Bi-text Topography and Quantitative Approaches of Parallel Text Processing

Maria Zimina

Bi-text Topography and Quantitative Approaches of Parallel Text Processing

Dr Zimina Maria
Centre of Textometrics SYLED-CLA^{2}T
Paris Sorbonne University – Paris 3
zimina@msh-paris.fr

Abstract:

This paper presents a series of experiments devoted to the development of new tools for multilingual textometric exploration of translation corpora. I propose to use *bi-text topography* to facilitate the study of lexical equivalences on a quantitative basis. The *map of parallel sections* allows for the visualization of the corpus cut into corresponding sections by raising one (or several) characters to the rank of parallel section delimiters. The exploratory results show that the use of quantitative methods (*characteristic elements computation*, *repeated segments extraction*, *multiple co-occurrences*, etc.) in combination with bi-text topography offers new means for automatic description of lexical equivalences in translation corpora. 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:

alignment, bi-text topography, parallel text processing, textometric analysis, translation correspondences.

1. Textometric analysis of multilingual text corpora

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. Historians, lawyers, philologists are getting used to working with new computer tools currently available for exploring intertextual correspondences between related parts of multilingual texts. Computer scientists explore language resources obtained from such corpora in order to improve the quality of machine translation software and the efficiency of search engines for the Web. Finally, translation resources obtained from multilingual texts are successfully used in different fields of linguistic research, ranging from *comparative linguistics* to *lexicography*, from *computer-assisted translation* to *foreign language learning and teaching*, from *discourse analysis* to *computational linguistics*, etc.

Multilingual text corpora have proved to be an invaluable source of translation data for terminology banks and electronic dictionaries. In the past ten to fifteen years new corpus-based tools and software have been developed for automatic extraction of translation resources and cross-language information retrieval.

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.

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 important research issue in the field. 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 frequently used to identify potential word candidates for lexical alignment.

Recent developments have shown that quantitative methods used in *textometric analysis* open up new horizons for identifying translation correspondences in bilingual texts (Martinez and Zimina 2002), (Zimina 2004b), (Zimina 2005 *forthcoming*). 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 new tools for multilingual textometric exploration of translation corpora.

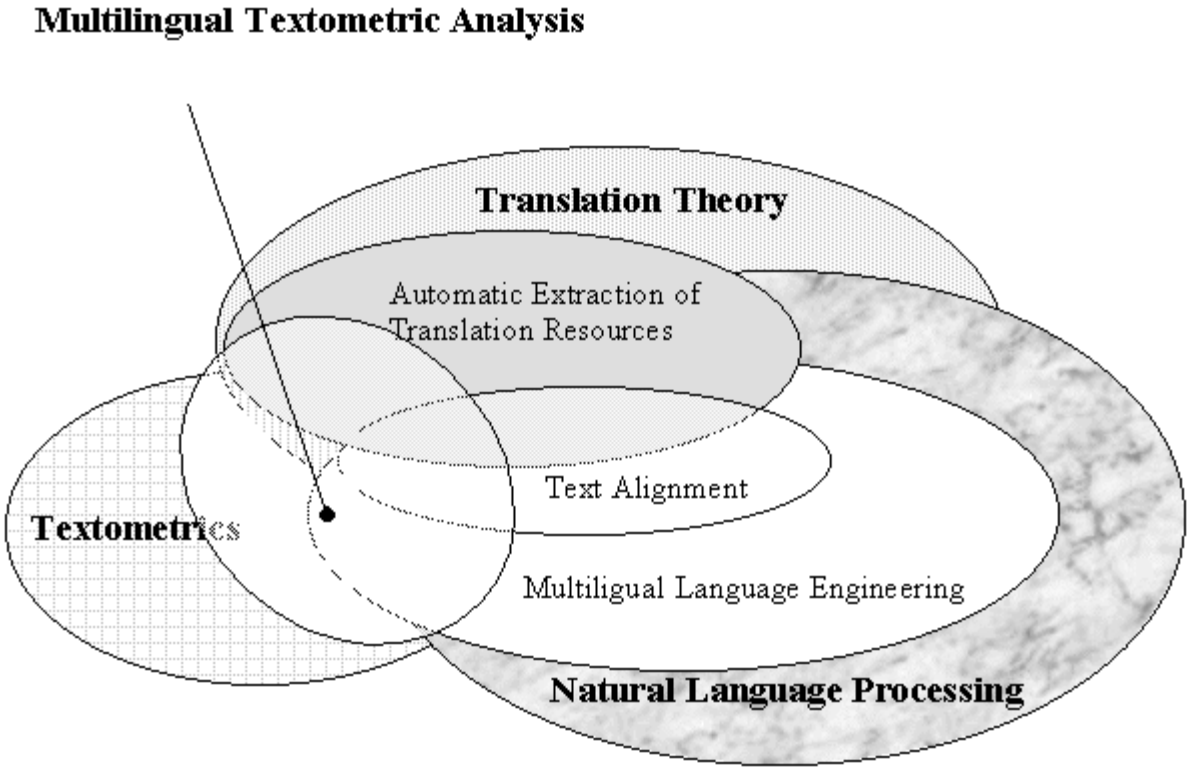

Figure 1: Scope of exploratory Multilingual Textometric Analysis

1.1 Choosing textual units

Multilingual textometric analysis is a new field of study bringing together the knowledge acquired in several related disciplines such as Translation Theory, Natural Language Processing (NLP) and Textometrics (see *Figure 1*).

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***) [1] allow the analyst not only to subdivide the text into graphical forms, but also to identify other types of textual units (see *Figure 2*):

- ***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.).

- ***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 the *Word-store*. This feature allows for the memorization of forms, segments, *Tgen(s)* for later use.

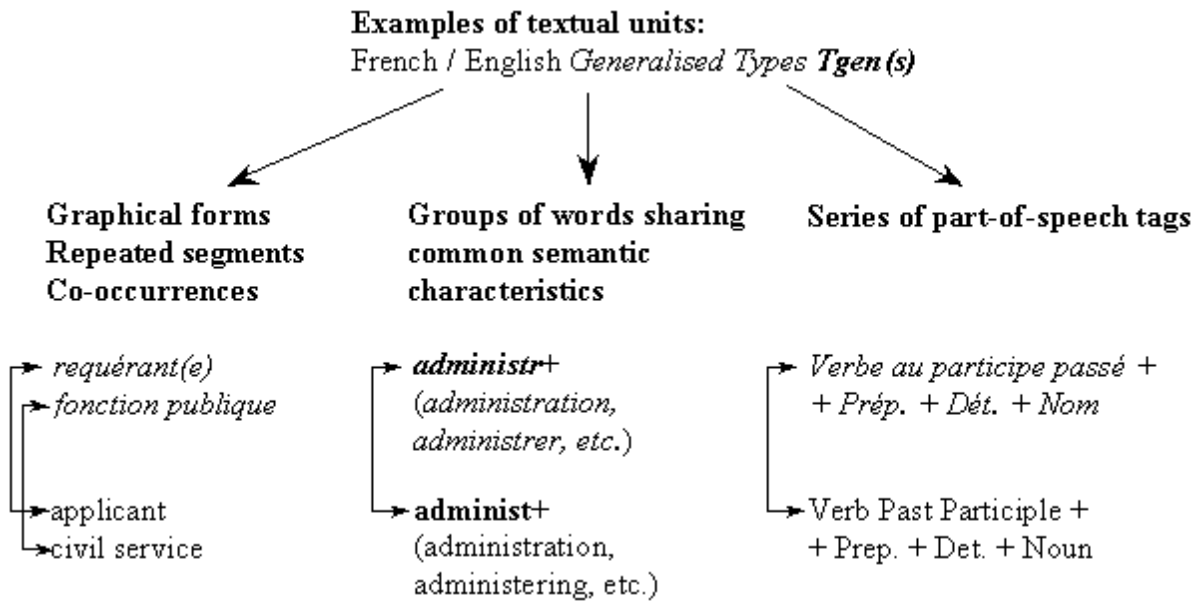

Figure 2: Examples of textual units *Tgen(s)*

1.2 Bilingual lexicon extraction

Exploratory results show that bilingual lexicon extraction from translation corpora can be sorted out by statistical study of distribution similarities between candidate terms representing mutual translation pairs (Fung and Church 1994), (Fung 2000). Different quantitative methods such as *hierarchical cluster analysis of graphical forms and repeated segments* of bilingual texts may allow an identification of translation correspondences on similar bases (Zimina 2000).

A list of one-to-one translation correspondences found in a bilingual lexicon is often limited to a given sample of corpus-based translation resources. Recent work has shown that this type of representation lacks flexibility when it comes to explore multiple translation correspondences between polysemous lexical units (Zimina 2004b). In this respect, the concept of *textometric browsing* based on *bi-text topography* offers new possibilities for the automatic description of multiple lexical equivalences in translation corpora.

1.3 Textometric browsing with the map of parallel sections

The concept of *textometric browsing* enables the user to move among the results produced by different quantitative methods and the original bi-text. The *map of parallel sections* allows for the visualization of the corpus cut into corresponding sections by raising one (or several) characters (e.g.: carriage return) 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 coloured.

In order to compare 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. The selected keys allow the user to compare corresponding textual fragments (sections, paragraphs, phrases, etc.).

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.). Existing textometric tools (such as *Lexico3* and *MkAlign*) [2] offer the possibility of promoting one or several delimiting characters to the rank of *section delimiters*. Such pre-coding allows for the study of the distribution of occurrences of any textual unit within the sections thus defined. [3]

Accordingly, *Figure 3* shows a fragment of the French/English parallel corpus ***Convention*** 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. [4]

Here are the explanations of elements used to codify the bi-text in the example on *Figure 3*:

- The key **text** is the code for the language (French or English).
- The paragraph character **§** marks the beginning of each aligned fragment (phrase) of the text.
- The character * identifies uppercase letters in the original document.

/.../ **<text="fr">§** du côté gibraltarien de la frontière, les fonctionnaires des douanes et de la police en service normal ne furent ni informés ni associés à la surveillance, au motif que cela impliquerait que l'information soit communiquée à un trop grand nombre de personnes.	/.../ **<text="en">§** on the *gibraltar side of the border, the customs officers and police normally on duty were not informed or involved in the surveillance on the basis that this would involve information being provided to an excessive number of people.
<text="fr">§ aucune mesure ne fut prise pour ralentir la file de voitures lors de leur entrée, ou pour examiner tous les passeports, car on craignait que cela puisse alerter les suspects.	**<text="en">§** no steps were taken to slow down the line of cars as they entered or to scrutinise all passports since it was felt that this might put the suspects on guard.
<text="fr">§ une équipe de surveillance distincte se trouvait cependant à la frontière et un groupe préposé à l'arrestation était posté dans le secteur de l'aéroport voisin.	**<text="en">§** there was, however, a separate surveillance team at the border and, in the area of the airfield nearby, an arrest group.
<texte="fr">§ le témoin *m, qui dirigeait une équipe de surveillance postée à la frontière, exprima sa déception au vu du manque apparent de coopération entre les divers groupes impliqués à *gibraltar, mais il comprit que les choses étaient ainsi organisées pour des questions de sécurité. /.../	**<text="en">§** witness *m who led a surveillance team at the frontier expressed disappointment at the apparent lack of co-operation between the various groups involved in *gibraltar but he understood that matters were arranged that way as a matter of security. /.../

Figure 3: Phrase-aligned French / English parallel corpus ***Convention*** (extract)

2. Extracting translation resources by textometric browsing

This section will describe a series of experiments that have been carried out in order to extract translation resources from the corpus ***Convention*** by means of textometric browsing. This approach allows the user to move among the results produced by different methods of textometric analysis and the original corpus. [5]

Certain procedures of textometric browsing, such as *bi-text topography*, which I will describe in the following, have not yet been included in the current version of *Lexico3*. These procedures will be available in the next version of *Lexico3*. The map of parallel sections is currently available within *MkAlign* editor.

2.1 Bi-text topography and text resonance

As we have shown in *section 1.1*, 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.

Statistical methods rely on measurements and counts based on objects resulting from identification of occurrences of textual units (forms, segments, generalised types) in the different parts of a text corpus. In bilingual corpora, it is convenient to identify corresponding parts of texts through *bi-text topography*.

Dragging textual unit(s) found in the dictionary of graphical forms (or in the list of repeated segments, the *Word-store*, etc.) onto the map of parallel sections, it is possible to produce a distribution of the selected textual unit(s) in different parts of the corpus. Colours on the map mark the sections containing at least one occurrence of the selected textual unit(s) (see *Figure 5*).

A corresponding set of sections in the other part of the bi-text in then selected through the process of *text resonance* (Lamalle and Salem 2002). *Characteristic elements computation* is used to discover the list of translation equivalents of the textual unit(s) used for initial topographic selection. The analysis of *characteristic elements* (in French: "spécificités") allows for an evaluation of the frequency of each of the textual units in each of the parts of the corpus (Lafon 1984).

The characteristic element diagnostics contains two indications (see *Figures 5-7*):

a) The sign (+ or –) indicating an over or under-use representation in the selected section(s) as compared to the entire corpus.

b) An exponent that indicates the degree of significance of the difference (an exponent equal to *x* means that the probability of a distribution difference more than or equal to the difference found was of the order 10^{-x}).

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.

2.2 Example: mapping multiple lexical correspondences in the corpus *Convention*

This section illustrates the principles of interactive textometric browsing in the corpus ***Convention*** through bi-text mapping of translation correspondences of the French term "fonctionnaires" (civil servants).

Step One (see *Figure 5*):

- The user selects the *Tgen* (from the dictionary of graphical forms, the list of repeated segments, the *Word-store*, etc.) and drags it to the map of parallel sections. For example, the "drag-and-drop" of the form "fonctionnaires" (F=49) onto the map, enables to colour automatically the sections in French, containing at least one occurrence of "fonctionnaires". It is possible to set two *probability thresholds*, producing more or less dark section colouring. For a simultaneous representation of two *Tgen(s)*, this process can be reiterated (with change of the mapping colour).

- Following the process of text resonance (see section 2.1), the activated selection in one of the corpus parts automatically produces a parallel selection of the equivalent sections in the other part of the corpus.

- The analysis of characteristic elements allows for an evaluation of the frequency of each of the textual units in each set of selected sections activated on the map. The results of characteristic elements computation are displayed in separate windows (one for the French part of the corpus and the other for the English one). The number of sections appears at the top of the window; the results can be saved using a button "Section" at the bottom of the window.

- In each window displaying characteristic elements, the first column presents the characteristic units in descending order. The next two columns show, respectively, the frequency of the textual unit in the entire corpus (*F*) and the sub-frequency of this unit in the selected set of sections (*f*). The *positive* and *negative* check buttons in the tab of characteristic elements enable to inverse the order of presentation of the list (on *Figure 5*, the list starts with positive characteristic units).

- The parallel lists of characteristic units show that the English form *servants* (f=31) and the repeated segment *civil servants* (f=29) come on the top of the list. We can then consider that the French term "fonctionnaires" and the English units *servants* and *civil servants* might be translation correspondences.

- Immediate context analysis enables to validate translation correspondences "fonctionnaires" – *servants / civil servants*. The context of related parallel sections is visualized by clicking on the squares representing these sections on the map.

Step Two (see *Figure 6*):

- The total frequency of the French term "fonctionnaires" (F=49) is higher than the sub-frequency of *servants* (f=31) in the corresponding sections of the English part of the corpus. This difference indicates that there exist other translations of "fonctionnaires" within the corpus.

- The “drag-and-drop” of the form *servants* (F=50) onto the map, enables to activate the sections in English, containing at least one occurrence of this form.

- Asymmetric colouring of the bi-text map reveals the sections of the English part of the corpus in which the term “fonctionnaires” is not translated by *servants*. Characteristic elements computation in this last set of sections enables to discover two other translations of “fonctionnaires”. The English units *officers* (f=10) and *officials* (f=7) come on the top of the list of characteristic units of the activated selection.

- As in **Step One**, immediate context analysis enables to validate translation correspondences “fonctionnaires” – *officers* and “fonctionnaires” – *officials*. The context of sections is visualized by clicking on the squares representing these sections on the map. By using the buttons (in the shape of hands), the user might go back or move forward to the next or preceding section or to the next/preceding occurrence of the selected *Tgen*. The mapping zone can be re-initialised at any time, after having recorded a graph in a report. These functions are already available in the present version of *Lexico3*.

Step Three (see *Figure 7*):

- The total frequency of the French term “fonctionnaires” (F=49) is slightly higher than the joined sub-frequency of the corresponding textual units in the English set of sections identified through textometric browsing (49 > 48). The difference shows that there exists at least one context of “fonctionnaires” for which no equivalence has been discovered in the previous exploration.

- The missing singular context is particularly difficult to grasp on quantitative bases. Nevertheless, it is possible to draw the attention of the user to the corresponding pair of sections where the French form “fonctionnaires” is present and the English *Tgen* composed of *servants*, *officers, officials* is absent.

- The missing context is visible on the map. 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 the missing translation of “fonctionnaires” (see *Figure 7*):

<text="fr">§ aux termes de /…/ la loi-cadre sur les **fonctionnaires** des länder /…/ seul peut être nommé fonctionnaire celui qui "offre la garantie qu’il prendra constamment fait et cause pour le régime fondamental libéral et démocratique au sens de la loi fondamentale."

<text="en">§ by virtue of /…/ the **civil service** (general principles) act for the länder, appointments to the civil service are subject to the requirement that the persons concerned "satisfy the authorities that they will at all times uphold the free democratic constitutional system within the meaning of the basic law".

3. Conclusions and upcoming developments

In this article, my goal was to present new tools for cross-language exploration of translation corpora. These tools are entirely based on quantitative methods of textometric analysis. The suggested approach offers new means for automatic description of lexical equivalences in translation corpora. It can be used to detect multiple translation correspondences of polysemous lexical units.

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.

Certain procedures for textometric analysis of multilingual translation corpora have not yet been included in the currently distributed version of textometric software *Lexico3*. New modules (such as *MkAlign*) are currently being developed within SYLED-CLA^{2}T Centre of Textometrics (Sorbonne University – Paris 3).

One of the upcoming developments will concern the inclusion of new procedures allowing for the identification of *networks of co-occurrences* in a text. A research methodology suggested by Martinez (2003) offers new means for parallel exploration of lexical constellations found in a bi-text (Martinez and Zimina 2002). The identification of parallel networks of co-occurrences in translation corpora might be conducted in conjunction with bi-text topography. It will then be possible to study lexical attractions in the neighbourhoods of mutual translation pairs discovered through textometric browsing, such as:

- the set of occurrences of the French form “fonctionnaires” translated in English by *servants* (number of contexts = 29).

- the set of occurrences of the French form “fonctionnaires” translated in English by *officers* (number of contexts = 9).

- the set of occurrences of the French form “fonctionnaires” translated in English by *officials* (number of contexts = 7). See section 2 *infra*.

On *Figure 8*, the computation of parallel networks of co-occurrences is carried out separately in each corpus part for each set of occurrences forming mutual translation pairs. The results of this computation provide a more precise context-based description of corresponding lexical items in translation corpora.

Parallel lexical networks represent a particularly valuable source of translation resources and make possible a thorough study of contextual differences in translation. Future work will have to produce more specific results and allow for further advances in this direction.

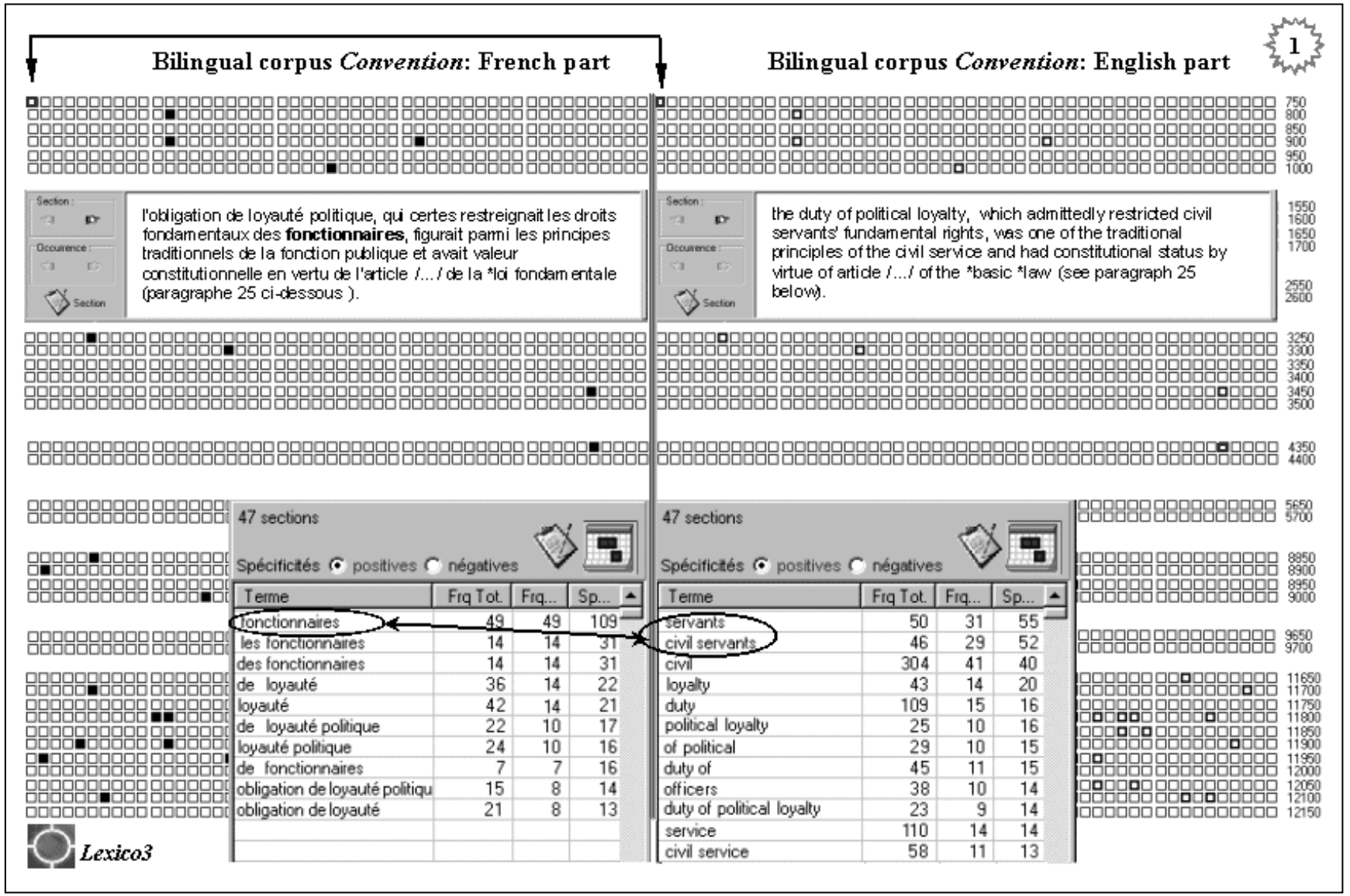

Figure 5: Map of parallel sections showing the distribution of the French form “fonctionnaires”

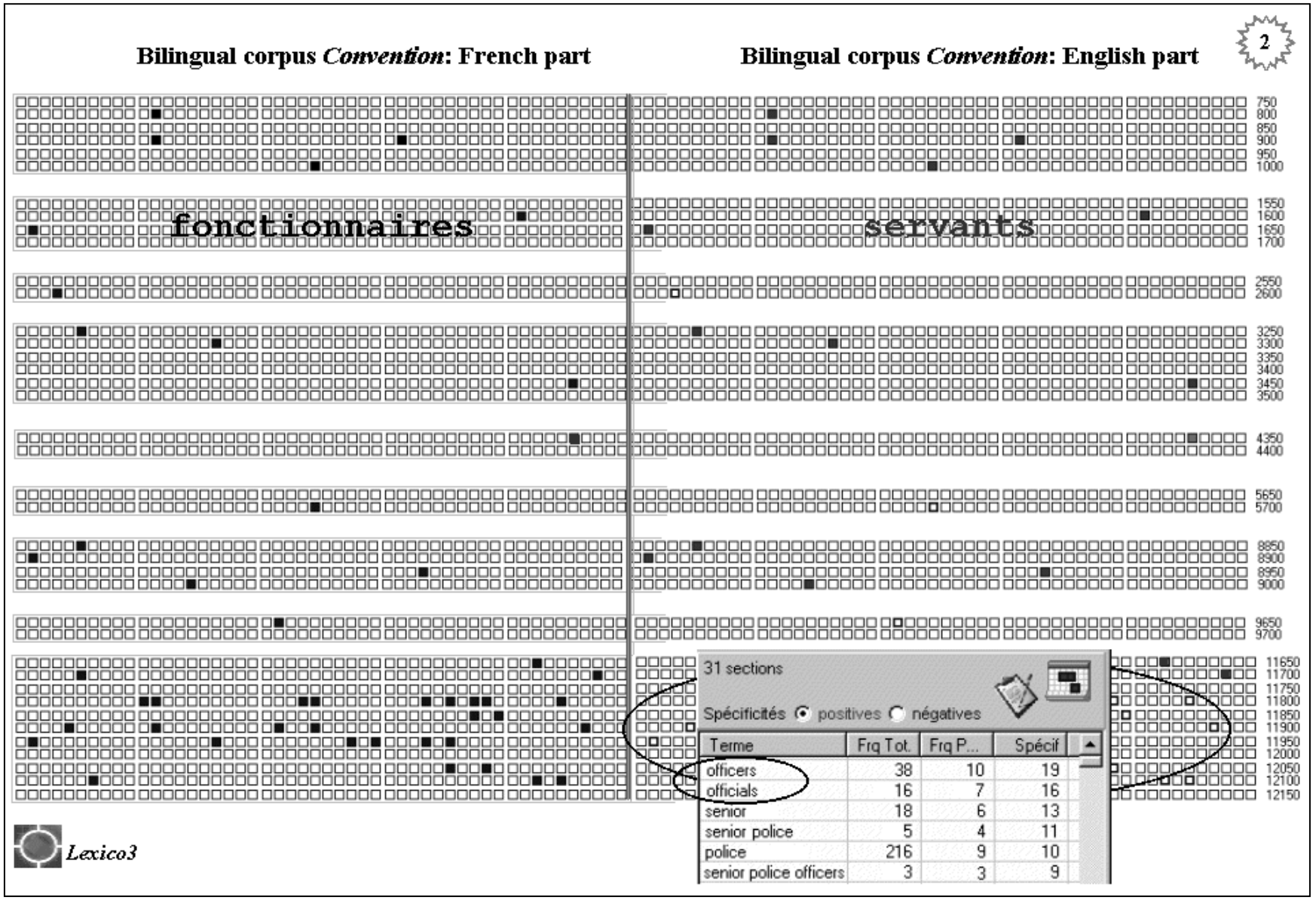

Figure 6: Distributions of the partially corresponding forms “fonctionnaires” / *servants*

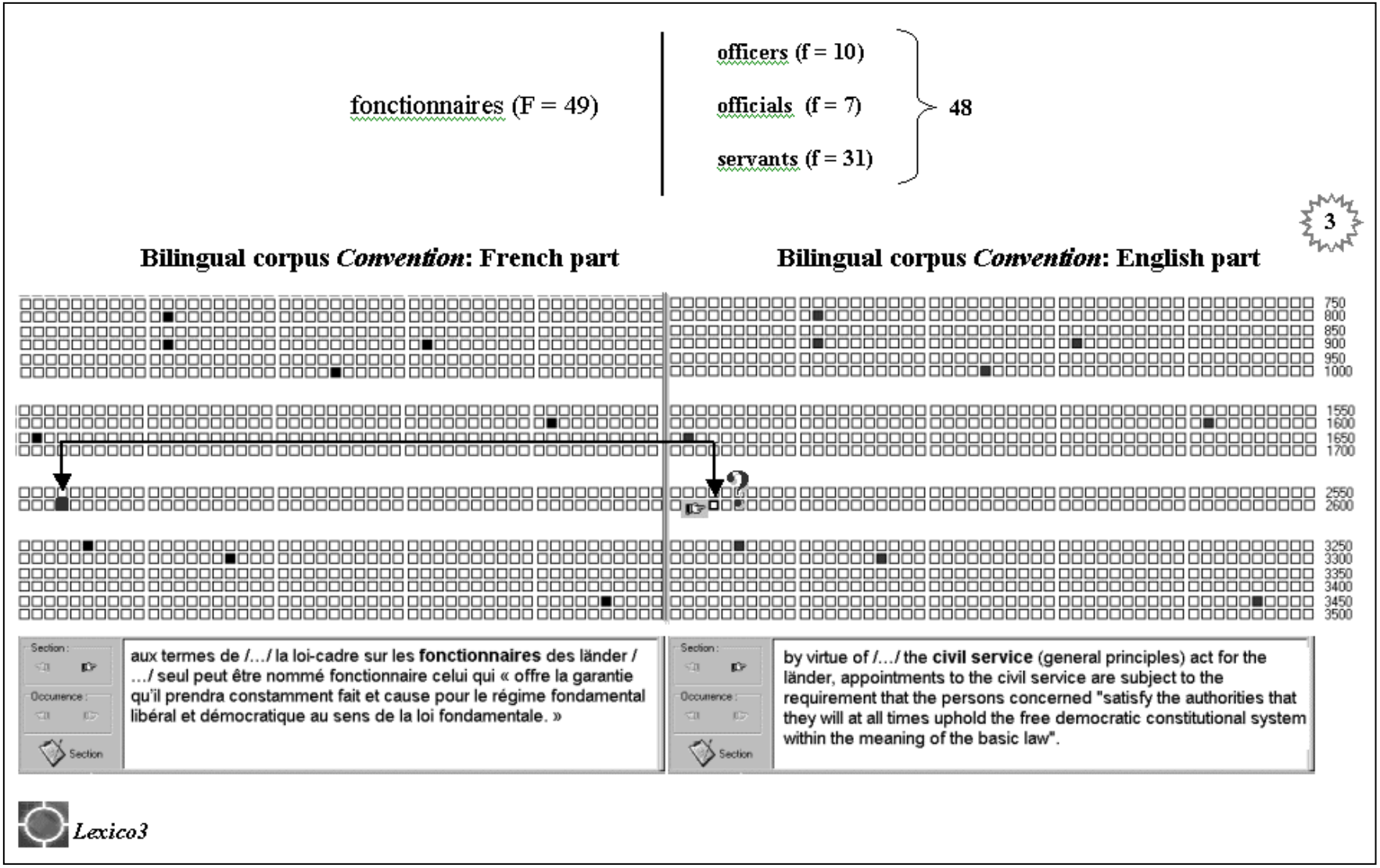

Figure 7: Localization of non-corresponding zones

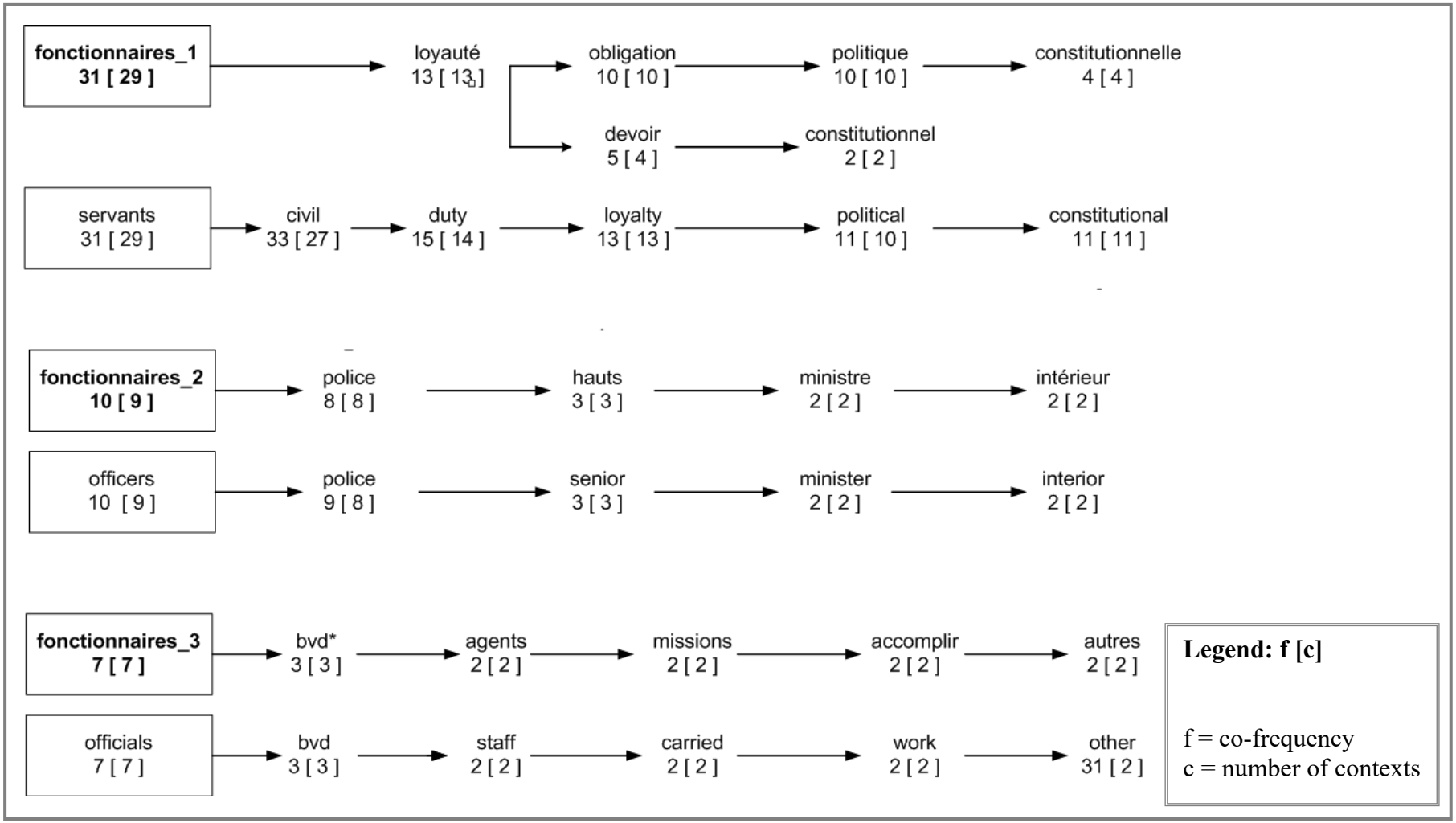

Figure 8: Lexical attractions in the neighbourhoods of mutual translation pairs discovered through textometric browsing (see *Figures 5-7*): "fonctionnaires" / *servants*, "fonctionnaires" / *officers*, "fonctionnaires" / *officials*

Notes

1. For further information on *Lexico3 Textometric toolbox* see the website of the SYLED-CLA^{2}T team at Paris Sorbonne University – Paris 3: http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/index-gb.htm

2. *MkAlign* is currently developed by Serge Fleury within SYLED-CLA^{2}T team at Paris Sorbonne University – Paris 3. Concerning corrections, updates, see the documents and manuals on the page http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm

3. In *comparable corpora*, the insertion of section delimiters is a much more complex task. Thematic proximity of related parts sorted out by statistical study of context word similarity between candidate translation pairs is one of the ways to deal with this issue (Fung 2000).

4. The corpus ***Convention*** was used in a variety of methodological studies within the research centre SYLED-CLA^{2}T (Paris Sorbonne University – Paris 3). Cf. References *infra*.

5. *Lexico3* software is based on object-oriented program architecture. The different interactive modules of this toolbox are able to exchange complex data items (forms, repeated segments, etc.). For instance, it is possible to send to the concordance module, or to any other modules, units established in the module of repeated segment, lists of forms and segments established in the characteristic elements modules, etc. As a result, veritable textometric browsing becomes possible.

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).

Lafon, P. (1984) *Dépouillements et statistiques en lexicométrie* (Genève-Paris : Slatkine-Champion).

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

Articles in Book:

Fung, P. (2000) A Statistical View on bilingual lexicon extraction: From Parallel corpora to non-parallel corpora, in J. Véronis (ed.) *Parallel Text Processing*: *Alignment and use of translation corpora* (Dordrecht: Kluwer Academic Publishers), 219-236.

Articles in Conference Proceedings:

Fung, P. and Church, K. (1994) K-vec: A New Approach for Aligning Parallel Texts. *COLING'94*, *Kyoto*, *1994*, 1096-1104.

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.

Martinez, W. and Zimina, M. (2002) Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. *JADT'02*, *Saint-Malo*, *2002*, 495-506.

Zimina M. (2000) Alignement de textes bilingues par classification ascendante hiérarchique. *JADT'00*, *Lausanne*, *2000*, 171-178.

Zimina M. (2004a) L'alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. *JADT'04*, *Louvain-la-Neuve*, *2004*, 1195-1202.

Zimina M. (2005 *forthcoming*) Exploration textométrique de corpus de traduction. *META'50*, *Montreal*, *2005*.

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://ed268-p3.no-ip.org/student/stmz

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