | 1005      | 434       |
| requin     | 6.9        | 40        | 29        |
| sur        | 4.7        | 175       | 84        |
| ses        | 4.6        | 106       | 55        |
| lame       | 4.3        | 10        | 9         |
| était      | 4.3        | 251       | 113       |
| étaient    | 4.0        | 40        | 24        |
| yeux       | 3.9        | 28        | 18        |
| nageoires  | 3.9        | 9         | 8         |
| avaient    | 3.9        | 32        | 20        |
| avait      | 3.8        | 177       | 81        |
| parlaient  | 3.7        | 6         | 6         |
| frappa     | 3.6        | 13        | 10        |
| dents      | 3.5        | 8         | 7         |
| continuait | 3.3        | 10        | 8         |
| Parfois    | 3.3        | 5         | 5         |
| harpon     | 3.2        | 20        | 13        |
| ainsi      | 3.1        | 7         | 6         |
| surface    | 3.1        | 29        | 17        |
| et         | 3.0        | 989       | 379       |
| leurs      | 3.0        | 19        | 12        |
| son        | 3.0        | 157       | 69        |
| poissons   | 3.0        | 34        | 19        |
| eau        | 2.9        | 73        | 35        |
| chair      | 2.9        | 15        | 10        |
| salée      | 2.8        | 4         | 4         |
| élevait    | 2.8        | 4         | 4         |
| cou        | 2.8        | 4         | 4         |

Figure 17 : Spécificités sur volet SOURCE (TRADUCTION2) : sections contenant le motif "vieil"

----- CIBLE -----  
 ----- (cible miroir) -----  
 Spécificités du vocabulaire (Annotation Forme) sur les sections CIBLE miroir de celles contenant le motif : vieil

Nombre d'occurrences du texte global = 29982  
 Nombre d'occurrences dans la partie visée = 10133  
 Seuil : 10  
 (Spécificités positives en haut de liste, négatives en bas)

| Forme         | Ind-Specif | Fq-Totale | Fq-Partie |
|---------------|------------|-----------|-----------|
| vieux         | 32.8       | 269       | 186       |
| Le            | 11.3       | 181       | 104       |
| requin        | 7.7        | 35        | 27        |
| vieil         | 7.0        | 16        | 15        |
| le            | 5.6        | 889       | 358       |
| avait         | 4.9        | 163       | 78        |
| sa            | 4.7        | 123       | 61        |
| la            | 4.5        | 826       | 326       |
| du            | 4.1        | 209       | 93        |
| lame          | 3.8        | 6         | 6         |
| l             | 3.7        | 525       | 209       |
| surface       | 3.7        | 15        | 11        |
| eau           | 3.7        | 80        | 40        |
| harpon        | 3.7        | 19        | 13        |
| ses           | 3.6        | 98        | 47        |
| étaient       | 3.6        | 43        | 24        |
| Les           | 3.5        | 46        | 25        |
| frappa        | 3.5        | 8         | 7         |
| jaillissaient | 3.4        | 5         | 5         |
| chair         | 3.4        | 12        | 9         |
| grosses       | 3.4        | 5         | 5         |
| était         | 3.2        | 251       | 104       |
| plusieurs     | 3.1        | 7         | 6         |
| gueule        | 3.0        | 9         | 7         |
| mâchoires     | 3.0        | 9         | 7         |
| aperçut       | 3.0        | 17        | 11        |
| côté          | 3.0        | 17        | 11        |
| vit           | 3.0        | 11        | 8         |
| laissait      | 2.9        | 4         | 4         |
| faux          | 2.9        | 4         | 4         |
| enfonce       | 2.9        | 4         | 4         |
| poissons      | 2.9        | 33        | 18        |
| pièce         | 2.9        | 4         | 4         |

Figure 18 : Spécificités sur volet CIBLE (TRADUCTION1) :  
 sections MIROIR des sections contenant le motif "vieil"

Dans la traduction de F. Bon, on trouve plutôt la séquence « le vieil homme » et dans celle de J. Dutourd, la séquence « le vieux ».

## 2. Base textométrique : alignement au format TMX

Un alignement au format TMX permet lui aussi de réaliser un calcul contrastif de cooccurrences sur 2 des volets intégrés dans ce type d'alignement. Nous prendrons ici comme exemple de fichier TMX, un alignement intégrant 4 volets de l'ouvrage "Le vieil homme et la mer" : le volet anglais original (noté en), les 2 volets traduits en français déjà mentionnés (noté fr1 et fr2) et un volet traduit en espagnol (noté es).

Le fichier a l'allure suivante (affichage partiel du fichier dans un navigateur, puis dans un éditeur (code source)):

| Volet | Anglais  | Français (Dutourd)   | Français (Bon)   | Espagnol  |
|-------|--|--|--|---|
| 1     | <p>He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally <i>salajo</i>, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty, and he always went down to help him carry either the coiled lines or the gaff and harpoon and the sail that was furled around the mast. The sail was patched with flour sacks and, furled, it looked like the flag of permanent defeat.</p> | <p>Il était une fois un vieil homme, tout seul dans son bateau qui pêchait au milieu du Gulf Stream. En quatre-vingt-quatre jours, il n'avait pas pris un poisson. Les quarante premiers jours, un jeune garçon l'accompagnait, mais au bout de ce temps, les parents du jeune garçon déclaraient que le vieux était décidément et sans remède <i>salajo</i> ce qui veut dire aussi qu'on peut l'être. On embarqua donc le gamin sur un autre bateau, lequel, en une semaine, ramena trois poissons superbes. Chaque soir le gamin avait la tristesse de voir le vieux rentrer avec sa barque vide. Il ne manquait pas d'aller à sa rencontre et l'aidait à porter les lignes serrées en spirales, la gaffe, le harpon, ou la voile roulée autour du mât. La voile était rapiécée avec de vieux sacs de farine ; ainsi repliée, elle figurait le drapeau en berne de la défaite.</p> | <p>Le vieil homme pêchait seul dans le Gulf Stream sur son canot depuis quatre-vingt-quatre jours sans avoir pris un poisson. Les quarante premiers jours, le garçon était venu avec lui. Mais après ces quarante jours, les parents du garçon lui avaient dit que le vieil homme était finalement et définitivement <i>salajo</i>, ce qui est la pire forme pour dire pas de chance, et selon leurs ordres, le garçon était parti sur un autre bateau, lequel avait pris trois bons poissons la première semaine. C'est le vendredi triste, le garçon, de voir le vieil homme revenir chaque soir le canot vide, et toujours il le rejoignait pour l'aider à porter les lignes enroulées, la gaffe, le harpon et la voile roulée autour du mât. Une voile rapiécée avec des sacs de farine qui pendait ainsi comme le drapeau d'une permanente défaite.</p> | <p>Éra un viejo que pescaba solo en un bote en el Gulf Stream y hasta ochenta y cuatro días que no cogía un pez. En los primeros cuarenta días había tenido consigo a un muchacho. Pero después de cuarenta días sin haber pescado los padres del muchacho le habían dicho que el viejo estaba definitivo y eternamente <i>salajo</i>, lo cual era la peor forma de la mala suerte, y por orden de sus padres el muchacho había salido en otro bote que cogió tres buenos peces la primera semana. El viernes triste el muchacho veía al viejo regresar todos los días con su bote vacío, y siempre bajaba a ayudarle a cargar los rollos de sardal o el bicho y el arpón y la vela enrollada al mástil. La vela estaba remendada con sacos de harina y enrollada, parecía una bandera en permanente derrota.</p> |
| 2     | <p>The old man was thin and gaunt with deep wrinkles in the back of his neck. The brown blotches of the benevolent skin cancer the sun brings from its reflection on the tropic sea were on his cheeks. The blotches ran well down the sides of his face and his hands had the deep creases scars from handling heavy fish on the cords. But none of these scars were fresh. They were as old as erosions in a fishless desert.</p>  | <p>Le vieil homme était maigre et sec, avec des rides comme des coups de couteau sur la nuque. Les taches brunes de cet inoffensif cancer de la peau que cause la réflexion du soleil sur la mer des Tropiques marquaient ses joues ; elles couvraient presque entièrement les deux côtés de son visage ; ses mains portaient les entailles profondes que font les filets au bout desquels se débattaient les lourds poissons. Mais aucune de ces entailles n'était récente elles étaient vieilles comme les aréoles d'un désert sans poissons.</p>  | <p>Le vieil homme était maigre et séca, avec des rides profondes rides dans l'arrière du cou. Sur ses joues, les taches brunes d'un cancer de la peau dans le cadre de la réflexion du soleil sur la mer des Tropiques. Les taches lui tombant de chaque côté de visage et ses mains gardaient les cicatrices profondément plissées des poissons hilés lourdement sur la corde. Mais aucune de ces cicatrices pour être récente. Elles étaient aussi vieilles que l'érosion dans le désert sans poisson.</p>   | <p>El viejo era flaco y desgarbado, con arrugas profundas en la parte posterior del cuello. Las manchas manchadas del benigno cáncer de la piel que el sol produce con sus reflejos en el mar tropical estaban en sus mejillas. Estas manchas cubrían por los lados de su cara hasta bastante abajo y sus manos tenían las hondas cicatrices que causa la manipulación de las cuerdas cuando se jatan los grandes peces. Pero ninguna de estas cicatrices era reciente. Eran tan viejas como las aréolas de un árido desierto.</p>  |

Figure 19 : Alignement TMX dans un navigateur (avec feuille de styles)

```

<?xml version="1.0" encoding="UTF-8" ?>
<xml:stylesheet type="text/xml" href="styles-tmx.xml" ?>
<tmx version="1.4">
<header adminlang="en" creationdate="20120223T162335Z" creationtool="mkAlign" creationtoolversion="2.00 (2.0b134)" datatype="xml" o-tmf="unknown" segtype="block" srclang="en"/>
<body>
<seg>
<tu xml:lang="en">
<p>He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salajo, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty, and he always went down to help him carry either the coiled lines or the gaff and harpoon and the sail that was furled around the mast. The sail was patched with flour sacks and, furled, it looked like the flag of permanent defeat.
</tu>
<tu xml:lang="fr1">
<p>Il était une fois un vieil homme, tout seul dans son bateau qui pêchait au milieu du Gulf Stream. En quatre-vingt-quatre jours, il n'avait pas pris un poisson. Les quarante premiers jours, un jeune garçon l'accompagnait, mais au bout de ce temps, les parents du jeune garçon déclaraient que le vieux était décidément et sans remède salajo ce qui veut dire aussi qu'on peut l'être. On embarqua donc le gamin sur un autre bateau, lequel, en une semaine, ramena trois poissons superbes. Chaque soir le gamin avait la tristesse de voir le vieux rentrer avec sa barque vide. Il ne manquait pas d'aller à sa rencontre et l'aidait à porter les lignes serrées en spirales, la gaffe, le harpon, ou la voile roulée autour du mât. La voile était rapiécée avec de vieux sacs de farine ; ainsi repliée, elle figurait le drapeau en berne de la défaite.
</tu>
<tu xml:lang="fr2">
<p>Le vieil homme était maigre et séca, avec des rides profondes rides dans l'arrière du cou. Sur ses joues, les taches brunes d'un cancer de la peau dans le cadre de la réflexion du soleil sur la mer des Tropiques. Les taches lui tombant de chaque côté de visage et ses mains gardaient les cicatrices profondément plissées des poissons hilés lourdement sur la corde. Mais aucune de ces cicatrices pour être récente. Elles étaient aussi vieilles que l'érosion dans le désert sans poisson.
</tu>
<tu xml:lang="es">
<p>Éra un viejo que pescaba solo en un bote en el Gulf Stream y hasta ochenta y cuatro días que no cogía un pez. En los primeros cuarenta días había tenido consigo a un muchacho. Pero después de cuarenta días sin haber pescado los padres del muchacho le habían dicho que el viejo estaba definitivo y eternamente salajo, lo cual era la peor forma de la mala suerte, y por orden de sus padres el muchacho había salido en otro bote que cogió tres buenos peces la primera semana. El viernes triste el muchacho veía al viejo regresar todos los días con su bote vacío, y siempre bajaba a ayudarle a cargar los rollos de sardal o el bicho y el arpón y la vela enrollada al mástil. La vela estaba remendada con sacos de harina y enrollada, parecía una bandera en permanente derrota.
</tu>
</seg>
</body>
</tmx>

```

Figure 20 : Alignement TMX (code source)

La figure précédente donne à voir 2 zones de texte avec pour chacune les 4 volets de langue (dans un ordre identique : en, fr1, fr2, es). Pour la première zone de texte dans le code source, seules les 2 premières langues sont affichées dans l'éditeur.

Un alignement TMX peut être schématisé ainsi :

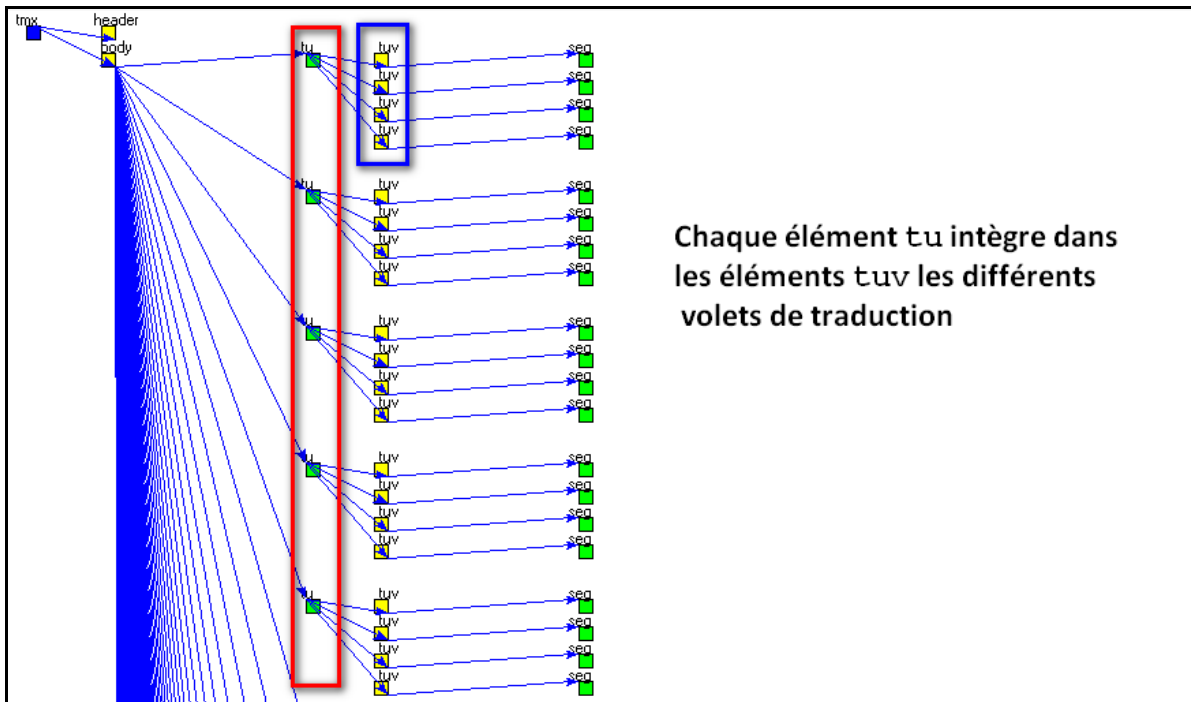


Figure 21 : Alignement TMX , arbre des éléments

Chaque élément  $tu$  de l'arbre XML encapsule les différents volets de traduction dans autant d'éléments  $tuv$ . L'alignement se lit de manière verticale pour chaque unité de langue ( $tu$ ) dans la figure précédente.

Pour contraster 2 volets, il faut donc pouvoir mettre au jour les 2 parties visées : pour la source, tous les éléments  $tuv$  d'une langue donnée contenant la forme visée ; pour la cible, les éléments  $tuv$  miroirs des précédents dans une autre langue.

Dans le *Trameur*, la *Carte des sections* peut mettre au jour un alignement des différents volets d'un fichier TMX :

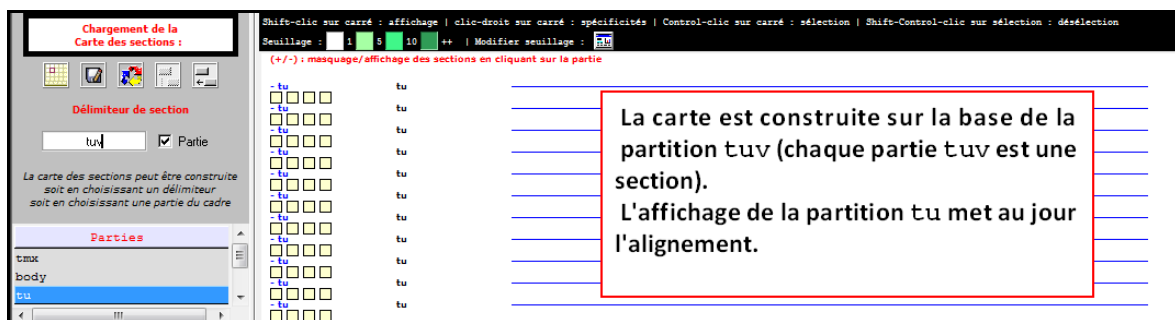


Figure 22 : Carte des sections d'un alignement TMX avec le Trameur

Ici la *Carte des sections* est générée à partir de la partition  $tuv$  i.e l'ensemble des parties  $tuv$  de la base textométrique construite à partir de l'alignement TMX. Ces parties intégrées dans le *Cadre* de la base textométrique sont identifiables par leurs positions sur la *Trame* : positions de début et de fin. Dans la carte construite, chaque colonne de sections coïncide avec un volet de l'alignement (1 colonne = 1 langue).

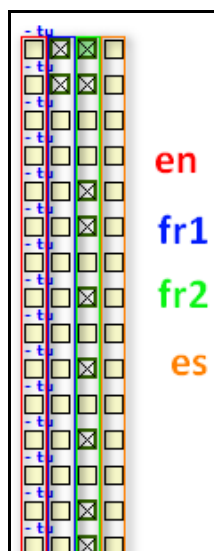


Figure 23 : 1 colonne = 1 volet = 1 langue

Identifier une colonne consiste donc à spécifier le volet visé *i.e* la langue des sections  $\tau_{uv}$  de cette colonne. Les langues sont identifiables par les attributs `lang` des balises  $\tau_{uv}$  (visibles dans le fichier source *supra*). La *Carte des sections* précédente est construite sur la partition  $\tau_{uv}$ , les volets visés doivent être identifiés par la valeur de la langue du volet source et celle du volet cible. Au final, identifier 2 volets distincts consiste à donner un couple de langues parmi celles encodées dans les balises  $\tau_{uv}$  du fichier TMX. Une section source donnée ayant une section cible miroir dans un voisinage de sections de longueur égal au nombre de volets de l'alignement (*i.e* le nombre de balises  $\tau_{uv}$  dans les éléments  $\tau_u$ ) et les positions respectives d'une section source et d'une section cible restant inchangées tout au long du fichier TMX. Au final, les couples de sections alignées sont aisément identifiables.

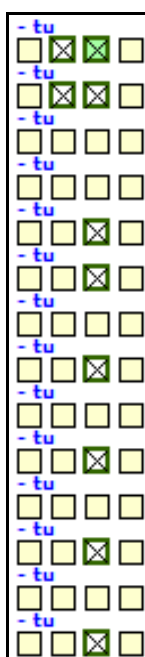


Figure 24 : Distribution d'une forme sur la carte de l'alignement

Une forme étant projetée sur la carte des sections, le calcul de ses cooccurrents consiste à spécifier les 2 volets à contraster. Si on projette la forme "vieil", a priori, seuls les volets fr2 et fr1 contiennent des sections avec cette forme (colonne 2 et 3 de la carte), la figure précédente donne à voir un extrait de cette distribution.

Le paramétrage du calcul des cooccurrents consiste à définir la valeur de 2 langues à contraster. Nous allons examiner ci-dessous les 2 couples de volets (fr2,fr1) et (fr2,en).

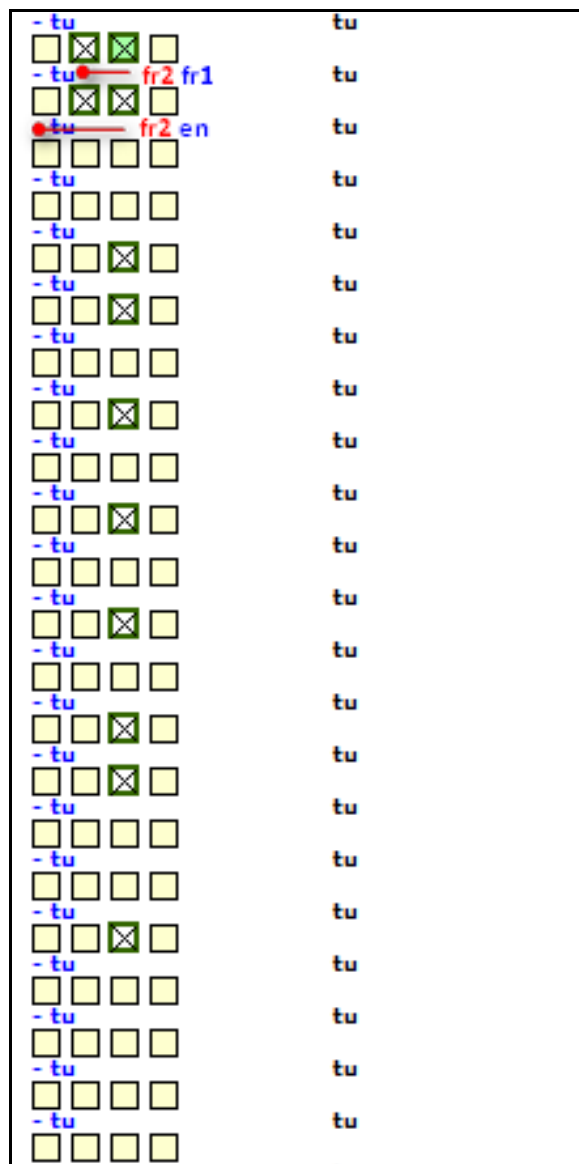


Figure 25 : Contraste de volets (2 à 2)

Pour le couple (fr2, fr1), le paramétrage du calcul est visible ci-dessous : la case à cocher TMX permet de préciser que la base chargée est un alignement au format TMX, les 2 zones de saisie V1 et V2 permettent de spécifier les langues à contraster (attributs lang des éléments `tu`), ici :

V1 <= xml:lang=fr2 (SOURCE)

V2 <= xml:lang=fr1(CIBLE)

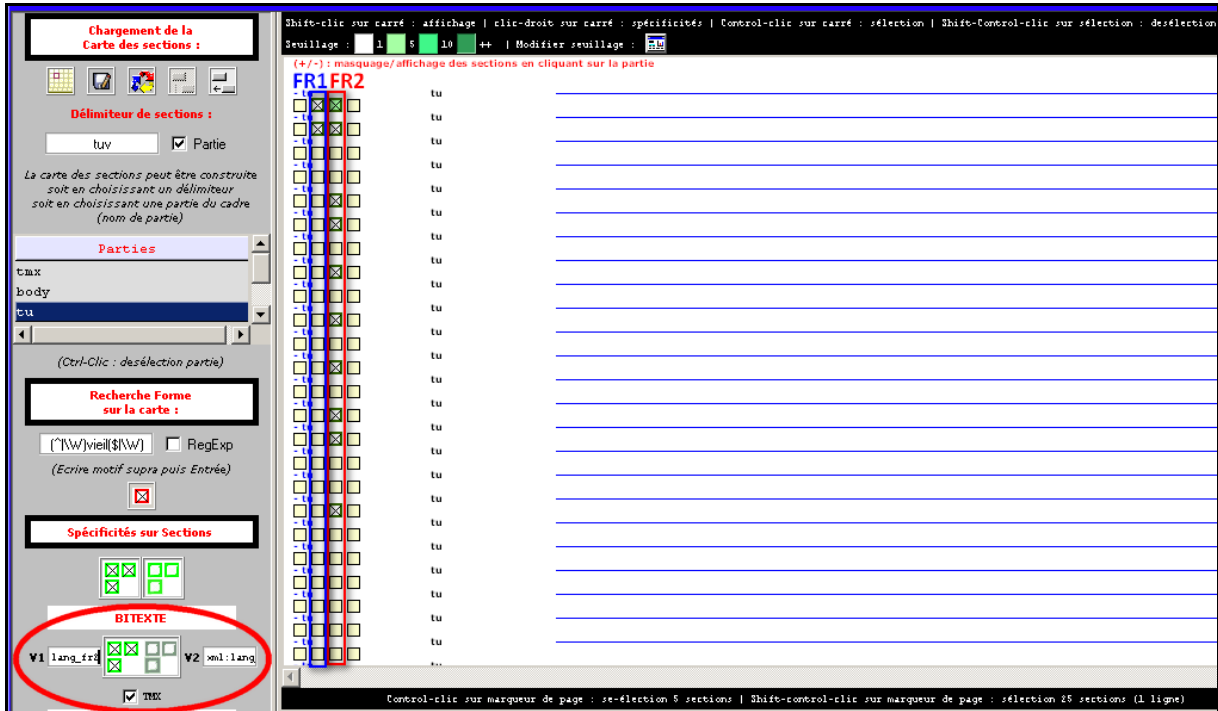


Figure 26 : Contraste de 2 volets (FR2, FR1)

Et le résultat du calcul à l'allure suivante :

**Spécificites du vocabulaire (Annotation Forme) sur les sections SOURCE contenant le motif : vieil**

Nombre d'occurrences du texte global = 29257  
 Nombre d'occurrences dans la partie visée = 10191  
 Seuil : 10

----- SOURCE -----  
 ----- (cible infra) -----  
 (Spécificites positives en haut de liste, negatives en bas)  
 Le fichier construit :  
 C: /DOCUME~1/FLBURY~1/MESDOC~1/MYDRO~1/WIP/WIP-TR~1/ALICNE~1/ALICNE~1/export/resultspecif-source-1415131422.txt

| Forme      | Ind-Specif | Eg-Totale | Eg-Partie |
|------------|------------|-----------|-----------|
| vieil      | **         | 185       | 185       |
| homme      | **         | 211       | 189       |
| Le         | 11.6       | 125       | 80        |
| le         | 8.8        | 1005      | 434       |
| requin     | 6.9        | 40        | 29        |
| sur        | 4.7        | 175       | 84        |
| ses        | 4.6        | 106       | 55        |
| lame       | 4.3        | 10        | 9         |
| était      | 4.3        | 251       | 113       |
| étaient    | 4.0        | 40        | 24        |
| yeux       | 3.9        | 28        | 18        |
| nageoires  | 3.9        | 9         | 8         |
| avaient    | 3.9        | 32        | 20        |
| avait      | 3.8        | 177       | 81        |
| parlaient  | 3.7        | 6         | 6         |
| frappa     | 3.6        | 13        | 10        |
| dents      | 3.5        | 8         | 7         |
| continuait | 3.3        | 10        | 8         |
| Parfois    | 3.3        | 5         | 5         |
| harpon     | 3.2        | 20        | 13        |
| ainsi      | 3.1        | 7         | 6         |
| surface    | 3.1        | 29        | 17        |
| et         | 3.0        | 989       | 379       |
| leurs      | 3.0        | 19        | 12        |
| son        | 3.0        | 157       | 69        |
| poissons   | 3.0        | 34        | 19        |
| eau        | 2.9        | 73        | 35        |
| chair      | 2.9        | 15        | 10        |
| salée      | 2.8        | 4         | 4         |
| élevait    | 2.8        | 4         | 4         |

Figure 27 : Spécificités volet SOURCE (FR2)

| CIBLE   |            |           |           |
|---|------------|-----------|-----------|
| Specificites du vocabulaire (Annotation Forme) sur les sections CIBLE miroir de celles contenant le motif : vieil |            |           |           |
| Nombre d'occurrences du texte global = 29982  |            |           |           |
| Nombre d'occurrences dans la partie visee = 10133   |            |           |           |
| Seuil : 10  |            |           |           |
| (Specificites positives en haut de liste, negatives en bas)   |            |           |           |
| Forme   | Ind-Specif | Fq-Totale | Fq-Partie |
| vieux   | 32.8       | 269       | 186       |
| Le  | 11.3       | 181       | 104       |
| requin  | 7.7        | 35        | 27        |
| vieil   | 7.0        | 16        | 15        |
| le  | 5.6        | 889       | 358       |
| avait   | 4.9        | 163       | 78        |
| sa  | 4.7        | 123       | 61        |
| la  | 4.5        | 826       | 326       |
| du  | 4.1        | 209       | 93        |
| lame  | 3.8        | 6         | 6         |
| l   | 3.7        | 525       | 209       |
| surface   | 3.7        | 15        | 11        |
| eau   | 3.7        | 80        | 40        |
| harpon  | 3.7        | 19        | 13        |
| ses   | 3.6        | 98        | 47        |
| étaient   | 3.6        | 43        | 24        |
| Les   | 3.5        | 46        | 25        |
| frappa  | 3.5        | 8         | 7         |
| jaillissaient   | 3.4        | 5         | 5         |
| chair   | 3.4        | 12        | 9         |
| grosses   | 3.4        | 5         | 5         |
| était   | 3.2        | 251       | 104       |
| plusieurs   | 3.1        | 7         | 6         |
| gueule  | 3.0        | 9         | 7         |
| mâchoires   | 3.0        | 9         | 7         |
| aperçut   | 3.0        | 17        | 11        |
| côté  | 3.0        | 17        | 11        |
| vit   | 3.0        | 11        | 8         |
| laissait  | 2.9        | 4         | 4         |
| faux  | 2.9        | 4         | 4         |

Figure 28 : Spécificités volet CIBLE (FR1)

On retrouve ci-dessus les résultats produits précédemment.

De même, pour le couple (fr2, en), le paramétrage du calcul est visible ci-dessous :

Figure 29 : Contraste de 2 volets (FR2, EN)

Et le résultat produit a l'allure suivante :



**Specificites du vocabulaire (Annotation Forme) sur les sections SOURCE contenant le motif : vieil**

Nombre d'occurrences du texte global = 29257  
 Nombre d'occurrences dans la partie visee = 10191  
 Seuil : 10

----- SOURCE -----  
 ----- (cible infra) -----  
 (Specificites positives en haut de liste, negatives en bas)  
 Le fichier construit :  
 C:/DOCUME~1/FLEURY~1/MESDOC~1/MYDROP~1/WIP/WIP-TR~1/ALIGNE~1/ALIGNE~1/export/resultspecif-source-1415131664.txt

| Forme      | Ind-Specif | Fq-Totale | Fq-Partie |
|------------|------------|-----------|-----------|
| vieil      | **         | 185       | 185       |
| homme      | **         | 211       | 189       |
| Le         | 11.6       | 125       | 80        |
| le         | 8.8        | 1005      | 434       |
| requin     | 6.9        | 40        | 29        |
| sur        | 4.7        | 175       | 84        |
| ses        | 4.6        | 106       | 55        |
| lame       | 4.3        | 10        | 9         |
| était      | 4.3        | 251       | 113       |
| étaient    | 4.0        | 40        | 24        |
| yeux       | 3.9        | 28        | 18        |
| nageoires  | 3.9        | 9         | 8         |
| avaient    | 3.9        | 32        | 20        |
| avait      | 3.8        | 177       | 81        |
| parlaient  | 3.7        | 6         | 6         |
| frappa     | 3.6        | 13        | 10        |
| dents      | 3.5        | 8         | 7         |
| continuait | 3.3        | 10        | 8         |
| Parfois    | 3.3        | 5         | 5         |
| harpon     | 3.2        | 20        | 13        |
| ainsi      | 3.1        | 7         | 6         |
| surface    | 3.1        | 29        | 17        |
| et         | 3.0        | 989       | 379       |
| leurs      | 3.0        | 19        | 12        |
| son        | 3.0        | 157       | 69        |
| poissons   | 3.0        | 34        | 19        |
| eau        | 2.9        | 73        | 35        |

Figure 30 : Spécificités volet SOURCE (FR2)

----- CIBLE -----  
 Specificites du vocabulaire (Annotation Forme) sur les sections CIBLE miroir de celles contenant le motif : vieil

Nombre d'occurrences du texte global = 26847  
 Nombre d'occurrences dans la partie visee = 9512  
 Seuil : 10  
 (Specificites positives en haut de liste, negatives en bas)

| Forme     | Ind-Specif | Fq-Totale | Fq-Partie |
|-----------|------------|-----------|-----------|
| old       | 49.4       | 247       | 200       |
| man       | 40.6       | 264       | 199       |
| The       | 12.1       | 194       | 115       |
| the       | 7.6        | 2119      | 858       |
| shark     | 7.2        | 46        | 33        |
| his       | 6.3        | 430       | 197       |
| her       | 6.0        | 29        | 22        |
| water     | 4.7        | 106       | 56        |
| were      | 4.5        | 138       | 69        |
| wide      | 4.4        | 13        | 11        |
| was       | 4.4        | 435       | 188       |
| hit       | 4.3        | 26        | 18        |
| blade     | 4.2        | 10        | 9         |
| had       | 3.8        | 203       | 93        |
| came      | 3.7        | 41        | 24        |
| as        | 3.6        | 191       | 87        |
| by        | 3.5        | 29        | 18        |
| jaws      | 3.5        | 17        | 12        |
| turtles   | 3.4        | 8         | 7         |
| s         | 3.4        | 77        | 39        |
| neck      | 3.3        | 5         | 5         |
| afraid    | 3.3        | 5         | 5         |
| flesh     | 3.3        | 5         | 5         |
| side      | 3.3        | 32        | 19        |
| rose      | 3.2        | 14        | 10        |
| harpoon   | 3.2        | 20        | 13        |
| surface   | 3.2        | 16        | 11        |
| and       | 3.0        | 1233      | 475       |
| closed    | 3.0        | 7         | 6         |
| birds     | 3.0        | 7         | 6         |
| they      | 3.0        | 99        | 47        |
| Sometimes | 3.0        | 7         | 6         |
| coming    | 2.9        | 15        | 10        |

Figure 31 : Spécificités volet CIBLE (EN)

### 3. Importer une base « d'alignements » multi-annotée

Les calculs présentés ci-dessus peuvent être réalisés sur tous les niveaux d'annotation disponibles dans la base textométrique chargée dans le *Trameur*.

Nous allons illustrer cette fonctionnalité en construisant une base multi-annotée et intégrant 2 volets alignés. Pour cela, nous procédons de la manière suivante :

- Chargement avec le *Trameur* de la base complète précédente (convention FR et EN) avec étiquetage en Français via *treetagger*
- Export de la base précédente : création d'une base multi-annotée (BASE-FR-EN) d'un bitexte, le volet FR est « bien » annoté, le volet EN ne l'est pas.
- Chargement avec le *Trameur* d'une base constituée uniquement du volet EN avec étiquetage en Anglais via *treetagger*
- Export de la base précédente : création d'une base multi-annotée du volet EN (BASE-EN)
- Dans un éditeur, suppression dans la base BASE-FR-EN de la *Trame* correspondant au volet EN et insertion « au même endroit » de la *Trame* de la base BASE-EN. Dans la base BASE-FR-EN, les 2 volets sont désormais « bien » annotés (chaque volet ayant son propre jeu d'annotations)

La base BASE-FR-EN peut finalement être importée dans le *Trameur*

La base résultante est visible dans la figure suivante (2 extraits de la *Trame* dans lesquels on donne à voir le début des 2 volets) : chaque item de la *Trame* dispose de 3 niveaux d'annotation : forme (balise <f>), lemme (balise <l>) et POS (balise <c>).

```

<items>
<item type="delim" pos="1"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="2"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="forme" pos="3"><f>convention</f><c>NOM</c><l>convention</l></item>
<item type="delim" pos="4"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="5"><f>de</f><c>PRP</c><l>de</l></item>
<item type="delim" pos="6"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="7"><f>sauvegarde</f><c>VER_pres</c><l>sauvegarder</l></item>
<item type="delim" pos="8"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="9"><f>des</f><c>PRP_det</c><l>du</l></item>
<item type="delim" pos="10"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="11"><f>droits</f><c>NOM</c><l>droit</l></item>
<item type="delim" pos="12"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="13"><f>de</f><c>PRP</c><l>de</l></item>
<item type="delim" pos="14"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="15"><f>l</f><c>VER_simp</c><l>l</l></item>
<item type="delim" pos="16"><f>'</f><c>DELIM</c><l>'</l></item>
<item type="forme" pos="17"><f>homme</f><c>NOM</c><l>homme</l></item>
<item type="delim" pos="18"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="19"><f>et</f><c>KON</c><l>et</l></item>
<item type="delim" pos="20"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="21"><f>des</f><c>PRP_det</c><l>du</l></item>
<item type="delim" pos="22"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23"><f>libertés</f><c>NOM</c><l>liberté</l></item>
<item type="delim" pos="24"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="25"><f>fondamentales</f><c>ADJ</c><l>fondamental</l></item>
<item type="delim" pos="26"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="forme" pos="27"><f>telle</f><c>PRO_DEM</c><l>tel</l></item>
<item type="delim" pos="28"><f> </f><c>DELIM</c><l>BLANK</l></item>
-u\-- BASE-FR-EN-3-annotations-utf8.xml (SGML)--L37-- 14-----
<item type="delim" pos="23950"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="23951"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="forme" pos="23952"><f>convention</f><c>NN</c><l>convention</l></item>
<item type="delim" pos="23953"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23954"><f>for</f><c>IN</c><l>for</l></item>
<item type="delim" pos="23955"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23956"><f>the</f><c>DT</c><l>the</l></item>
<item type="delim" pos="23957"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23958"><f>protection</f><c>NN</c><l>protection</l></item>
<item type="delim" pos="23959"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23960"><f>of</f><c>IN</c><l>of</l></item>
<item type="delim" pos="23961"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23962"><f>human</f><c>JJ</c><l>human</l></item>
<item type="delim" pos="23963"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23964"><f>rights</f><c>NNS</c><l>right</l></item>
<item type="delim" pos="23965"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23966"><f>and</f><c>CC</c><l>and</l></item>
<item type="delim" pos="23967"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23968"><f>fundamental</f><c>JJ</c><l>fundamental</l></item>
<item type="delim" pos="23969"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23970"><f>freedoms</f><c>NNS</c><l>freedom</l></item>
<item type="delim" pos="23971"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="forme" pos="23972"><f>as</f><c>IN</c><l>as</l></item>
<item type="delim" pos="23973"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23974"><f>amended</f><c>VVD</c><l>amend</l></item>
<item type="delim" pos="23975"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23976"><f>by</f><c>IN</c><l>by</l></item>
<item type="delim" pos="23977"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="23978"><f>protocol</f><c>NN</c><l>protocol</l></item>
-u\-- BASE-FR-EN-3-annotations-utf8.xml (SGML)--L23991--52%-----

```

Figure 32 : Base textométrique multiannotée d'un alignement de textes bilingues

Une fois importée, les mêmes calculs peuvent être réalisés sur la *Carte des sections* mettant au jour l'alignement.

On pourrait bien entendu disposer d'une base multi-annotée et intégrant 2 volets alignés en procédant d'une autre manière : on pourrait par exemple projeter (via un programme ou à la main) des annotations sur la *Trame* d'une base intégrant 2 volets alignés.

## 4. Calculs contrastifs sur une base textométrique de textes alignés

Au delà du calcul des cooccurrents sur 2 volets alignés d'une base textométrique *via* la *Carte des sections*, il est possible de mener des calculs textométriques sur chacun des volets en utilisant le *Gestionnaire de sélections* et notamment les modules permettant de lancer des calculs sur telle ou telle partie de la base de travail.

Sur la base précédente par exemple, les modules de calcul encadrés dans la figure suivante peuvent être lancés sur un des volets de la base après sélection du volet visé dans la liste des parties de la partition volet.

| Partition | valeur | Partie | début | fin   | fq item | FQ item | Fq Max | orme Max |
|-----------|--------|--------|-------|-------|---------|---------|--------|----------|
| volet     | FR     |        | 1     | 23948 | 1439    | 11211   | 690    | de       |
| volet     | EN     |        | 23949 | 46489 | 1227    | 10564   | 1003   | the      |

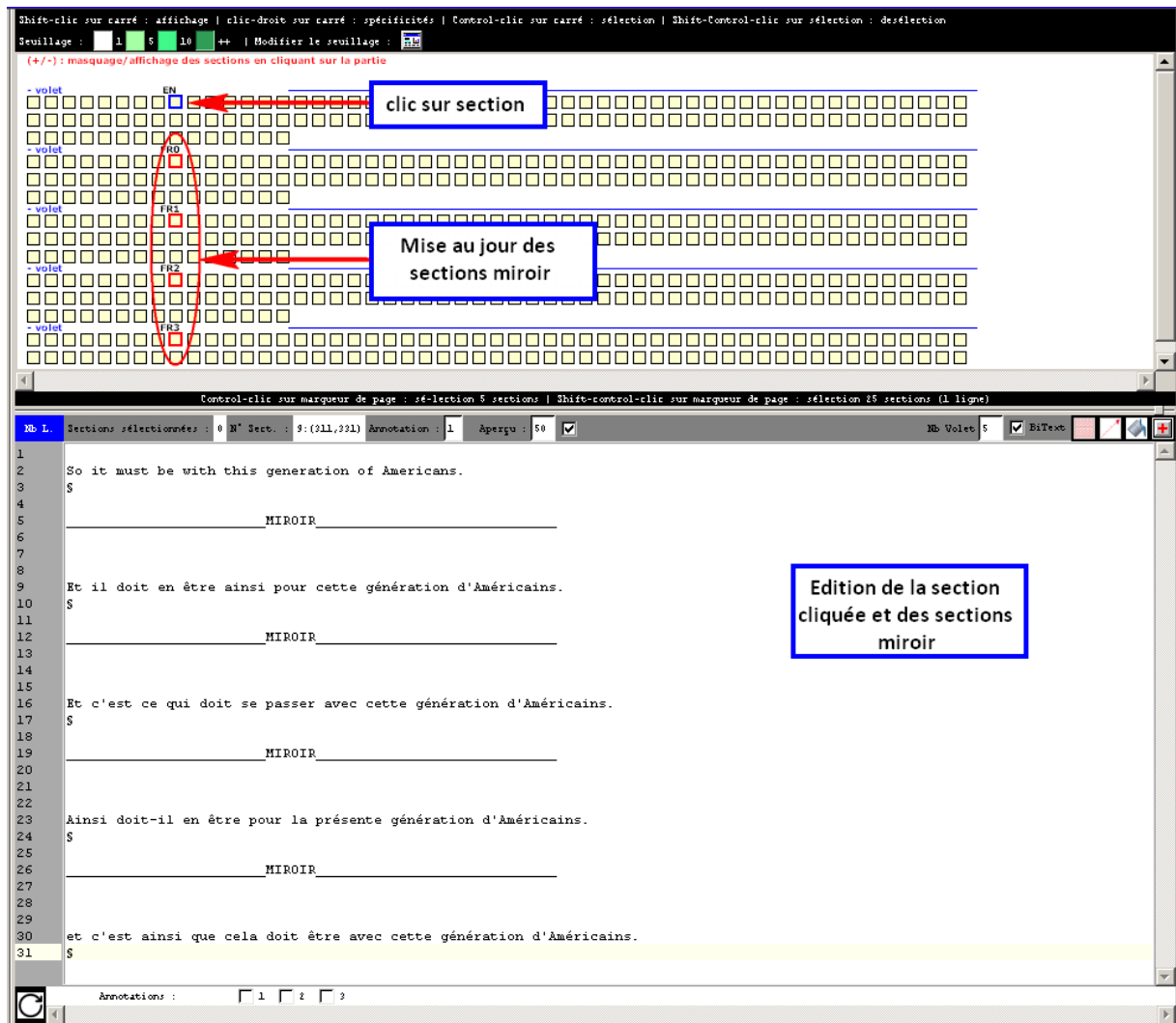
Figure 33 : Calculs contrastifs sur une base textométrique d'alignements

## 5. Outils et méthodes d'exploration d'un alignement

Dans la suite de ce document, 2 bases textométriques sont utilisées :

1. [Base bilingue ParTUT2Trameur \(français/anglais\)](#). Les 2 volets de cette base disposent de 9 niveaux d'annotation (Alignement de treebank). Cette base permet de tester les traitements spécifiques pour les bases textométriques intégrant 2 textes. [Alignement de Treebank pour le Trameur](#) (pdf) : présentation du processus de transcodage de 2 treebanks ParTUT pour construire une base Textométrique.
2. [Base bilingue "Investiture Obama" \(français/anglais\)](#). Cette base regroupe le discours original en anglais prononcé par B. Obama le 20 janvier 2009 à Washington, publié sur le site de The New York Times ([volet EN](#)), et 4 traductions françaises de ce discours (volets FR0-1-2-3). Les traductions ont été récupérées sur le site officiel de la Maison Blanche ([volet FR0](#)), sur les sites des journaux français Le Monde ([volet FR1](#)) et Libération ([volet FR2](#)), ainsi que sur le site de RFI ([volet FR3](#)). Les 5 volets de cette base disposent de 3 niveaux d'annotation (forme, lemme, catégorie). Cette base permet de tester les traitements spécifiques pour les bases textométriques intégrant 2 textes alignés.

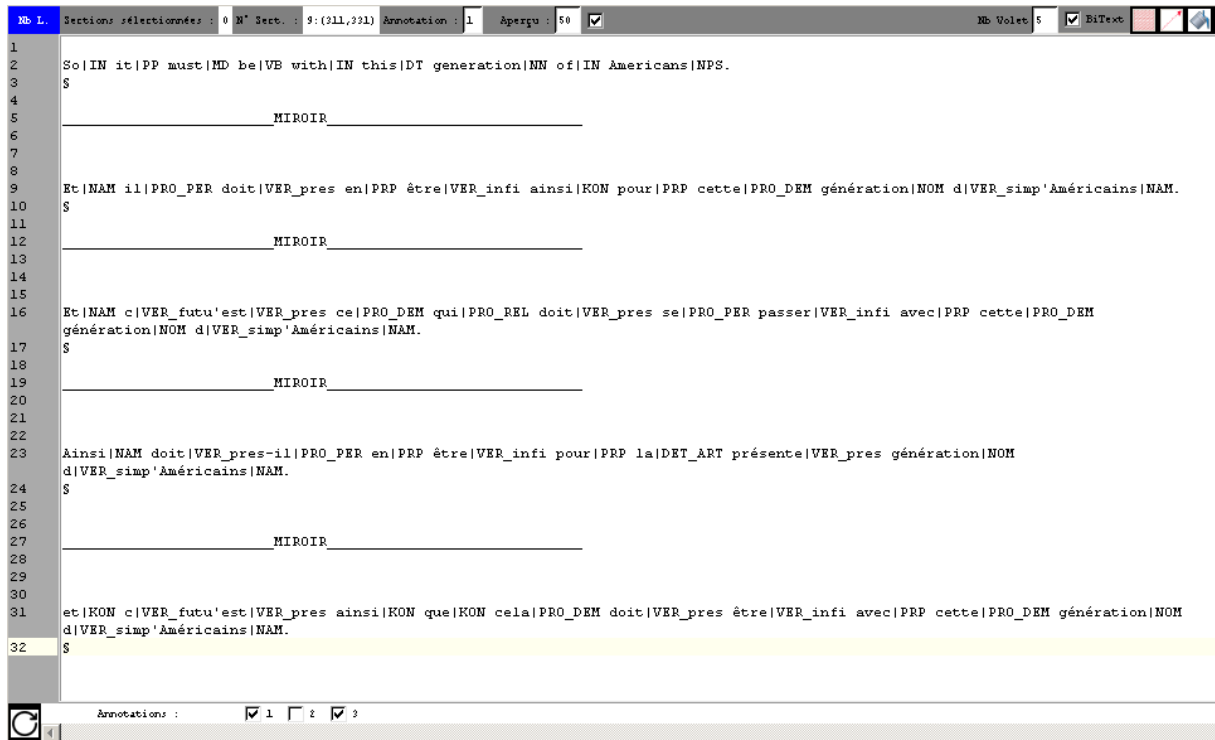
### 5.1 Edition d'un alignement multimulti



## 5.2 Opérations dans une section multi-volets

Toutes les opérations disponibles dans l'éditeur de section peuvent désormais être réalisées sur une section multi-volets.

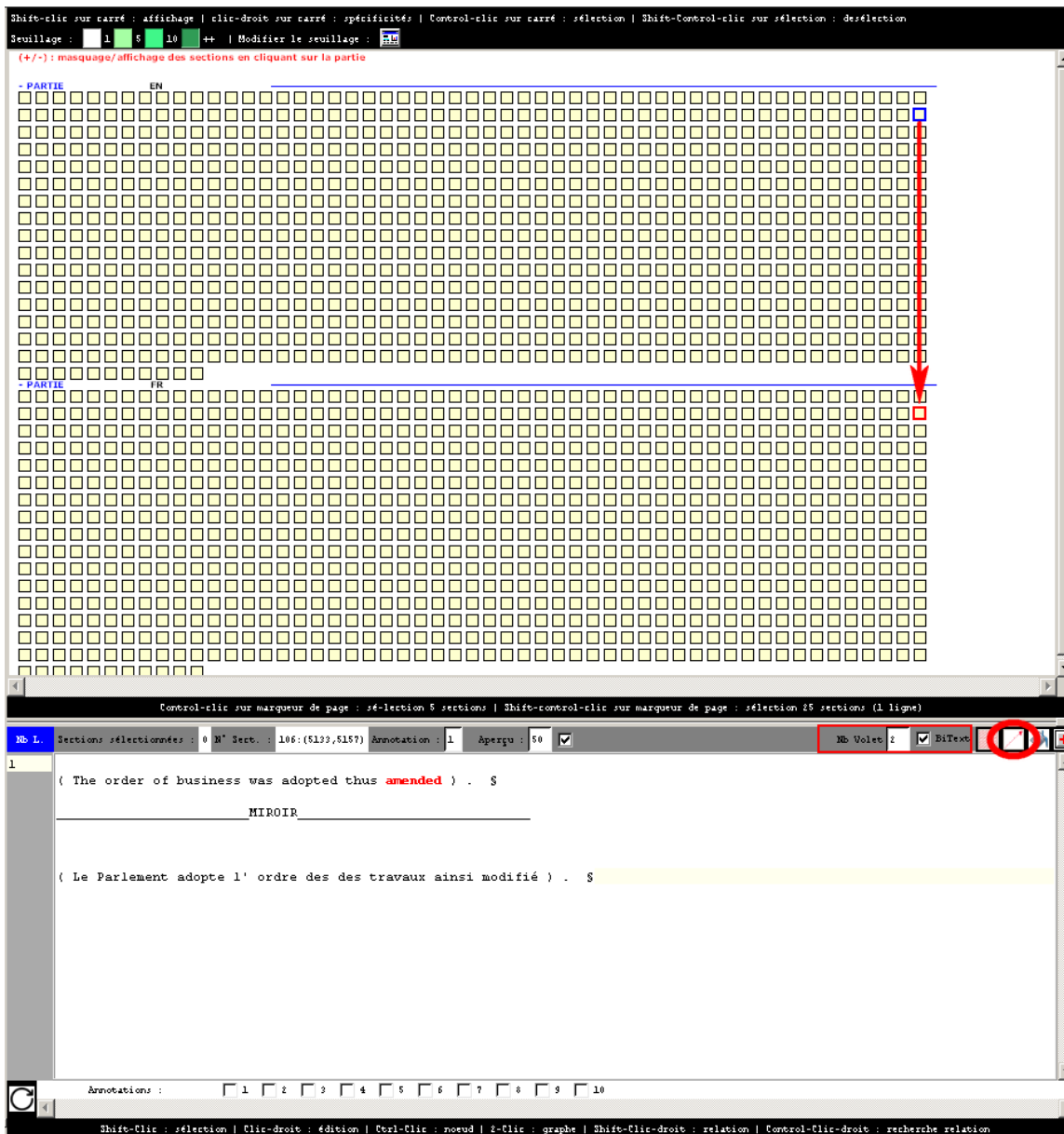
Ci-dessous, l'affichage de 3 niveaux d'annotations sur les 5 sections éditées conjointement :



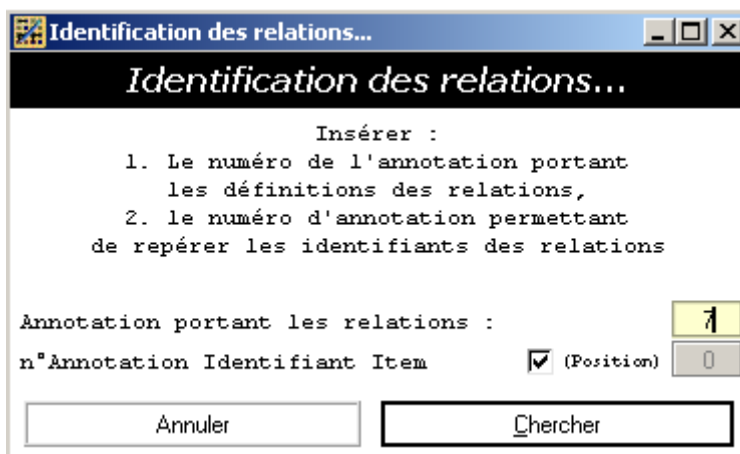
### 5.3 Graphe des relations de dépendances sur une section multi-volets

Cette fois-ci, on considère que la base parTUT est chargée dans le Trameur. Cette base intègre un alignement de treebanks (cf documentation en ligne) : un volet EN et un volet FR. On édite conjointement une section et son miroir :





Le bouton « Affichage des relations de dépendance dans la section » (cerclé de rouge ci-dessus) permet de mettre au jour le graphe de relations sur les 2 zones textuelles alignées ici :



Le Trameur : documentation complémentaire

