



HAL
open science

Première approche textométrique de l'analyse contrastive du corpus trilingue anglais-français-allemand GUIDES

William Martinez, Maria Zimina

► **To cite this version:**

William Martinez, Maria Zimina. Première approche textométrique de l'analyse contrastive du corpus trilingue anglais-français-allemand GUIDES. 2007. hal-01371290

HAL Id: hal-01371290

<https://hal.science/hal-01371290>

Preprint submitted on 25 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Fédération de recherche des Sciences de la Ville

PROJET *GUIDE DES GUIDES*

**PREMIERE APPROCHE TEXTOMETRIQUE DE L'ANALYSE CONTRASTIVE DU
CORPUS TRILINGUE ANGLAIS-FRANÇAIS-ALLEMAND
*GUIDES***

William MARTINEZ, Maria ZIMINA-POIROT

SYLED-CLA2T

Université de la Sorbonne nouvelle – Paris 3

SOMMAIRE

1. Textométrie et corpus multilingues.....	4
1.1. Introduction	4
1.2. La spécificité de l'approche textométrique	4
1.3. Textométrie multilingue	5
1.4. Navigation textométrique intertextuelle.....	5
2. Première approche textométrique de l'échantillon textuel issu du corpus anglais-français-allemand <i>GUIDES</i>	6
2.1. Présentation du corpus	6
2.2. Préparation du corpus pour l'analyse statistique.....	6
Numérisation des documents	6
Codage des données	7
Création des partitions.....	8
2.3. Prise en compte des formats d'échange	9
2.4. Analyse contrastive des volets trilingues.....	9
Repérages des unités textuelles complexes.....	9
Les statistiques par partie	9
Le calcul des spécificités lexicales.....	10
L'analyse factorielle des correspondances	12
3. Conclusions et perspectives	13
ANNEXES	14
Bibliographie.....	25
Sites Web :	25
Ouvrages :	25
Articles de revue :	25
Articles dans des actes de congrès internationaux :	26

Ouvrages techniques : 27

1. Textométrie et corpus multilingues

1.1. Introduction

Les méthodes quantitatives convoquées dans les explorations de l'échantillon numérisé des textes du corpus *GUIDES* (cf. *Section 2*) sont issues des recherches menées depuis plusieurs années au sein du *Centre de Lexicométrie et d'Analyse Automatique des Textes (SYLED-CLA2T)* de l'Université de la Sorbonne nouvelle – Paris 3).

Les préoccupations de cette équipe comprennent à la fois le développement de la méthodologie de l'étude quantitative des textes et la réalisation de logiciels d'analyse textométrique pour diverses communautés de chercheurs concernés par le domaine des études textuelles.

Les axes de recherches développés dans les travaux des membres du *SYLED-CLA2T* couvrent plusieurs problématiques :

- Développement de méthodes consacrées à l'étude de corpus textuels étendus.
- Prise en compte de catégorisations (morphosyntaxiques, sémantiques, etc.) dans les traitements lexicométriques.
- Alignement de textes multilingues.
- Etude de la sémantique des unités lexicales.
- Réflexions sur les relations entre texte et contexte et sur l'émergence de normes dans l'activité langagière.
- Intertextualité et mondes discursifs.

1.2. La spécificité de l'approche textométrique

Dans ce qui suit, on utilisera le terme *textométrie* pour se référer à l'ensemble des méthodes quantitatives permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes [Lebart et Salem, 1994].

L'analyse textométrique s'appuie sur des comptages réalisés à partir du repérage des occurrences d'unités lexicales dans les différentes parties d'un corpus. Globalement, elle regroupe trois types de méthodes [Salem, 1987] :

- Les méthodes documentaires qui opèrent une simple réorganisation de la surface textuelle.

- Les méthodes qui opèrent, pour des textes pris isolément, des comptages et des calculs d'indices statistiques.
- Les méthodes statistiques « contrastives » qui produisent des résultats portant sur le vocabulaire de chacun des textes par rapport à l'ensemble des textes réunis dans un corpus à des fins de comparaison.

La spécificité de la démarche textométrique réside dans le statut privilégié conféré aux données textuelles. Cette approche ne prend que très peu d'appui sur des savoirs *a priori* (linguistiques, pragmatiques, etc.). La textométrie porte une attention particulière au *texte* dans son intégrité et permet de prendre en considération sa structure ordonnée dont les régularités se manifeste au plan syntagmatique et paradigmatic.

1.3. Textométrie multilingue

Dans un contexte multilingue, la textométrie offre plusieurs perspectives de recherche pour de multiples dimensions d'analyse de corpus dans des langues différentes (l'alignement d'unités textuelles en correspondances de traduction, l'acquisition de ressources traductionnelles, l'extraction de lexiques terminologiques, la synthèse de l'information, etc.).

Les questions de repérage/mise en relation de structures textuelles spécifiques présentes au fil des textes réunis à des fins de comparaisons, ont déjà fait l'objet de plusieurs études textométriques : [Zimina, 2000] ; [Martinez et Zimina, 2002] ; [Zimina, 2004]. Appuyés par de nouveaux développements logiciels, ces recherches sont en train de s'étendre à l'étude des phénomènes de résonances textuelles et intertextualité [Lamalle et Salem, 2002] ; [Salem, 2004 ; 2006] ; [Zimina, 2005] ; [Fleury et Zimina, 2007].

1.4. Navigation textométrique intertextuelle

Les comparaisons de structures récurrentes dans les corpus qui présentent des similitudes du point de vue de leurs contextes de production (*corpus parallèles* ou *comparables*, *séries textuelles chronologiques*, etc.), attire l'attention sur la circulation d'unités textuelles étendues qui reflètent l'existence de concepts partagés par la communauté produisant ces textes [Salem, 1987]. Ces formations discursives communes pourraient être repérées à travers l'analyse des répétitions segmentales partagées par un groupe de textes [Salem, 2006] ou encore par l'analyse de structures lexicales complexes qui se réalisent, par exemple, dans les réseaux de concurrences chaînées [Martinez, 2000] ; [Martinez et Zimina, 2002] ; [Leblanc et Martinez, 2005].

Les fonctionnalités d'exploration textométrique intégrées aux logiciels *Lexico3* [Lamalle *et al.*, 2004], *COOCS* [Martinez, 2005], *MkAlign* [Fleury et Zimina, 2006] développés par les membres du *SYLED-CLA2T*, offrent un large panel d'outils pour les comparaisons intertextuelles. Les procédures informatiques intégrées au sein de ses outils interactifs permettent d'envisager une étude approfondie des phénomènes textuels complexes, tels que la variation des traductions d'un terme en fonction des contextes, le repérage thématique d'équivalences traductionnelles, la découverte des constellations lexicales parallèles, etc.

2. Première approche textométrique de l'échantillon textuel issu du corpus anglais-français-allemand GUIDES

2.1. Présentation du corpus

- *Murray's Handbook for Travellers on the Continent. Part 1. Being a guide to Holland, Belgium* (1875).
- *Baedeker « Belgien und Holland »* (1873) [français et allemand].

2.2. Préparation du corpus pour l'analyse statistique

La préparation du corpus textuel représente une étape essentielle à la réussite des procédures analytiques subséquentes. Il faut en amont du traitement statistique du texte définir des règles et des normes qui garantissent la rapidité et la régularité d'exécution des programmes informatiques d'analyse lexicale. La transformation des guides en un corpus informatique formaté pour le traitement textométrique compte deux phases principales :

- La *numérisation* des documents qui les transforme de papier en fichier informatique.
- Le *précodage* du fichier texte suivant des normes de transcription spécifiques pour l'adapter aux dépouillements quantitatifs.

Numérisation des documents

La transformation des données commence par la saisie optique de chaque page au moyen d'un scanner dont les clichés sont ensuite soumis à un programme de reconnaissance optique de caractères qui traduit les graphismes en mots (cf. *ANNEXES* : **Figure 1**). Les erreurs de reconnaissance sont possibles et il est alors nécessaire d'éditer les textes obtenus afin d'effectuer une correction manuelle. On retire à cette occasion les éléments que l'on estime

sans intérêt pour l'analyse lexicale (signatures, illustrations, etc.) et on uniformise le format d'écriture (éventuellement standardisation des noms propres, conversion des majuscules en minuscules, etc.). A l'issue de ces diverses opérations de transcodage on obtient un texte 'en machine' exploitable par les programmes d'analyse textométrique.

Précodage des données

L'objectif du précodage des données est de rendre possible la lecture informatique du texte à différents niveaux de son organisation structurelle suivant l'enchaînement des éléments lexicaux et contextuels qui le composent. En effet, la particularité de l'approche textométrique est d'aborder le texte au niveau de son support physique en laissant le choix de le traiter mot à mot, phrase par phrase ou encore suivant d'autres niveaux de précision.

La lecture informatique du corpus doit permettre d'identifier d'une part les *unités lexicales* et d'autre part les *unités contextuelles*. Pour rendre possible cette double lecture du texte il est nécessaire de transformer ce dernier par l'introduction de deux types de marqueurs : les *signes délimiteurs* et les *clefs de balisage*.

Unités lexicales

Pour l'étude du corpus **GUIDES** nous avons retenu la *forme graphique* comme unité de dépouillement et de mesure systématique. La détection automatique des formes graphiques s'effectue à partir de la division fondamentale des caractères composant le texte en deux catégories : les caractères délimiteurs et les caractères non-délimiteurs. Outre l'espace, considéré comme caractère séparateur par défaut, la liste de délimiteurs en vigueur dans le cadre de l'analyse du corpus **GUIDES** comprend les marques habituelles de ponctuation forte et faible, telles que [. ! ? , ; : ...].

En plus de l'analyse des formes graphiques, l'approche textométrique permet d'étudier la répartition dans les textes d'unités lexicales complexes : *lemmes*, *segments répétés*, *cooccurrences*, *types généralisés* (cf. *Section 2*) [Lamalle et Salem, 2002]. Cependant, une segmentation en formes graphiques constitue une première étape nécessaire qui permet tout à la fois :

- d'obtenir une première estimation des principales caractéristiques lexicométriques du corpus : nombre d'occurrences, de formes, d'hapax, fréquence maximale (cf. **ANNEXES : Figure 4**) ;
- de réaliser les premières typologies sur les parties du corpus (cf. **ANNEXES : Figures**

10-12).

- de localiser les erreurs qui subsistent après les premières corrections.

Unités contextuelles

Le corpus **GUIDES** propose un deuxième niveau de segmentation avec le découpage contextuel en phrases et en paragraphes. On pourra retenir pour le marquage méta-textuel des paragraphes le signe ‘§’ et pour celui des phrases un signe tel que ‘\$’ pour éviter toute ambiguïté avec les points d’abréviation particulièrement fréquents dans le texte. L’extrait suivant illustre ce codage :

```
§
médecins : dr. verhaeghe, dr. van mullem et dr. cosyn.
§
bains$ cartes de bains, 60 cent. la pièce$ les cabines sont conduites à l'eau non pas par
des chevaux mais à bras d'hommes$ on donne lors du départ quelques francs de pourboire pour
le séchage du linge et les chaises sur la plage$
tentes sur la plage pour se mettre à l'abri contre le vent et le soleil (non contre la
pluie), 1 fr. par jour$
costumes de bain de tricot pour hommes, 5 fr. ; hauts, avec surtout, 8 fr.$
§
```

Une fois définies ces unités contextuelles, il devient possible d’opérer des analyses dans chaque ‘fenêtre de mots’ et s’intéresser par exemple à l’univers lexical d’une forme vedette pour ensuite extraire du texte une phrase ou un paragraphe illustrant une attraction entre mots (cf. **Section 2.4** : L’analyse des cooccurrences).

Création des partitions

On rencontre dans un texte un certain nombre de délimitations inhérentes telles que des délimitations chronologiques, des délimitations marquant un changement d’auteur ou encore indiquant la séquence des parties. Au cours de l’étude textométrique, on compare les fréquences du vocabulaire dans une partie par rapport à une autre en modifiant les découpages du corpus. Pour que ces comparaisons s’effectuent efficacement, le texte doit comporter des balises informatives indiquant ces délimitations logiques sous une forme codée compréhensible par la machine.

Le corpus **GUIDES** peut faire l’objet d’un codage fondé sur plusieurs de ses coordonnées de production : *ville, page, année*, etc. Chaque clef découpe le corpus en un certain nombre de sections et permettent d’affiner le calcul statistique pour les analyses portant sur un bloc isolé du texte. Toutes les clefs se présentant sous la forme *<type=contenu>*, on codera par exemple un texte décrivant la ville de Liège par une clef telle que **<ville=Liège>**.

Etant donné que les documents font tous l’objet d’un codage fondé sur leurs coordonnées de production, il devient possible d’effectuer des requêtes lexicométriques sur le corpus d’une

manière ciblée. Par exemple : *Quels sont les segments répétés caractéristiques des discours de l'année 1877 ? Quelles formes sont sur-employées lorsque l'on décrit la ville de Louvain, ou sous-employées lorsque l'on décrit les villes du Sud ? Ou encore, quels sont les contextes où l'on évoque les mots « place » et « hotel » ?*

2.3. Prise en compte des formats d'échange

Le coût important imposé par la saisie des textes sur un support lisible par un ordinateur incite à réfléchir à la constitution d'un corpus annoté permettant à plusieurs équipes de chercheurs d'effectuer leurs propres expériences à partir des mêmes textes saisis selon des standards très généraux [Habert *et al.*, 1997]. Dans la période actuelle, plusieurs normes tendent à s'imposer pour le stockage des corpus textuels. Le format *XML*, par exemple, présente l'intérêt d'être un langage informatique de balisage générique dont l'objectif est de faciliter l'échange automatisé de contenus entre systèmes d'informations hétérogènes - notamment sur Internet (cf. *ANNEXES : Figure 3*).

2.4. Analyse contrastive des volets trilingues

Repérages d'unités textuelles complexes

Comme c'est montré sur la figure 5 (cf. *ANNEXES*), les fonctionnalités des logiciels de textométrie permettent de constituer des types rassemblant les occurrences de formes graphiques différentes liées par une propriété commune. On peut ainsi, moyennant certaines précautions, rassembler le pluriel et le singulier d'une même forme, les flexions d'un même verbe, des formes qui possèdent un lien sémantique, etc. Les formes ainsi regroupées peuvent ensuite être manipulées comme des entités uniques (*Tgens*). Les *Tgens* sont constitués en faisant une recherche sur plusieurs formes, en introduisant des chaînes de caractères qui correspondent à des préfixes, des suffixes ou des suites de caractères graphiques.

Au-delà des dépouillements en formes graphiques, les méthodes textométriques permettent de recenser les figements lexicaux par l'analyse des *segments répétés* (suite de formes graphiques identiques attestées plusieurs fois dans le texte, cf. *ANNEXES : Tableau 6*) et le calcul des réseaux de *cooccurrences* (couples de formes présentes dans les mêmes contextes : phrase, sections, etc.) [Salem, 1987, 2006] ; [Martinez 2000, 2005].

Les statistiques par partie

L'analyse du corpus peut s'appuyer sur la mise en opposition de certaines parties par rapport à d'autres, lorsqu'il s'agit, par exemple, de comparer les variations de vocabulaire utilisé dans

les descriptions des villes ou espaces urbains.

On peut ainsi étudier la ventilation des unités textuelles choisies dans les différentes parties du corpus **GUIDES**. Sur la figure 7 (cf. *ANNEXES*), ces parties correspondent aux villes décrites dans l'échantillon du corpus. Selon notre hypothèse, les concentrations des unités textuelles utilisées dans les différentes parties pourraient fournir des indices sur les éléments textuels qui se correspondent d'une ville à l'autre ou d'une langue à l'autre. Les divergences sont toutes aussi intéressantes. Par exemple, les guides en anglais comportent plusieurs mentions correspondant au système bancaire. Cette information est absente dans les guides en français et en allemand (cf. *ANNEXES* : **Figure 7**).

Au-delà des partitions logiques repérables dans le texte (situation géographique des villes, accès et transports urbains, bâtiments publics), celui-ci contient des marques de rupture qui constituent sa respiration (phrases, paragraphes, etc.). Comme nous l'avons déjà signalé en décrivant les principes du précodage des données, il est possible de promouvoir un caractère délimiteur (ou plusieurs d'entre eux) au rang de délimiteur de section. Cette opération permet d'étudier ensuite la répartition des occurrences d'une unité textuelle quelconque parmi les sections ainsi constituées. Par exemple, sur la figure 8 (cf. *ANNEXES*), on matérialise les sections (unités contextuelles égales aux phrases du volet anglais du corpus **GUIDES**) correspondant aux ventilations des segments *the most* et *the largest*. L'affichage de la partition « ville » permet de naviguer dans le texte des guides afin de cerner les traits saillants de chaque agglomération.

Le calcul des spécificités lexicales

A partir des informations issues des dépouillements quantitatifs en formes graphiques, on peut déterminer les formes les plus caractéristiques de chaque sous-partie déclarée du corpus (déterminées d'après les clefs de partition, cf. *ANNEXES* : **Figure 2**). Le *modèle statistique hypergéométrique* [Lebart et Salem, 1994] permet de calculer un *indice de spécificité* pour chaque forme dans chaque sous-partie du texte en distinguant les formes qui sont plus employées que ne le prévoit la probabilité et celles qui sont sous-employées. On dresse ainsi un tableau résumant les mots qui caractérisent les descriptions de chaque ville soit parce qu'ils sont très souvent employés soit très rarement.

Les spécificités positives et négatives montrent comment on évoque pour chaque ville des thèmes différents : (cf. *ANNEXES* : **Figure 9**) :

- **Blankenberghe** : ses qualités de station balnéaire (*plage, bain, mer*) et ses tarifs (*fr*), au détriment de la *ville* même.
- **Heyst** : son industrie au capital de plusieurs *millions* et les *villages* voisins de quelques *kil*.
- **Liège** : illustrée dans le détail par des plans (*pl*), la ville se caractérise par la *meuse* et son incontournable *théâtre*
- **Louvain et Tournai** : ces villes se distinguent respectivement par leur *chapelle* et leur *transept* qui monopolisent la description touristique au détriment de considérations plus mercantiles (*fr*).

Une fois ces spécificités mises en évidence par la statistique on peut rechercher leurs contextes de réalisation. Par exemple autour du pôle *théâtre* on s'aperçoit que dans les descriptions de la ville de Liège l'édifice sert le plus souvent (8 occurrences sur 10) de référence pour situer d'autres l'établissement. On le vérifie dans les concordances suivantes :

, rue de l'harmonie, 7, près du	théâtre , très-bon ; belle vue du haut de la m
hamal, 6, tous deux derrière le	théâtre \$ - hôt. de la pommelette (pl. e),
drale, 17 ; d'Allemagne, place du	théâtre , 8; schiller (pl. d), même place (
; deux fontaines (pl. h), près du	théâtre , à g. en montant ; café de la renais
; café du diva, tous deux près du	théâtre ; café de la renaissance, passage lem
ière, s'étend jusqu'à la place du	théâtre , en faisant une grande courbe. le qua
\$ sur la grande et belle place du	théâtre est le théâtre royal (pl. 43), con
r dans les flots \$ de la place du	théâtre partent, au s.- o., plusieurs des ru

L'analyse des cooccurrences

Le modèle hypergéométrique employé pour déterminer les formes spécifiques positives et négatives d'une sous-partie du corpus permet également d'opérer un calcul des affinités lexicales autour d'un pôle lexical donné. Le calcul des *poly-cooccurrences* livre une liste des attractions lexicales autour d'une forme dans une unité contextuelle [Martinez 2000, 2005]. Muni des informations rapportées par l'exploration cooccurrentielle on peut extraire du texte les contextes de réalisation de ces poly-cooccurrences.

Par exemple, on parvient à recenser dans le volet allemand du corpus l'attraction lexicale *stadt* → *der* → *und* → *die* → *maas* attestée à trois reprises dans des contextes différents :

1 - Contexte n°1362	(56 formes dont 10 vedettes)	Densité info.=0.18
die alte thurmreiche stadt an einen berg sich anlehnend, der bis zur höhe mit häusern bedeckt ist, die breite maas (meuse), welche hier die ourthe aufnimmt, die vielen schornsteine, das beste zeugnis für die grossartige gewerbthätigkeit der bewohner, das fruchtbare und durch seine gemüsebau berühmte thal, das alles bietet ein bild von nicht gewöhnlicher landschaftlicher schönheit.		
2 - Contexte n°1374	(17 formes dont 6 vedettes)	Densité info.=0.35
viele sind in solcher nähe der stadt, dass die grubengänge unter den strassen und der maas liegen.		
3 - Contexte n°1605	(36 formes dont 9 vedettes)	Densité info.=0.25
über die grosse thurmreiche stadt, die gewerbreichen thäler der maas, der ourthe und vesdre; aufwärts begrenzen die ardenner gebirge den blick, abwärts reicht er bis zum petersberg bei mastricht und dann in das limburgische flachland hinein.		

On peut préciser ces résultats en introduisant dans le calcul une partition du corpus telle que celle de la clef « ville ». Le calcul des cooccurrences rapporte alors les associations lexicales dans chaque sous-partie du corpus et permet de dresser une évolution de l'univers lexical d'un pôle au fil du texte.

Par exemple, on discerne, autour de la forme française *ville* (64 occurrences), une constellation différente d'unités lexicales (noms, adjectifs, etc.) pour les villes de **Blankenberghe** (*moins, chère, fr, exemple, paie, vie*); **Heyst** (*habitants, bains, station, aspect, industriel*); **Liège** (*meuse, tout, entre*); **Tournai** (*rue, bâtiments, cygne*). Dans la description de Blankenberghe c'est l'aspect économique qui prime, alors qu'à Heyst, on souligne l'histoire industrielle de la ville et la qualité de ses structures d'accueil, bains et station. A Liège, la Meuse sert de point de référence tandis qu'à Tournai c'est la rue du cygne qui permet de repérer les établissements recommandés.

L'analyse factorielle des correspondances

L'étude simultanée des distributions lexicales dans les trois volets du corpus **GUIDES** corpus peut être enrichie par des analyses et des visualisations appuyées sur les méthodes de la *statistique multidimensionnelle*. Cette approche permet de produire des descriptions synthétiques des systèmes de correspondances présents dans le corpus.

Sur chacun des graphiques (cf. **ANNEXE : Figures 10-12**), le plan factoriel engendré par les deux premiers axes de l'*analyse factorielle des correspondances* (AFC) fournit une représentation schématique des informations contenues dans les *tableaux lexicaux* (tableaux à double entrée dont les lignes sont constituées par les ventilations des différentes formes dans les parties du corpus). Les facteurs sont représentables sur un plan graphique sous forme d'axes croisés où l'on situe les parties du corpus les unes par rapport aux autres. Les figures 6-8 présentent les premiers plans factoriels calculés pour chacun des volets du corpus

GUIDES divisé en parties suivant la clef « ville » (cf. *ANNEXE : Figure 2*).

L'approche factorielle de l'échantillon du corpus **GUIDES** confirme une organisation textuelle semblable au sein des volets trilingues. On observe un agencement des points qui distingue nettement les villes balnéaires de Blankenberghe et Heyst du reste de l'échantillon.

Les résultats de cette première exploration montrent que les méthodes factorielles ouvrent des perspectives de recherche prometteuses pour la synthèse globale de l'information contenue dans les guides rédigés dans des langues différentes.

3. Conclusions et perspectives

Au terme de cette première étude textométrique du corpus **GUIDES**, nous constatons que ce matériau textuel trilingue est extrêmement riche et offrent plusieurs pistes de recherche pour une analyse contrastive sur corpus (comparaisons interlingues, évolution du vocabulaire dans le temps, étude contrastive d'espaces urbains, etc.). L'approche textométrique nous semble pertinente pour une analyse contextuelle fine du vocabulaire attesté dans les guides.

Les premiers constats sur l'échantillon récolté par saisie manuelle ont permis d'esquisser les chantiers futurs. Plusieurs d'autres explorations pourraient être envisagées si l'on parvient à consolider un véritable corpus électronique des guides.

ANNEXES

19. Blankenberghe et Heyst.

Hôtels. — *Sur la digue*, à dr. de la montée principale, le *Grand Hôtel des Familles, nouvelle maison fondée par actions, ayant environ 200 chambres, des bains de mer chauds bien organisés, de beaux jardins où se donnent quelquefois des concerts: ch. du côté de la mer, à partir de 10 fr. environ, moins chère de l'autre côté. A g. de la montée principale, les hôt. Godderis, (table d'hôte à 1 h., souper à 7 h., pension, 7 à 15 fr.); Rocher de Cancale, petit; du Cursaal (120 ch., v. ci-dessous); Victoria; de la Plage; *de l'Océan, pas cher. Plus loin à l'O., l'hôt. du Phare.

Dans le village: Grand Hôtel D'hondt, surtout fréquenté par la bourgeoisie belge (pens., 7 fr., la moitié pour les enfants); Dr. Verhaeghe, excellente maison sans enseigne (pens., 8 fr.); Maison des Bains, chez Marchand, bonne (pens., 7 fr.); hôt. de France, (table d'hôte à 1 h. 1/2; pens., 7 fr. 50); du Chemin de fer, Van Mullem, avec bains chauds (1 fr. 50 au cachet); du Lion d'Or, tous rue de l'Eglise; *de Bruges, (dîn., 2 fr. 50); du Couvent; d'Allemagne; de l'Europe; de Gand; de la Paix et du Parc, autant d'hôtels nouveaux.

Restaurants. — *Sur la digue*, outre les hôtels nommés ci-dessus, le Cursaal, local nouveau et élégant, du style mauresque, où l'on ne peut entrer que si l'on est abonné (ch. de 5 à 20 fr.; concerts et bals). La salle de concert se trouve à côté du *café-restaurant. Dîner à 1 h. 1/2, 2 fr. 50; à 5 h., 3 fr.; souper à 7 h.; beaucoup de journaux. Café-restaurant du Pavillon des Bains (table d'hôte, beaucoup de journaux, 40 c. le verre de bière); Pauwels D'hondt, pas cher. — *Dans le village:* l'hôtel du Rhin, où l'on trouve aussi des logements. — La vie est moins chère à Blankenberghe qu'à Ostende; les prix y sont sur la digue ce qu'ils ne sont que dans la ville de l'autre côté; on paie, par exemple, 2 fr. 50 à 3 fr. pour le dîner et 2 fr. pour le souper.

Logements garnis. — *Sur la digue*, entre autres: Maison Segaert, avec un grand balcon (24 ch.); Maison de Gobart (ch. à 1 lit, sur la mer, avec balcon, 4 à 10 fr., service compris; à 2 lits, 2 fr. de plus; pension); *Succursale du Dr. Verhaeghe (v. ci-dessus; pension avec ch. sur la mer 10 fr. 50, moins chère du côté de la ville). Il s'est élevé de plus ici dans ces derniers temps tout une rangée de maisons à louer, plus ou moins élégantes, pour des familles. — *Dans le village:* de nombreux appartements dans toutes les rues; prix d'une chambre à 1 lit, en moyenne, 2 à 3 fr.; à 2 lits, 3 à 4 fr. Toutefois, de la mi-juillet à la fin d'août, il est prudent de s'assurer d'avance un logement de ce genre, ou du moins d'arriver assez tôt à Blankenberghe pour pouvoir retourner au besoin, le soir, à Bruges.

Médecins: Dr. Verhaeghe, Dr. van Mullem et Dr. Cosyn.

Bains. Cartes de bains, 60 cent. la pièce. Les cabines sont conduites à l'eau non pas par des chevaux, mais à bras d'hommes. On donne lors du départ quelques francs de pourboire pour le séchage du linge et les chaises sur la plage.

Tentes sur la plage pour se mettre à l'abri contre le vent et le soleil (non contre la pluie), 1 fr. par jour.

Costumes de bain de tricot pour hommes, 5 fr.; hauts, avec surtout, 8 fr.

Barques: 5 fr. la course; en société, 1 fr. par personne. — **Anes**, dont les propriétaires sont importuns: pour Heyst (p. 140), 2 à 3 fr., et pour d'autres excursions sur la plage.

Les feuilles *La Plage* et *La Vigie de la Côte* donnent la liste des étrangers et annoncent les heures du flux et du reflux. La première paraît tous les jeudis et dimanches, la seconde les dim. seulement.

Coquillages, aux abords des dunes et dans le village.

Bains chauds à l'hôtel des Familles (v. ci-dessus), chez le Dr. van Mullem, rue de l'Eglise, et chez P. van Wulpen de Langhe, Bakkerstraat, 8; 2 fr. 50 le bain, 1 fr. pour les enfants.

Blankenberghe, à 4 h. N.-E. d'Ostende et à 3 h. au N. de Bruges, est un village de pêcheurs d'environ 2,350 hab., composé de petites

Figure 1. Fac-simile d'une page du corpus *GUIDES* [Baedeker, 1873].

<langue=français><année=1873><page=139_141><ville=blankenberghe_heyst>

§

19. blankenberghe et heyst

§

hôtels. - sur la digue, à dr. de la montée principale, le *grand hôtel des familles, nouvelle maison fondée par actions, ayant environ 200 chambres, des bains de mer chauds bien organisés, de beaux jardins où se donnent quelquefois des concerts : ch. du côté de la mer, à partir de 10 fr. environ, moins chère de l'autre côté. a g. de la montée principale, les hôt. godderis, (table d'hôte à 1h., souper à 7h., pension, 7 à 15 fr.) ; rocher de cancale, petit ; du cursaal (120 ch., v. ci-dessous) ; victoria ; de la plage ; *de l'océan, pas cher. plus loin, à l'o., l'hôt. du phare.

§

dans le village : grand hôtel d'hondt, surtout fréquenté par la bourgeoisie belge (pens., 7 fr., la moitié pour les enfants) ; dr. verhaeghe, excellente maison sans enseigne (pens., 8 fr.) ; maison des bains, chez marchand, bonne (pens., 7 fr.) ; hôt. de france, (table d'hôte à 1 h. ½ ; pens., 7 fr. 50) ; du chemin de fer, van mullem, avec bains chauds (1 fr. 50 au cachet) ; du lion d'or, tous rue de l'église ; *de bruges, (dîn., 2 fr. 50) ; du couvent ; d'Allemagne ; de l'europe ; de gand ; de la paix et du parc, autant d'hôtels nouveaux.

§

restaurants. - sur la digue, outre les hôtels nommés ci-dessus, le cursaal, local nouveau et élégant, du style mauresque, où l'on ne peut entrer que si l'on est abonné (ch. de 5 à 20 fr. ; concerts et bals). la salle de concert se trouve à côté du *café-restaurant. dîner à 1 h. ½, 2 fr. 50 ; à 5 h., 3 fr. ; souper à 7 h. ; beaucoup de journaux. café-restaurant du pavillon des bains (table d'hôte, beaucoup de journaux, 40 c. le verre de bière) ; pauwels d'hondt, pas cher. - dans le village : l'hôtel du rhin, où l'on trouve aussi des logements. - le vie est moins chère à blankenberghe qu'à ostende ; les prix y sont sur la digue ce qu'ils ne sont que dans la ville de l'autre côté ; on paie, par exemple, 2 fr. 50 à 3 fr. pour le dîner et 2 fr. pour le souper.

§

logements garnis. - sur la digue, entre autres : maison segaert, avec un grand balcon (24 ch.) ; maison de gobart (ch. à un lit, sur la mer, avec balcon, 4 à 10 fr., service compris ; à 2 lits, 2 fr. de plus ; pension) ; *succursale du dr. verhaeghe (v. ci-dessus ; pension avec ch. sur la mer, 10 fr. 50, moins chère du côté de la ville). il s'est élevé de plus ici dans ces derniers temps tout une rangée de maisons à louer, plus ou moins élégantes, pour des familles. - dans le village : de nombreux appartements dans toutes les rues ; prix d'une chambre à 1 lit, en moyenne, 2 à 3 fr. ; à 2 lits, 3 à 4 fr. toutefois, de la mi-juillet à la fin d'août, il est prudent de s'assurer d'avance un logement de ce genre, ou du moins d'arriver assez tôt à blankenberghe pour pouvoir retourner au besoin, le soir, à bruges.

§

médecins : dr. verhaeghe, dr. van mullem et dr. cosyn.

§

bains. cartes de bains, 60 cent. la pièce. les cabines sont conduites à l'eau non pas par des chevaux mais à bras d'hommes. on donne lors du départ quelques francs de pourboire pour le séchage du linge et les chaises sur la plage.

tentes sur la plage pour se mettre à l'abri contre le vent et le soleil (non contre la pluie), 1 fr. par jour.

costumes de bain de tricot pour hommes, 5 fr. ; hauts, avec surtout, 8 fr.

§

barques : 5 fr. la course ; en société, 1 fr. par personne. Anes

/.../

Figure 2. Retranscription informatique du document – Codage pour *Lexico3*.


```

<?xml version="1.0" encoding="Latin1"?>

<CORPUS>

<CORPUSHEAD>
<CORPUSSOURCE>"Sciences de la ville - Univ. Denis Diderot"</CORPUSSOURCE>
<CORPUSTITLE>"Guides Baedeker"</CORPUSTITLE>
<CORPUSAUTHOR>"Verlag Karl Baedeker"</CORPUSAUTHOR>
<CORPUSCOMPILATIONDATE>"01032007"</CORPUSCOMPILATIONDATE>
</CORPUSHEAD>

<PAGE=139>

<DOCUMENT Number="0" Language="Français" Title="Blankenberghe et Heyst" Date=" 1873">

<TITLE>
19. Blankenberghe et Heyst
</TITLE>

<PARAGRAPH Number="0">

    <SUBTITLE>
    hôtels.
    </SUBTITLE>

    <PHRASE Number="0">
    - sur la digue, à dr. de la montée principale, le *grand hôtel des familles,
nouvelle maison fondée par actions, ayant environ 200      chambres, des bains de mer chauds
bien organisés, de beaux jardins où se      donnent quelquefois des concerts : ch. du côté
de la mer, à partir de 10      fr.
    </PHRASE>

    <PHRASE Number="1">
    environ, moins chère de l'autre côté. a g. de la montée principale, les      hôt.
godderis, (table d'hôte à 1h., souper à 7h., pension, 7 à 15 fr.) ;      rocher de cancale,
petit ; du cursaal (120 ch., v. ci-dessous) ;      victoria ; de la plage ; *de l'océan,
pas cher.
    </PHRASE>

    <PHRASE Number="2">
    plus loin, à l'o., l'hôt. du phare.
    </PHRASE>

</PARAGRAPH>

<PARAGRAPH Number="1">

    <PHRASE Number="3">
    dans le village : grand hôtel d'hondt, surtout fréquenté par la      bourgeoisie belge
(pens., 7 fr., la moitié pour les enfants) ; dr. verhaeghe,      excellente maison sans
enseigne (pens., 8 fr.) ; maison des      bains, chez marchand, bonne (pens., 7 fr.) ;
hôt. de france, (table      d'hôte à 1 h. ½ ; pens., 7 fr. 50) ; du chemin de fer, van
mullem, avec bains chauds (1 fr. 50 au cachet) ; du lion d'or, tous rue de l'eglise ; *de
bruges, (din., 2 fr. 50) ; du couvent ; d'Allemagne ; de l'Europe ;      de gand ; de la
paix et du parc, autant d'hôtels nouveaux.
    </PHRASE>

</PARAGRAPH>
(...)

```

Figure 3. Retranscription informatique du document – Codage XML.

Langue	Nombre d'occurrences	Nombre de formes	Fréquence maximale	Nombre d'hapax
anglais	7 417	2 272	718 (<i>the</i>)	1 455
français	1 0719	2 813	783 (<i>de</i>)	1 743
allemand	8 049	3 016	305 (<i>der</i>)	2 120

Figure 4. Corpus *GUIDES* : caractéristiques lexicométriques de l'échantillon numérisé.

The screenshot shows a window titled "Groupes de formes" with the following fields and controls:

- Nom du groupe : basili/cath/kirch+
- Le motif : kirch
- est contenu dans ce que je recherche
- Ajouter
- Rechercher (lightbulb icon)
- Enregistrer (floppy disk icon)
- Supprimer (trash can icon)
- Charger (upload icon)

Below the controls is a table with two columns: "Forme" and "Fréquence".

Forme	Fréquence
basilika	2
basilica	1
cathedrale	4
cathédrale	3
cathédrale	1
kirche	14
peterskirche	4
jacobskirche	3
kirchen	3
gertrudenkirche	2
martinskirche	2
paulskirche	2
andreakirche	1
domkirche	1
frauenkirche	1
jesuitenkirche	1
kirchendaches	1
kirchwäter	1
kreuzkirche	1
lambertuskirche	1
laurentiuskirche	1
michaelskirche	1
pfarrkirche	1
prachtkirchen	1
quintinskirche	1
stiftskirche	1

Figure 5. Groupes de formes rassemblant les occurrences liées par une propriété commune.

Tableau 6. Inventaire des segments répétés en anglais (extrait).

2	----	----	----	----	the 11th
2	----	----	----	----	the 12th
3	----	----	----	----	the 16th
	2	----	----	----	the 16th cent
2	----	----	----	----	the belgian
3	----	----	----	----	the bishop's
3	----	----	----	----	the bishops
7	----	----	----	----	the bold
2	----	----	----	----	the branch
2	----	----	----	----	the canal
	2	----	----	----	the capital of
2	----	----	----	----	the carved
5	----	----	----	----	the centre
	3	----	----	----	the centre of
3	----	----	----	----	the ch
	2	----	----	----	the ch of st
2	----	----	----	----	the chapel
2	----	----	----	----	the chapels
11	----	----	----	----	the choir
	4	----	----	----	the choir is
3	----	----	----	----	the church
3	----	----	----	----	the citizens
	2	----	----	----	the citizens were
3	----	----	----	----	the city
	2	----	----	----	the convent of st
	2	----	----	----	the death of
3	----	----	----	----	the duke
	2	----	----	----	the duke of
	2	----	----	----	the empr charles v
4	----	----	----	----	the english
2	----	----	----	----	the enterprise
2	----	----	----	----	the events
3	----	----	----	----	the existing
	2	----	----	----	the existing building is
3	----	----	----	----	the finest
	2	----	----	----	the finest in belgium
3	----	----	----	----	the first
8	----	----	----	----	the french
5	----	----	----	----	the front
2	----	----	----	----	the gates
2	----	----	----	----	the government
2	----	----	----	----	the grand
	2	----	----	----	the grande place
4	----	----	----	----	the ground
	2	----	----	----	the h de ville
2	----	----	----	----	the haven
	3	----	----	----	the high altar
2	----	----	----	----	the highest
2	----	----	----	----	the hotels
3	----	----	----	----	the houses
2	----	----	----	----	the invading
	3	----	----	----	the junction of the
3	----	----	----	----	the king
	3	----	----	----	the l bank
6	----	----	----	----	the largest
	3	----	----	----	the largest in belgium
2	----	----	----	----	the last
3	----	----	----	----	the late
4	----	----	----	----	the liégeois
2	----	----	----	----	the longdoz
2	----	----	----	----	the lower
2	----	----	----	----	the main
9	----	----	----	----	the meuse
2	----	----	----	----	the mines
2	----	----	----	----	the mitre
2	----	----	----	----	the modern
11	----	----	----	----	the most
	2	----	----	----	the most interesting
	2	----	----	----	the name of
2	----	----	----	----	the natural
7	----	----	----	----	the nave
2	----	----	----	----	the niches
3	----	----	----	----	the novel
8	----	----	----	----	the old

	2	----	----	the old hall is curious
3	----	----	----	the oldest
3	----	----	----	the other
3	----	----	----	the ourthe
	2	----	----	the painted glass in
2	----	----	----	the pictures
6	----	----	----	the place
	3	----	----	the place du théâtre
2	----	----	----	the pont
2	----	----	----	the present
2	----	----	----	the princes
3	----	----	----	the principal
3	----	----	----	the railway
2	----	----	----	the real
2	----	----	----	the rest
2	----	----	----	the richest
8	----	----	----	the river
2	----	----	----	the roads
2	----	----	----	the rue
2	----	----	----	the sambre
7	----	----	----	the same
2	----	----	----	the sand
2	----	----	----	the scenery
4	----	----	----	the schelde
4	----	----	----	the sea
2	----	----	----	the space
2	----	----	----	the stat
3	----	----	----	the stream
2	----	----	----	the streets
2	----	----	----	the theatre
2	----	----	----	the tower
23	----	----	----	the town
	2	----	----	the town hall
	2	----	----	the town is
	2	----	----	the town liége
	2	----	----	the town the
2	----	----	----	the transepts
5	----	----	----	the university
2	----	----	----	the upper
2	----	----	----	the valley
2	----	----	----	the vast
6	----	----	----	the vesdre
	2	----	----	the vesdre the
2	----	----	----	the view
	3	----	----	the w front
2	----	----	----	the walloons
2	----	----	----	the water
	2	----	----	the way to
	2	----	----	the windings of the
	2	----	----	the work of
2	----	----	----	the works

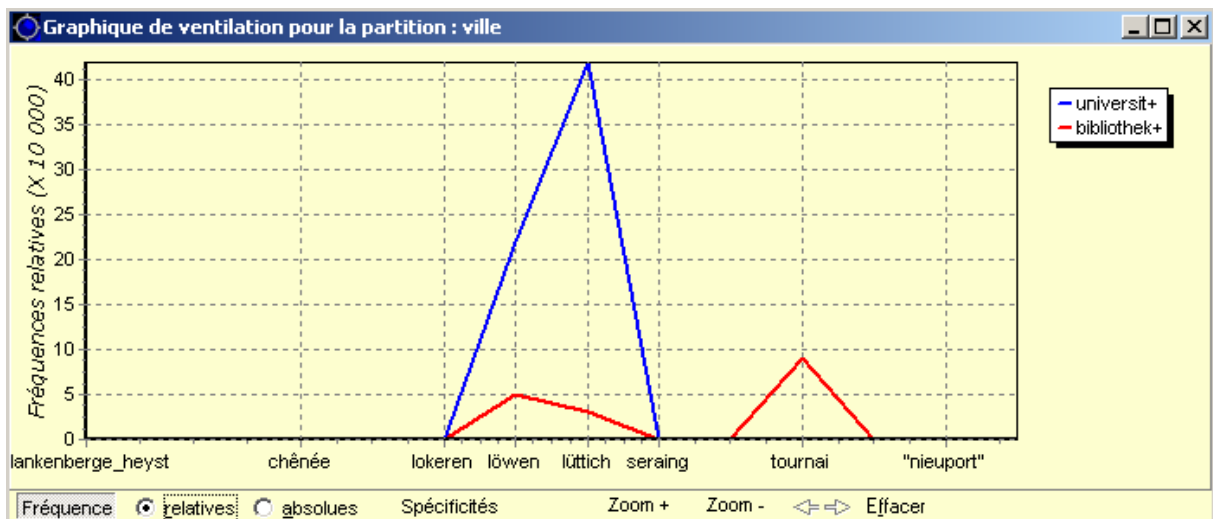
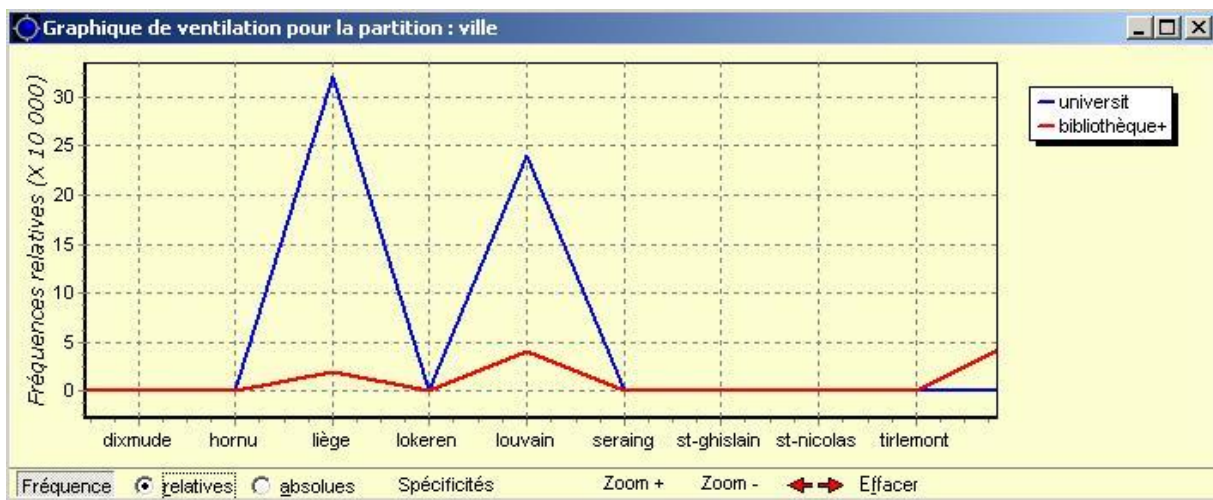
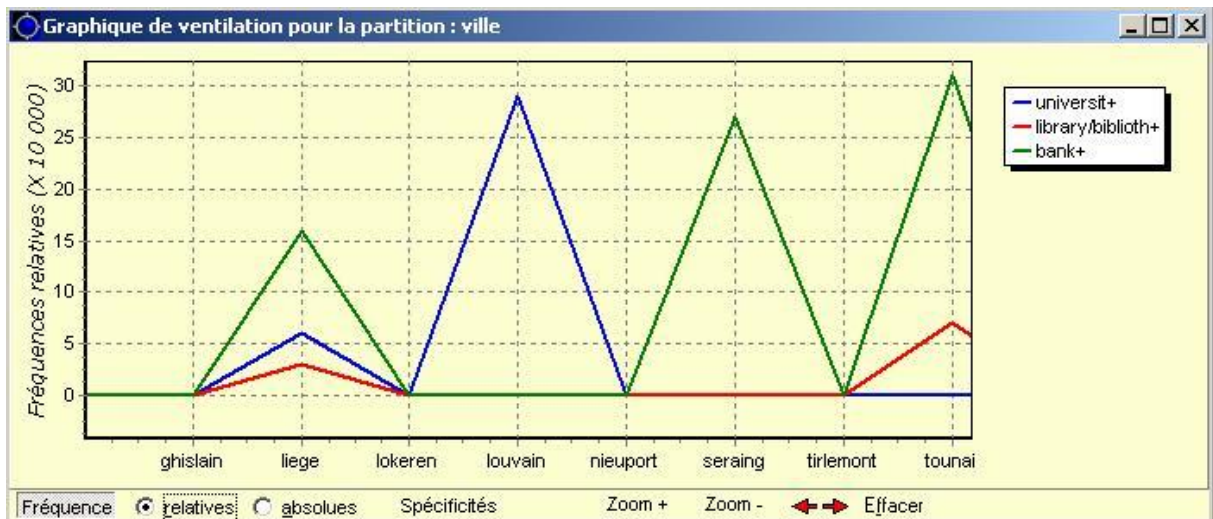


Figure 7. Analyse comparative des ventilations lexicales.

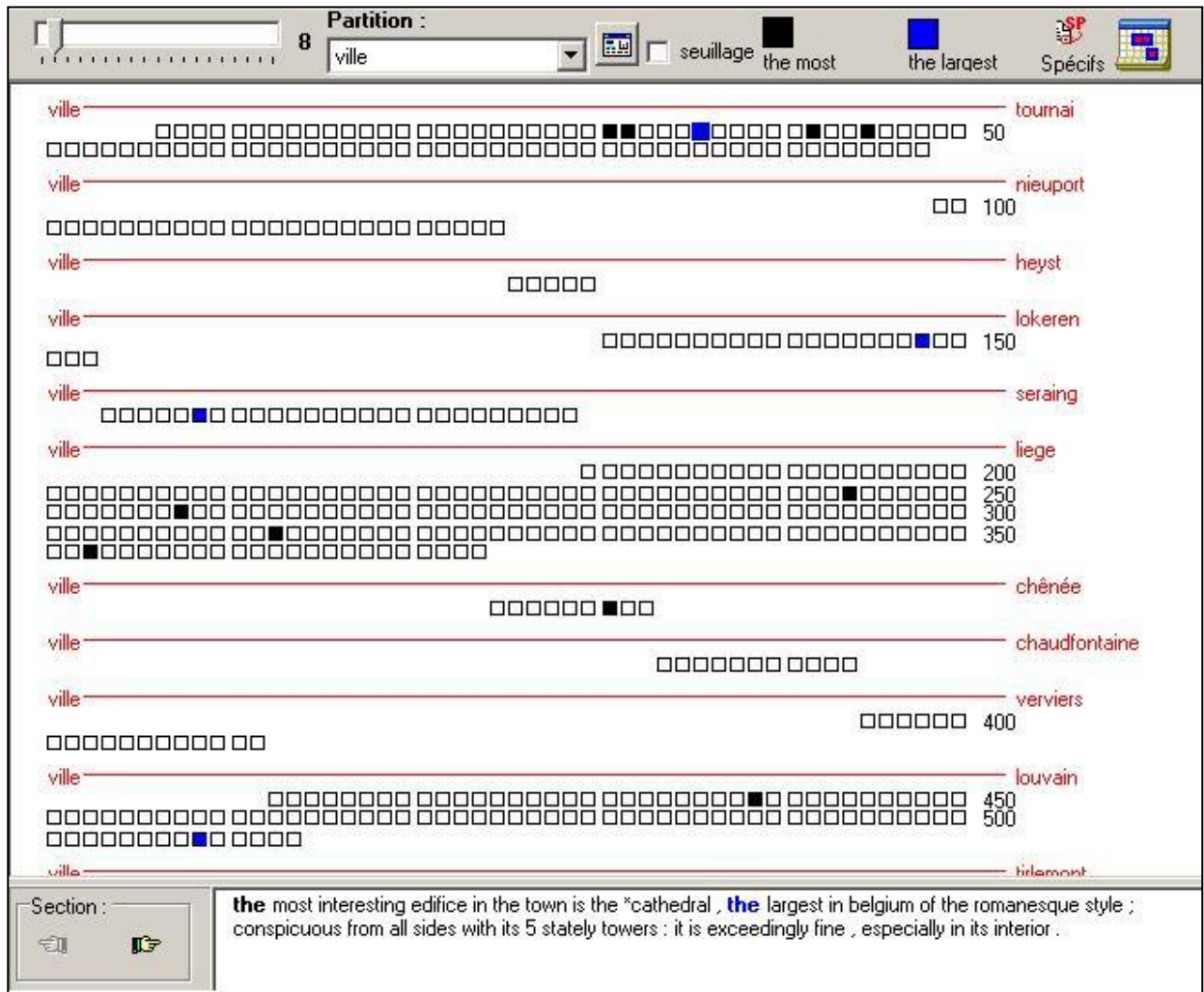


Figure 8. Carte des sections : localisation des unités complexes (*segments répétés*) dans les phrases (*sections*) du corpus partitionné.

Spécificités positives									
Blankenberghe		Heyst		Liège		Louvain		Tournai	
forme	sp.	forme	sp.	forme	sp.	forme	sp.	forme	sp.
<i>fr</i>	+12	<i>kil</i>	+6	<i>pl</i>	+10	<i>chap</i>	+5	<i>salle</i>	+7
<i>plage</i>	+6	<i>millions</i>	+6	<i>théâtre</i>	+5	<i>porte</i>	+4	<i>plus</i>	+4
<i>bains</i>	+6	<i>hab</i>	+5	<i>meuse</i>	+4			<i>transept</i>	+4
<i>ch</i>	+4	<i>hôt</i>	+5					<i>petite</i>	+4
<i>mer</i>	+4	<i>min</i>	+5					<i>on</i>	+4
<i>maison</i>	+4	<i>seraing</i>	+4						
<i>pens</i>	+4	<i>village</i>	+4						
<i>enfants</i>	+4								
<i>village</i>	+4								
<i>/.../</i>									
Spécificités négatives									
Blankenberghe		Heyst		Liège		Louvain		Tournai	
forme	sp.	forme	sp.	forme	sp.	forme	sp.	forme	sp.
<i>ville</i>	-2	<i>st</i>	-3	<i>hab</i>	-3			<i>fr</i>	-3
				<i>hôtel</i>	-3			<i>h</i>	-2
				<i>petite</i>	-2				
				<i>salle</i>	-2				
<i>/.../</i>									

Figure 9. GUIDES : Spécificités lexicales de chaque ville [volet français]

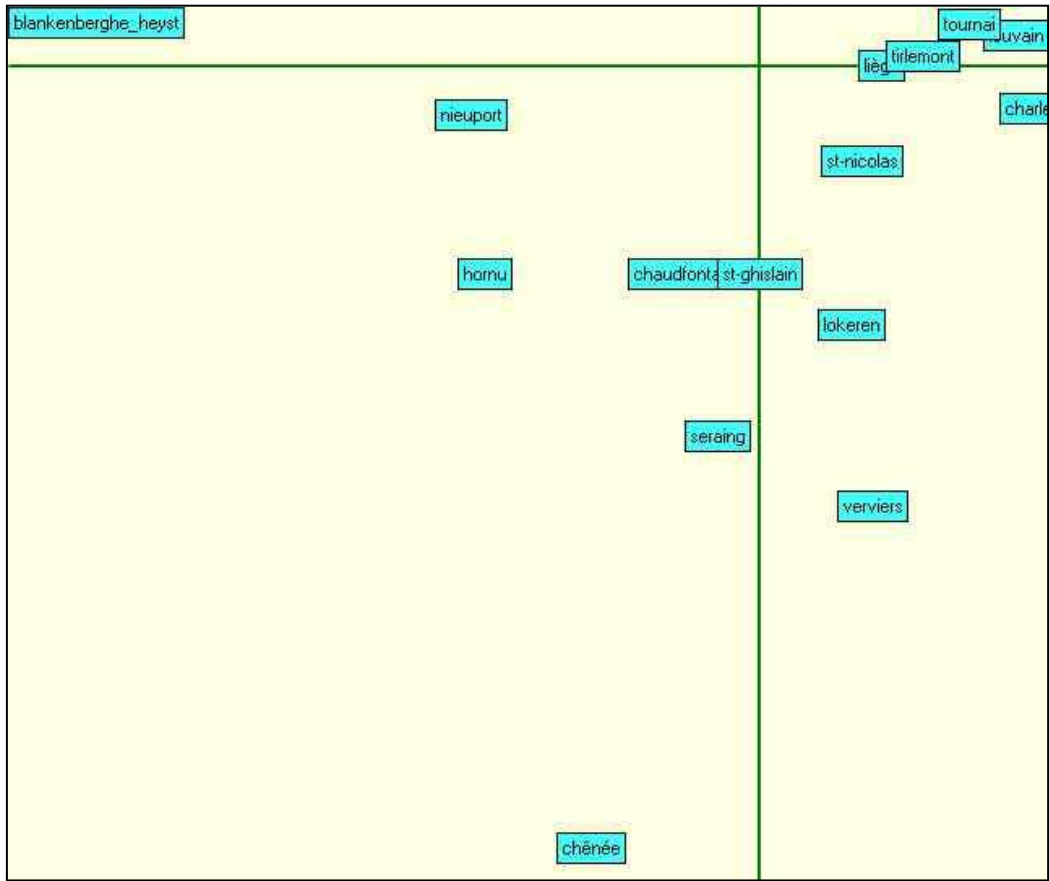


Figure 10. GUIDES : L'analyse factorielle des correspondances [volet français]

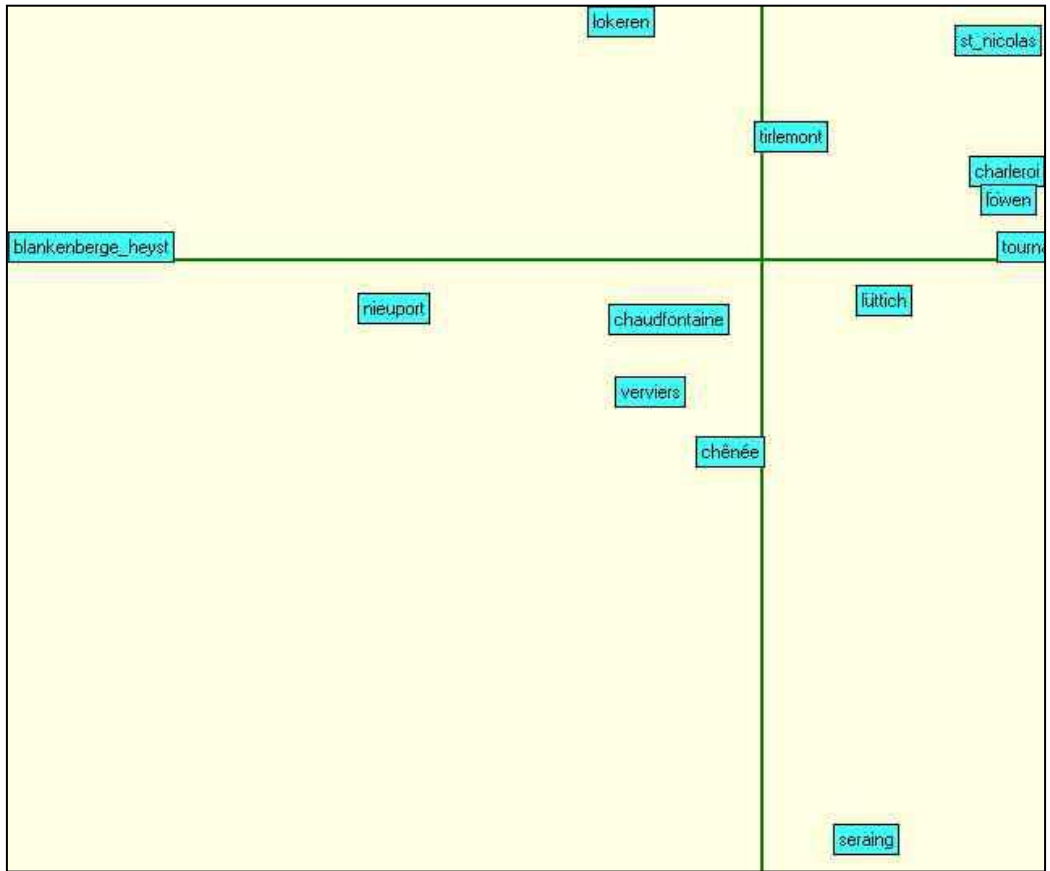


Figure 11. GUIDES : L'analyse factorielle des correspondances [volet allemand]

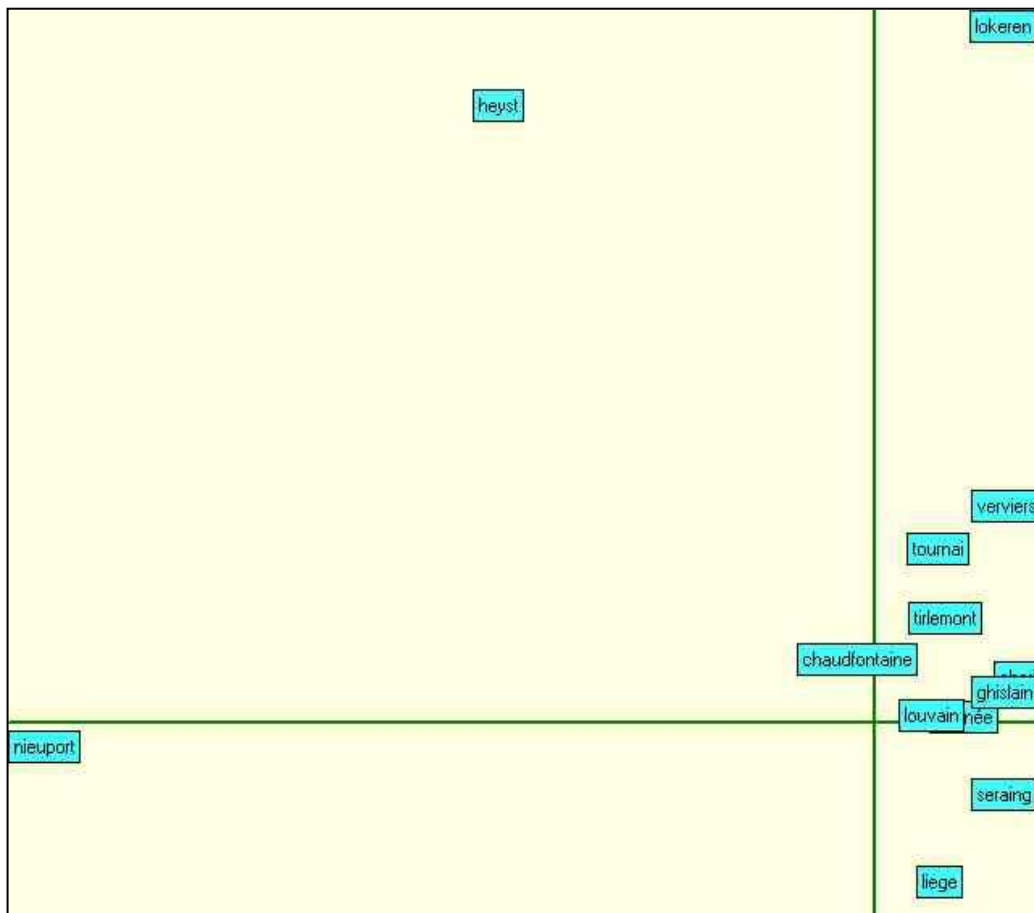


Figure 12. GUIDES : L'analyse factorielle des correspondances [volet anglais]

Bibliographie

Sites Web :

Portail Traitement Automatique du Langage (SYLED-CLA2T, Université Paris 3 - Sorbonne nouvelle) :

<http://tal.univ-paris3.fr>

Répertoires d'Outils et de Ressources (SYLED-CLA2T, Université Paris 3 - Sorbonne nouvelle) :

<http://www.cavi.univ-paris3.fr/ilpga/syled/outils-cla2t.htm>

<http://tal.univ-paris3.fr/outils-tal-p3.html>

Lexico3 (logiciel conçu pour le traitement lexicométrique de textes comportant plusieurs centaines de milliers d'occurrences) :

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

La revue ***LEXICOMETRICA*** :

<http://www.cavi.univ-paris3.fr/lexicometrica/>

Blog TAL-P3 (Journal de lectures, de liens, d'activités pour les étudiants du secteur TAL) :

<http://tal.univ-paris3.fr/blogtal>

JADT, Journées d'Analyse de Données Textuelles (liens et actes) :

<http://www.cavi.univ-paris3.fr/lexicometrica/jadt>

Ouvrages :

HABERT, B., NAZARENKO, A., SALEM, A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin / Masson.

LEBART, L., SALEM, A. (1994). *Statistique textuelle*. Paris : Dunod.

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

Articles de revue :

FLEURY, S., ZIMINA, M. (2007) « Exploring translation corpora with *MkAlign*. » In Bokor G. (Ed.) Revue électronique *Translation Journal*, janvier 2007.

<http://accurapid.com/journal/39mk.htm>

HABERT, B., SALEM, A. (1995). « L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. » *TAL* n°36(1-2), p. 249-276.

LEBLANC, J-M., MARTINEZ, W. (2005). « Positionnements énonciatifs dans les vœux présidentiels sous la cinquième République. » *CORPUS* n° 4 « Les corpus politiques : objet, méthode et contenu. », décembre 2005.

<http://corpus.revues.org/document347.html>

ZIMINA, M. (2002) « Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. » Revue électronique *LEXICOMETRICA* n° spécial « Corpus alignés » [Véronis J. (Ed.)].

<http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6/Zimina.pdf>

Articles dans des actes de congrès internationaux :

LAMALLE, C., SALEM, A. (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. » In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles JADT'02*, Saint-Malo, 13-15 mars 2002, p. 403-411.

http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/lamalle_salem.pdf

MARTINEZ, W. (2000). « Mise en évidence de rapports synonymiques par la méthode des cooccurrences. » In *Actes des 5èmes Journées d'Analyse Statistique des Données Textuelles JADT'00*, Ecole Polytechnique Fédérale de Lausanne (Suisse), 9-11 mars 2000, pp. 171-178.

<http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/78/78.pdf>

MARTINEZ, W., ZIMINA, M. (2002). « Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. » In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles JADT'02*, Saint-Malo, 13-15 mars 2002, p. 495-506.

http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/martinez_zimina.pdf

SALEM, A. (2004). « Introduction à la résonance textuelle. » In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles JADT'04*, Louvain La Neuve, 10-12 mars 2006, p. 986-992.

http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_096.pdf

SALEM, A. (2006). « Proximités segmentales. » In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles JADT'06*, Besançon, 19-21 avril 2006, p. 986-

992.

<http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/II-075.pdf>

ZIMINA, M. (2000) « Alignement de textes bilingues par classification ascendante hiérarchique. » In *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles JADT'00*, Ecole Polytechnique Fédérale de Lausanne (Suisse), 9-11 mars 2000, pp. 171-178.

<http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/77/77.pdf>

ZIMINA M. (2004) « L'alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. » In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles JADT'04*, Louvain-la-Neuve (Belgique), 10-12 mars 2004, pp. 1195-1202.

http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_118.pdf

ZIMINA, M. (2005). « Topographie bi-textuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles. », In *Actes des 7es Journées scientifiques du Réseau de chercheurs 'Lexicologie, Terminologie, Traduction'*, Bruxelles, 8-10 septembre 2005, p. 175-186.

<http://perso.univ-lyon2.fr/~thoiron/JS%20LTT%202005/pdf/Zimina.pdf>

Ouvrages techniques :

FLEURY, S., ZIMINA, M. (2006) « MkAlign ». *Manuel d'utilisation*. SYLED-CLA2T, Université de la Sorbonne nouvelle - Paris 3.

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

LAMALLE, C., MARTINEZ, W., FLEURY, S., SALEM, A., FRACCHIOLLA, B., KUNCOVA, A., LANDE, B., MAISONDIEU, A., POIROT-ZIMINA, M. (2004) *Lexico3 Textometric toolbox User's manual*. SYLED-CLA2T, Université de la Sorbonne nouvelle – Paris 3.

<http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuelsL3/L3-usermanual.pdf>

MARTINEZ, W. (2005) « COOCS – Outils lexicométriques pour l'analyse des cooccurrences. » *Manuel d'utilisation*. SYLED-CLA2T, Université de la Sorbonne nouvelle – Paris 3.

<http://www.cavi.univ-paris3.fr/ilpga/individus/martinez/>