

Exploring translation corpora with *MkAlign*

Serge Fleury, Maria Zimina

Centre of Textometrics *SYLED-CLA²T*
Paris Sorbonne University – Paris 3

Abstract

This paper presents a series of experiments devoted to the development of a new tool for multilingual textometric exploration of translation corpora. We propose to use bi-text topography to facilitate the study of lexical equivalences on quantitative bases. The suggested approach opens up new horizons for interactive exploration of translation resources of multilingual texts in a variety of fields of study: translation, foreign language learning and teaching, bilingual terminology, lexicography, etc.

Key-words: bi-text map, quantitative analysis, translation correspondences.

Introduction

In a constantly changing information society, researchers and practitioners are continually faced with growing volumes of multilingual text data of all kinds: electronic archives of translated texts, multilingual databases, international web sites, etc. Different communities are increasingly interested in multilingual text processing for a variety of reasons. In this respect, development of computer tools for exploring intertextual correspondences between related parts of multilingual texts is an important research issue.

Considerable progress has been made in the field of parallel text alignment and bilingual lexicon extraction (Véronis, 2000). Current text alignment algorithms perform quite successfully on the sentence level. However, there is a need to continue research in finer-grained text alignment. At the same time, huge volumes of non-parallel, yet comparable corpora are currently available in almost any field of knowledge. In this respect, the challenge is to discover links between different parts of such corpora on the word level (Déjean and Gaussier, 2002).

Automatic discovery of lexical correspondences in multilingual texts is closely connected to empirical study of the translation process. The development of translation description models is an intricate task. In order to deal with the inherent complexity of translation correspondences, current computer systems extend the notion of multilingual text processing to deal with multi-level language structures. Linguistic and/or pragmatic knowledge of different nature is used to identify potential word candidates for lexical alignment which remains quite difficult.

Recent developments have shown that quantitative methods used in *textometric analysis* open up new horizons for identifying translation correspondences in bilingual texts (Zimina 2004ab), (Zimina 2005ab). Most of these methods have not been exploited in the field of multilingual text processing to their full potential. The present article outlines a series of experiments devoted to the development of a new textometric tool for creating, editing and exploring translation corpora: *MkAlign* (Fleury and Zimina, 2006).

1. Textometric analysis of multilingual texts

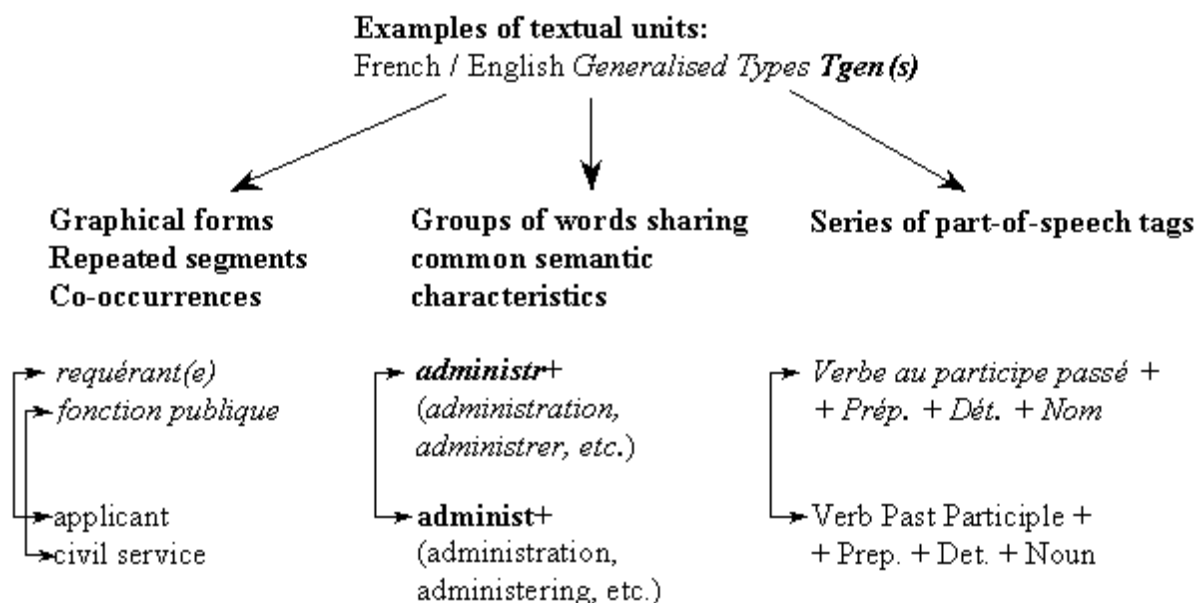
In a French-speaking community, the term *textometric analysis* (in French: “analyse textométrique”) covers a series of methods that enable the researcher to formally reorganise textual sequences and to conduct statistical analysis based on the *vocabulary* of a corpus of texts (Salem 1987), (Lebart, Salem and Berry 1997).

The vocabulary is a set of distinct graphical forms found in a corpus. A *graphical form* is a series of *non-delimiting characters* bounded by two *delimiting characters*. The occurrences of graphical forms are entirely defined by the list of delimiting characters chosen by the user. Once the list of delimiting characters is established (e.g.: .,:;!/?/_\ '""()[]{}\$\$ and the *space* character), other characters become non-delimiting characters. Any series of non-delimiting characters bounded by delimiting characters is considered an *occurrence* (token). A form is then identified as a *type* corresponding to identical occurrences in a corpus of texts.

Abrupt changes that occur in the distribution of a graphical form in different contexts (parts) of a corpus may raise questions concerning the identification of other related graphical units (different manifestations of the same lemma, forms related on the semantic level, etc.). Textometric tools (such as **Lexico3** and **COOCS**)¹ allow the analyst not only to subdivide the text into graphical forms, but also to identify other types of textual units (see *Figure 1*):

- **Repeated Segments** (Salem 1987): series of consecutive forms found in the corpus with frequency greater than or equal to 2.
- **Co-occurrences**: simultaneous, but not necessarily contiguous, presence of occurrences of two forms in a given context (phrase, section, etc.).
- **Multiple co-occurrences** (Martinez 2003): lexical networks formed by simultaneous presence of occurrences of several related forms in a given context (phrase, section, etc.).
- **Generalised Types** or **Tgen(s)** (Lamalle and Salem 2002): textual units defined by the user with the help of tools that permit automatic regrouping of occurrences in the text (e.g.: occurrences of forms starting with a given sequence of characters, such as *administ+*: administration, administrative, administer, etc.). The resulting “object” can then be processed like a “usual” form. Tools based on *regular* (or *rational*) *expressions* look-up facilities, frequently used in computing, considerably simplify the search for such groups.

The *Tgen(s)* selection has been largely implemented in **Lexico3** textometric toolbox (Lamalle *et al.*, 2004). In order to facilitate the creation of *types* that collect occurrences of different graphical forms according to a common characteristic, the user might work with dynamic lexical storage facilities, such as *Word-store*. This feature allows for the memorization of forms, segments, *Tgen(s)* for later use.



*Figure 1: Examples of textual units **Tgen(s)***

2. Textometric browsing with a bi-text map

As we have shown on figure 1, the concept of *type/token* relationship might be extended to provide a much broader definition of textual units or generalised types *Tgen(s)*. By following these principles, it becomes possible to consider a “spatial” approach to localisation of textual units within the text corpora. The concept of textometric browsing enables the user to move among the results produced by different quantitative methods and the original bi-text (Lamalle and Salem, 2002; Lamalle *et al.*, 2004).

In bilingual corpora, it is convenient to visually identify corresponding parts of texts through *bi-text topography* (Zimina 2004ab; 2005ab). In order to visualize corresponding parts, the bi-text must include tags that indicate the parallel structure of the corpus. The insertion of *keys* is crucial in the preparation of the corpus. Such pre-coding permits the study of the distribution of occurrences of a given textual unit within the sections thus defined. The selected keys allow the user to compare corresponding textual fragments (sections, paragraphs, phrases, etc., cf. *Figure 2*).

In parallel text processing, the insertion of section delimiters can be performed through parallel matching of corresponding parts in different languages: logical partitions (author, year, date, etc.) and marks for breathing (sentences, paragraphs, etc.).

The **MkAlign** *bi-text map* allows for the visualization of the corpus cut into corresponding sections by raising one (or several) characters (e.g.: ‘§’) to the rank of *parallel section delimiters*. This visualisation permits the user to produce an automatic selection of sections in one of the monolingual parts of the bi-text where any textual unit under study (word, collocation, repeated segment, etc.) is found. The selected sections of the map are highlighted. At any moment, the user is allowed to reiterate a topographic selection in any corpus part for further investigation of translation correspondences on the word level. In order to describe how textometric browsing works, we shall provide some corpus-based examples.

3. Mapping lexical correspondences in parallel contexts with *MkAlign*

This section illustrates some principles of interactive textometric browsing in parallel contexts. For illustration purposes, we shall use a piece of French-English parallel corpus *Convention*.²

Step One (see *Figure 3-4*):

- The user picks up any *Tgen* from the dictionary of graphical forms (*DICT*) or the list of available textual units (*LISTES*) by right mouse click.
- It is also possible to create an entirely new *Tgen* using *regular expressions* within *Recherche Source/Recherche Cible* zone of the bi-text map (*MAP*). In our example, we have decided to represent simultaneously distributions of the French type ***gouvern+*** [government, gouverner, etc.] and the English type ***govern+*** [government, governing, etc.].
- “Crossed” squares of the map display text sections containing at least one occurrence of the selected types. The content of relevant sections is visualized in the lower part of the window by clicking on the squares representing these sections on the map.
- Following the process of *text resonance* (Lamalle and Salem 2002), activated section(s) in one of the corpus parts automatically produce a parallel selection of the equivalent section(s) in the other corpus part. The mapping zone can be re-initialised at any time, after having recorded a graph in a report.

Step Two (see *Figures 5-6*):

- Symmetric colouring of the map displays the bi-text sections (corresponding contexts) in which the French type ***gouvern+*** is translated by the English type ***govern+***.
- Asymmetric colouring of the bi-text map reveals sections in which the French type ***gouvern+*** does not correspond to the English type ***govern+***. These asymmetric distributions of corresponding textual units (breaking points) are even more interesting for translation study than the cases of perfect symmetry (Zimina, 2005b).
- Identification of non-corresponding sections enables to check and correct alignment via bi-text editor (*ALIGN*) and to localize omissions or unusual translation correspondences:
gouvernement du district ~ regional council
la législation sur la fonction publique ~ legislation governing the civil service.

As a rule, these singular contexts are particularly difficult to reveal through traditional bilingual lexicon extraction methods due to their low repetition frequency and/or unusual semantic or syntactic properties.

Our “topographic approach” of translated texts enables to draw the attention of the user to very subtle translation phenomena through a relatively straightforward technique of bi-text map exploration based on distributional analysis. The related text is visualized by clicking on the squares representing these sections on the map. It becomes possible to go through the text displayed in the toolbox in order to discover meaning of translation correspondences.

Step Three (see Figures 6-7):

- Specific bi-text sections highlighted on the map might be exported in XML format through report creation (*EXPORT-XML*). For example, figure 7 shows an aligned bi-text fragment generated automatically from initial parallel corpus. For this particular filtering, only bi-text sections containing the French type *gouvern*+ have been activated on the map.

Upcoming research will help to extend existing features of *MkAlign* towards non-parallel yet comparable corpora. We are currently working on contextual vectors identification to capture corresponding areas in related texts. In this respect, *MkAlign* offers many possibilities of report generation through exporting and importing source and/or target corresponding text zones in different formats: *xml*, *html*, *txt*. In other words, the user “captures” special areas of bilingual corpora according to particular distributional criteria (absence or presence of certain lexical items or word groups). Generated sub-corpora are then re-imported into the bi-text editor (*ALIGN*) for cross-check, editing and alignment. These interactive text management facilities are already available in the currently distributed v. 1.038 *MkAlign*. Future work will help to identify specific application scenarios and allow for further advances in this direction.

Conclusions

Bilingual lexicon extraction from translation corpora lacks flexibility when it comes to explore multiple translation correspondences between polysemous lexical units.

In this article, we presented a new tool for cross-language exploration of bilingual corpora: *MkAlign*. This tool is based on quantitative methods of textometric analysis. The concept of textometric browsing is central in corpus investigation. It is unique in that it allows the user to maintain control over the entire corpus exploration, from initial segmentation to the extraction and editing of text resources. The units that are then counted automatically originate entirely from the list of delimiters provided by the user, with no need for outside dictionary resources.

The suggested approach offers new means for context-based study of translation corpora and for detection of multiple translation correspondences.

mkAlign@CLA2T-P3 1.038

mkAlign 1.038 CLA2T/SYLED [U. DE PARIS 3, Sorbonne nouvelle] <http://tal.univ-paris3.fr>

HOME PARAM ALIGN MAP DICT PARAL3 LISTES EXPORT-XML EXPORT EXPORT-L3

<p>Segmenteur</p> <p>\$</p> <p>Recherche Source</p> <p>...</p> <p>Recherche Cible</p> <p>...</p> <p></p> <p></p>	<p>Invokant les de la convention, il se plaignait de ne pas avoir eu accès, dans le cadre d'une procédure pouvant conduire à sa condamnation à une peine d'emprisonnement, à un tribunal indépendant et impartial.</p> <p>§</p>	<p>relying on of the convention, he complained that in proceedings which could result in his being sentenced to a term of imprisonment he had not had access to an independent and impartial tribunal.</p> <p>§</p>	86	<input checked="" type="checkbox"/> R
	<p>il voyait en outre dans l'obligation de porter la ceinture de sécurité une atteinte à sa vie privée contraire aux de la convention.</p> <p>§</p>	<p>he also complained of an interference in his private life on account of the requirement to wear a safety-belt, contrary to of the convention.</p> <p>§</p>	87	<input checked="" type="checkbox"/> R
	<p>la commission a retenu la requête quant au grief soulevé sur le terrain de l' et l'a déclarée irrecevable pour le surplus.</p> <p>§</p>	<p>the commission declared the application admissible as to the complaint raised under and declared it inadmissible as to the remainder.</p> <p>§</p>	88	<input checked="" type="checkbox"/> R
	<p>dans son rapport (), elle conclut à la violation de l' en ce qui concerne l'accès à un tribunal (unanimité) et estime qu'aucune question distincte ne se pose sur le terrain de l' quant au défaut d'audience (unanimité).</p> <p>§</p>	<p>in its report of (), it expressed the opinion that there had been a violation of as regards access to a court (unanimously) and that no separate issue arose under as to the failure to hold a hearing (unanimously).</p> <p>§</p>	89	<input type="checkbox"/> W
	<p>le texte intégral de son avis et de l'opinion concordante dont il s'accompagne figure en annexe au présent arrêt (1).</p> <p>§</p>	<p>the full text of the commission's opinion and of the concurring opinion contained in the report is reproduced as an annex to this judgment (1).</p> <p>§</p>	90	<input type="checkbox"/> W

18 / 388

0 1934 3 << >> 18

Page 18

mkAlign@CLA2T-P3 1.038

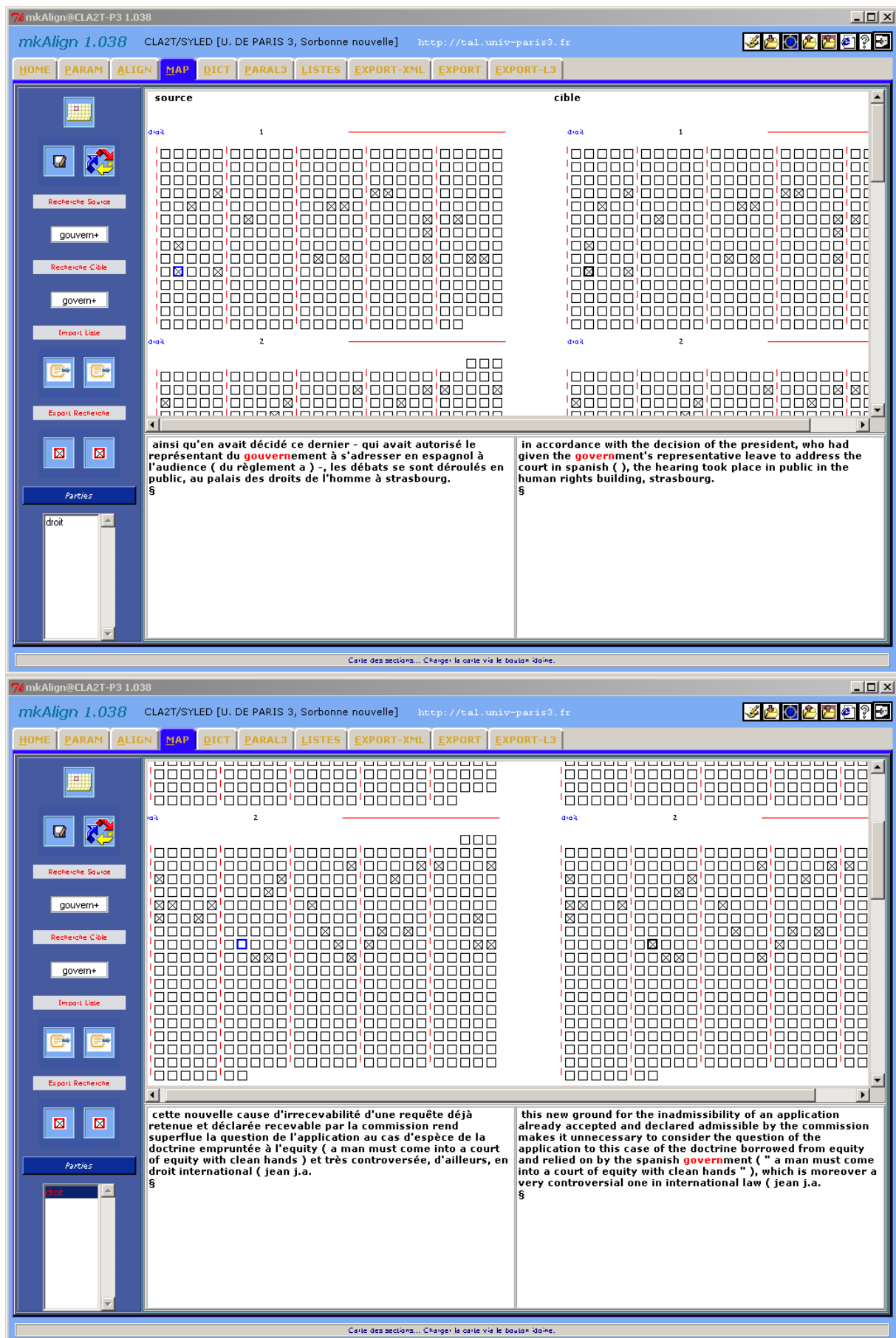
mkAlign 1.038 CLA2T/SYLED [U. DE PARIS 3, Sorbonne nouvelle] <http://tal.univ-paris3.fr>

HOME PARAM ALIGN MAP DICT PARAL3 LISTES EXPORT-XML EXPORT EXPORT-L3

Freq	Forme	Freq	Forme
(clic-droit sur une forme -> visualisation dans la représentation cartographique de l'alignement)			
2785	de	4689	the
2163	la	2359	of
1576	l	1538	to
1163	à	1212	in
1068	le	938	a
853	et	909	and
835	d	822	that
771	en	507	court
584	les	421	was
564	que	414	by
532	une	408	for
529	des	374	it
528	du	352	not
463	un	350	as
455	a	342	be
442	il	332	had
399	au	327	applicant
388	cour	316	is
383	par	300	or
378	dans	300	s
337	qu	290	on
326	pour	258	with
310	pas	235	case
307	ne	221	been
252	ou	217	this
242	elle	216	which
240	son	203	convention
230	sur	199	an
227	n	187	his
214	requérant	167	he
206	s	167	law
198	convention	164	see
189	se	158	appeal
188	ce	157	her
187	avait	153	civil
186	ci	143	from
182	qui	142	have
179	est	136	at
156	aux	129	administrative
156	était	129	its
147	sa	128	proceedings
146	droit	126	paragraph
130	paragraphe	120	any
129	si	120	judgment
123	été	117	above

Page 18

Figures 2-3: Bi-text segmentation, alignment and editing with *MkAlign*



Figures 4-5: Locating distribution similarities and breaking points with *MkAlign*

CLA²T [U. DE PARIS 3, Sorbonne nouvelle]
mkAlign Export

Dimanche 03 Decembre 2006 20:48:47

Fichier source initial : C:\Program Files\MkAlign38\mkAlign-1.38-win32\corpus\trad_jurFR.txt
Fichier cible initial : C:\Program Files\MkAlign38\mkAlign-1.38-win32\corpus\trad_jurEN.txt
Forme graphique (RegExp) utilisée pour générer l'export : gouvern+
Segmenteur choisi pour l'alignement initial : §

MKA-ed	SOURCE	CIBLE
80	l'instrument de ratification de la convention, déposé par le gouvernement autrichien, contient notamment une réserve ainsi libellée : §	the instrument of ratification of the convention deposited by the austrian government contains, inter alia, a reservation worded as follows : §
91	conclusions presentées à la cour par le gouvernement §	final submissions to the court by the government §
92	dans son mémoire, le gouvernement invite la cour à dire. §	in their memorial the government asked the court to hold that §
103	le gouvernement n'en disconvient pas. §	this was not disputed by the government. §
114	il ne ferait aucun doute, en effet, qu'en désignant dans ladite réserve les "mesures de privation de liberté", le gouvernement autrichien visait aussi les procédures menant à celles-ci. §	there could be no doubt that by the reference in that reservation to "measures for the deprivation of liberty" the austrian government had meant to include proceedings resulting in such measures. §
145	le gouvernement combat cette thèse, tandis que la commission y souscrit en substance. §	the government contested this view, whereas the commission accepted it. §
147	il faut que la décision d'une autorité administrative ne remplissant pas elle-même les exigences de l' de la convention - comme c'est le cas en l'espèce de la direction de la police fédérale et du gouvernement du land (paragraphes 6 et 7 ci-dessus) - subisse le contrôle ultérieur d'un " organe judiciaire de pleine juridiction " (voir notamment, mutatis mutandis, les arrêts albert et, série a., öztürk précité, p-22, et, série a., §	albert and of, series a., ; öztürk, previously cited, p-22, ; and of, series a., §
170	d'après le gouvernement , la cour n'a pas compétence pour annuler les condamnations prononcées par les juridictions nationales et ordonner le remboursement des amendes. §	the government contended that the court had no jurisdiction to quash convictions pronounced by national courts or to order repayment of fines. §
212	l'affaire a été déférée à la cour par le gouvernement du royaume d'espagne (" le gouvernement "), puis par la commission européenne des droits de l'homme (" la commission "), dans le délai de trois mois qu'ouvrent les de la convention. §	the case was referred to the court by the government of the kingdom of spain (" the government ") and by the european commission of human rights (" the commission "), within the three-month period laid down by and of the convention. §

Figures 6-7: Browsing in parallel contexts and XML report generation with *MkAlign*

Notes

1. On *Lexico3* and *COOCS* Tools : <http://www.cavi.univ-paris3.fr/ilpga/syled/outils-cla2t.htm>
2. The corpus *Convention* is composed of the *European Convention for the Protection of Human Rights and Fundamental Freedoms* as well as a series of related protocols and judgements of the European Court of Human Rights. This corpus was used in a variety of methodological studies within the research centre *SYLED-CLA²T* (Paris 3 University). See, for instance, (Zimina, 2005b).

References

Books:

Lebart, L., Salem, A. and Berry L. (1997) *Exploring Textual Data* (Boston: Kluwer Academic Publishers).

Salem, A. (1987) *Pratique des segments répétés : essai de statistique textuelle* (Paris : Klincksieck).

Véronis, J. (ed.) (2000) *Parallel Text Processing: Alignment and use of translation corpora* (Dordrecht: Kluwer Academic Publishers).

Articles in Journals:

Déjean H, Gaussier, É. (2002) Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicométrica*, no. 'Corpus alignés'. Available on-line from <http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm>

Articles in Conference Proceedings:

Lamalle, C. and Salem, A. (2002) Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *JADT'02, Saint-Malo, 2002*, 403-412. Available on-line from <http://www.cavi.univ-paris3.fr/lexicometrica/jadt/>

Zimina, M. (2004a) L'alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. *JADT'04, Louvain-la-Neuve, 2004*, 1195-1202. Available on-line from <http://www.cavi.univ-paris3.fr/lexicometrica/jadt/>

Zimina, M. (2005a) Bi-text Topography and Quantitative Approaches of Parallel Text Processing. *Corpus Linguistics Conference Series*, Vol. 1, no. 1 (Centre for Corpus Research, Birmingham University). Available on-line from <http://www.corpus.bham.ac.uk/PCLC/>

PhD Theses:

Martinez, W. (2003) *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels* (PhD Thesis, Paris Sorbonne University – Paris 3).

Zimina, M. (2004b) *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles* (PhD Thesis, Paris Sorbonne University – Paris 3). Available on-line from http://www.cavi.univ-paris3.fr/ilpga/ed/student/stmz/ED268-PagePersoMZ_fichiers/stmz/page8.htm

On-line publications:

Lamalle, C., Martinez, W., Fleury, S., Salem, A., Fracchiolla, B., Kuncova, A., Lande, B., Maisondieu, A. and Poirot-Zimina, M. (2004) Lexico3 Textometric toolbox User's manual (Centre of Textometrics *CLA²T*, Paris Sorbonne University – Paris 3). Available on-line from <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuelsL3/L3-usermanual.pdf>

Fleury, S., Zimina, M. (2006) MkAlign. Manuel d'utilisation (Centre of Textometrics *CLA²T*, Paris Sorbonne University – Paris 3). Available on-line from <http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm>

Zimina, M. (2005b) Equivalences traductionnelles. *Rapports d'analyse : Navigations textométriques avec Lexico3* (Centre of Textometrics *CLA²T*, Paris Sorbonne University – Paris 3). Available on-line from <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/navigations/navigation-bitexte.pdf>