Que faire des corpus multilingues parallèles ? Une expérience

Jean-Claude Deroubaix

Groupe de recherche sur les acteurs internationaux et leurs discours
(GRAID-Institut de Sociologie) – Université Libre de Bruxelles
44 av. Jeanne – 1050 Bruxelles – Belgique
deroubaix@swing.be

Abstract

With the movement going to a globalisation of political systems, we are confronted more and more with political text written in more than one language. This proliferation was first a challenge for translators. Some progress have been done to facilitate the manipulation of these texts in the context of translation. We propose to examine how it is possible to use traditional tools of lexicometrics to describe and analyse this kind of text from the point of view of the social scientist.. We'll examine also the benefits we could obtain from these attitude.

Résumé

Face au mouvement de mondialisation des systèmes politiques, nous sommes confrontés de plus en plus souvent à des textes politiques rédigés en plusieurs langues. Cette prolifération fut d'abord un défi pour les traducteurs. Des progrès ont été réalisés pour faciliter le traitement de ces textes dans le contexte de la traduction. Nous nous proposons d'examiner comment les outils classiques de la lexicométrie peuvent être utilisés pour décrire et analyser ce genre de corpus du point de vue des sciences sociales. Nous examinerons aussi les bénéfices que l'on peut attendre d'un tel point de vue.

Mots-clés : discours politique, statistique lexicale, corpus multilingues.

1. Les corpus multilingues

Lors des JADT[1] de Nice, ma communication (Deroubaix, 1998) avait porté sur la nécessité de construire des outils d'analyse performants des corpus multilingues. En effet, l'émergence puis le renforcement d'accords institutionnels et politiques qui dépassent largement le cadre de simples accords classiques de relations internationales diplomatiques pour constituer des sociétés politiques couvrant de grandes régions du monde (Union européenne, ALENA) ou plus radicalement le monde entier (OMC) engendrent dès lors des textes politiques qui s'imposent souvent comme supérieurs aux normes nationales. Ces textes sont généralement multilingues, ayant valeur légale, dans chacune des langues originelles (onze langues pour l'Union européenne par exemple et pour l'instant). En outre ils sont traduits dans d'autres langues si nécessaire en vue d'être compris par l'ensemble des populations qui y sont soumises.

La multiplication de tels textes soulève de nombreuses questions.

Les premières concernent essentiellement les traducteurs qui ont à faire face à une croissance exponentielle de texte et à la nécessité d'automatiser au moins partiellement leur travail,

[1] Pour la réalisation de cette communication nous avons utilisé la suite logicielle LEXICO1 pour Mac développée au Laboratoire de Lexicologie politique de Saint-Cloud par André Salem. Les analyses factorielles ont été réalisées avec les programmes de l'ADDAD, mis à la disposition des utilisateurs de LEXICO1.

d'une part, et d'autre part de maintenir une cohérence traductive difficile à obtenir lorsqu'il s'agit d'effectuer des traductions entre systèmes politiques parfois bien différents[2].

Une seconde liste de questions concernent plus particulièrement l'analyse de ces textes politiques auxquels il peut paraître un peu trop simple d'appliquer des analyses scientifiques sur une seule des versions linguistiques sans prendre en compte le fait qu'il ne s'agit en l'occurrence que d'analyse portant sur un extrait d'un corpus multilingue par nature.

Les questions proprement traductives ont suscité évidemment de nombreux travaux tant sur la manière de constituer des mémoires traductives que sur la réalisation semi-automatique de glossaires ou même pour aller en amont de ces deux démarches pour réaliser des outils d'alignement de corpus. (cf., par exemple, la bibliographie dans Kraif, 2001).

2. Les questions d'analyse lexicale

Un moindre intérêt a été porté à l'analyse lexicale multilingue. Pourtant, les outils classiques de l'analyse lexicométrique peuvent fournir des résultats intéressants lorsqu'ils sont appliqués à des corpus multilingues. En effet, peu de chose distinguent formellement les corpus multilingues des corpus traditionnellement explorés par les analyses lexicométriques. L'élément fondamental sur lequel les calculs statistiques de la lexicométrie vont être réalisés est le tableau lexical entier c'est-à-dire un tableau des fréquences des formes lexicales (ou des lemmes, éventuellement) attestées dans ce corpus et ventilées dans l'ensemble des parties du corpus.

C'est sur ce tableau que vont se calculer les spécificités, vont se construire des classifications des parties ou des formes lexicales ou se calculer divers indices de richesse de vocabulaire par exemple. Voici le début d'un tel tableau. Il s'agit de la distribution des occurrences des mots de plus grande fréquence dans le corpus constitué des 11 textes intitulés « Grandes orientations de politique économique »[3] adoptés par le Conseil européen. Il s'agit du corpus français.

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
de	143	143	232	179	285	500	970	1381	2010	1927	1966
la	104	123	168	134	187	324	555	876	1232	1093	1256
des	84	78	150	111	163	282	523	668	979	852	966

Cependant, dans le cas d'un corpus multilingue, nous disposons d'autant de tableaux lexicaux qu'il y a de langues utilisées dans le corpus parallèle. Ainsi pouvons-nous construire :

En italien :

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
di	105	101	143	111	218	402	726	1018	1331	1415	1384
e	69	75	98	95	139	203	413	658	839	977	897
la	52	69	83	62	96	132	257	442	632	669	656

[2] Ainsi est-il malaisé de traduire les concepts qui organisent la sécurité sociale de chacun des pays membres de l'UE sans agir sur leur signification. Cf. le sens de « cotisation sociale » dans Friot (2003).

[3] Ces documents portent évidemment des noms différents dans les versions linguistiques différentes. Nous désignerons désormais ce corpus, et les textes qui le constituent, avec l'abréviation française GOPE quelle que soit la langue considérée.

En anglais :

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
the	202	201	280	239	368	527	981	1537	1913	2179	1934
of	90	79	145	113	173	254	484	818	1087	1226	1029
in	79	79	114	110	189	243	479	752	1074	1160	993

En allemand :

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
die	138	146	155	135	195	322	579	711	1215	1282	1120
der	131	102	154	139	193	287	569	950	1044	1201	1175
und	71	77	110	110	150	236	505	672	934	1009	965

La juxtaposition de ces tableaux lexicaux nous donne un nouveau tableau lexical qui correspond à celui que nous aurions obtenu en soumettant à la segmentation et au comptage un corpus multilingue dont les parties seraient chacune constituée des différentes versions linguistiques du même texte disposées à la queue leu leu.

	GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
1	de	143	143	232	179	285	500	970	1381	2010	1927	1966
2	la	104	123	168	134	187	324	555	876	1232	1093	1256
3	des	84	78	150	111	163	282	523	668	979	852	966
…	…	…	…	…	…	…	…	…	…	…	…	…
1487	visent	0	0	1	0	0	0	1	1	2	2	3
1488	di	105	101	143	111	218	402	726	1018	1331	1415	1384
1489	e	69	75	98	95	139	203	413	658	839	977	897
1490	la	52	69	83	62	96	132	257	442	632	669	656
…	…	…	…	…	…	…	…	…	…	…	…	…
3066	vincolanti	0	0	0	0	1	0	3	1	3	2	0
3067	the	202	201	280	239	368	527	981	1537	1913	2179	1934
3068	of	90	79	145	113	173	254	484	818	1087	1226	1029
3069	in	79	79	114	110	189	243	479	752	1074	1160	993
…	…	…	…	…	…	…	…	…	…	…	…	…
4420	week	0	0	0	0	0	0	1	3	2	4	0
4421	die	138	146	155	135	195	322	579	711	1215	1282	1120
4422	der	131	102	154	139	193	287	569	950	1044	1201	1175
4423	und	71	77	110	110	150	236	505	672	934	1009	965
…	…	…	…	…	…	…	…	…	…	…	…	…
6017	zunehmende	0	0	0	0	0	2	0	1	2	4	1

En pratique, cependant, il importe de recourir à une segmentation séparée des textes selon la langue en vue d'éviter quelques problèmes d'homographies translinguistiques[4].

[4] Par exemple lors du traitement du corpus bilingue des déclarations gouvernementales belges (Deroubaix, 1997) nous nous sommes trouvé devant le problème, trivial, de devoir distinguer *de* préposition en français, du

L'intérêt d'une telle démarche réside dans la possibilité de donner comme point de référence aux analyses, la totalité du corpus multilingue et non celle d'une seule de ses parties. Ainsi dans l'application d'une analyse factorielle des correspondances, nous pouvons décrire le nuage des parties immédiatement par rapport à toutes les langues, nous pouvons aussi situer les différentes versions linguistiques de chacune des parties les unes par rapport aux autres. Quant au nuage des formes, il nous sera loisible entre autres de repérer les relations des formes entre elles indépendamment de leur langue. Ces avantages peuvent aussi être mis à profit lors de la construction de typologies par classification automatique.

3. Les GOPE et leurs formes les plus fréquentes

Dans nos démocraties politiques, il est de grands textes politiques qui suscitent les commentaires de la presse et font débat entre les citoyens ; parmi ces textes figurent certes les déclarations solennelles des gouvernements nationaux, les programmes de ces mêmes gouvernements et les programmes politiques des partis.. Les citoyens en prennent connaissance immédiatement ou par le biais des résumés fournis par la presse. Chacun d'entre-eux pressent que ces textes vont avoir une influence certaine sur la vie politique d'abord, sur la vie quotidienne ensuite. Les grandes orientations de politique économique adoptées par le conseil européen ne jouissent évidemment pas de la même publicité ni de la même attention du citoyen ou de la presse. Pourtant il s'agit de l'énoncé du programme de politique économique que l'UE attend de voir concrétiser dans l'année par les États membres. Ce programme est très concret : il contient des recommandations générales et des recommandations pays par pays. Les programmes gouvernementaux nationaux sont subordonnés au suivi de ces recommandations. C'est la raison pour laquelle, au GRAID, nous avons décidé d'étudier le vocabulaire de ces textes, car du fait même qu'ils se présentent sous la forme de recommandations, ils forment un pont, un vecteur de la circulation lexicale des termes du pouvoir (cf. Gobin, 2003), un lieu essentiel pour comprendre comment se dit et se fait la politique aujourd'hui.

Observons d'abord les 15 termes les plus fréquents de ces GOPE (en ayant éliminé de la liste les mots-outils, prépositions et articles essentiellement).

Formes lexicales les plus fréquentes dans les GOPE (avec leur rang)			
En français	*En italien*	*En anglais*	*En allemand*
22 emploi	19 lavoro	9 labour	27 öffentlichen
24 travail	25 mercato	13 market	29 maßnahmen
25 marché	26 crescita	14 growth	32 insbesondere
26 croissance	28 bilancio	17 employment	33 mitgliedstaaten
32 devrait	29 occupazione	19 economic	34 jahr
33 taux	35 stati	21 budgetary	36 bip
34 mesures	39 politiche	22 public	43 eu
35 budgétaire	40 membri	23 policy	44 2001
36 œuvre	41 particolare	27 government	49 2002
39 publiques	43 mercati	30 member	54 wirtschaft
40 marchés	44 tasso	32 states	56 2000

de article masculin ou féminin défini en néerlandais. En traitant séparément les deux corpus et en construisant le tableau lexical par juxtaposition, nous évitons ces confusions.

41 états	45 disoccupazione	33 markets	61 finanzen
42 politiques	46 pubbliche	34 gdp	64 jahren
43 membres	48 dovrebbe	36 measures	65 wachstum
46 chômage	50 livello	38 unemployment	67 unternehmen

Ce qui est frappant et ce quelle que soit la langue est l'importance centrale accordée à l'emploi et au travail dans ce qui n'est somme toute qu'un texte économique qui n'est pas sensé se pencher sur les politiques sociales. Des analystes ont déjà soulevé cette dépendance des politiques sociales européennes par rapport aux compétences économiques de l'Union, cependant la répétition brute des lexèmes travail et emploi dans ces GOPE confirme cette subordination. Les Grandes orientations de politique économique traite moins d'industries (=économie réelle) que de l'emploi comme variable d'ajustement des politiques économiques et monétaires.

Pourtant outre cette symétrie entre les différentes versions linguistiques, il convient de souligner aussi quelques divergences : le rang d'*emploi* et de *travail* ne sont pas identiques, le français se distinguant de l'anglais et de l'italien. Or nous savons, par l'étude des GOPE en français, que *travail* est fortement lié à *marché* par le syntagme « marché du travail » alors que la majorité des utilisations de *emploi* le trouve lié à *taux* dans le segment « taux d'emploi ». L'allemand du fait des déclinaisons qui multiplient les formes lexicales d'un même lemme et de son utilisation de mots composés soulève des difficultés plus grandes pour l'analyse des formes les plus fréquentes puisque un lemme comme *arbeit* n'apparaît sous l'espèce de *arbeitsmarkt* qu'au 113ème rang, *arbeitsmarkt* est rendu dans les autres langues par un polyforme comme *marché du travail*.

4. Une analyse des correspondances d'un corpus multilingue

Nous avons réalisé l'analyse des correspondances du tableau lexical du corpus multilingue (quatre langues) des GOPE en ne retenant que les formes dont la fréquence était égale ou supérieure à 10 c'est à dire 6017 formes.

GOPE	Nbre d'occurrences	Nbre de formes	Nbre de formes F>=10
français	160464	6442	1487
Italien	152665	7067	1579
anglais	133422	4746	1354
allemand	126108	10075	1597
corpus	572659	28330	6017

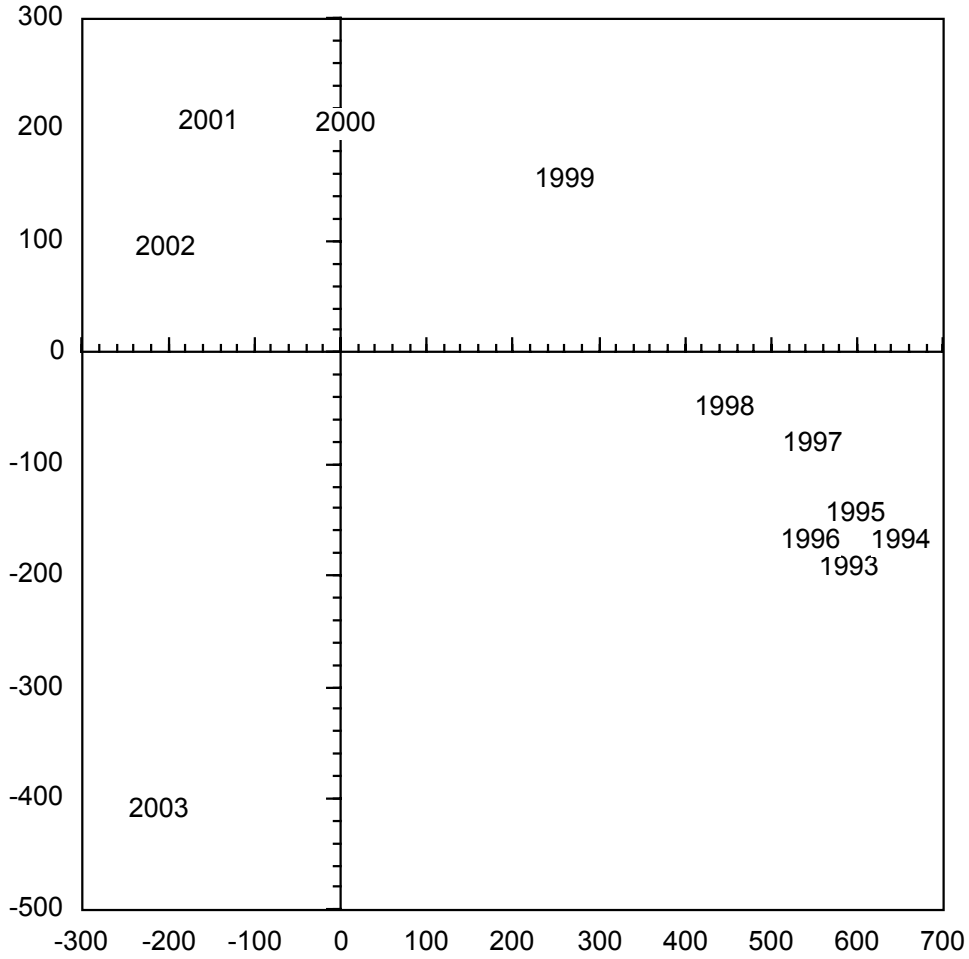
Le tableau lexical tronqué contient onze colonnes correspondant aux onze publications des GOPE de 1993 à 2003.

Le premier plan de l'AFC nous donne du nuage des GOPE une image en forme de parabole partant de 1993 en bas à gauche montant vers 2001 et redescendant ensuite dans le quadrant inférieur droit.

Nous savons (Deroubaix, 1997 ; Salem, 1997) qu'une telle image est la marque d'une série textuelle chronologique, c'est-à-dire d'un ensemble de discours dont le vocabulaire se renouvelle peu à peu, acquérant à chaque nouvelle publication une fraction de vocabulaire nouveau et en délaissant à chaque fois une autre fraction.

Il n'en reste pas moins qu'une lecture de ce plan factoriel du point de vue du vocabulaire ainsi acquis et rejeté est plein d'enseignement pour l'étude de la circulation lexicale et la modification des politiques énoncées. et qu'il est possible aussi d'examiner les ruptures et les écarts à la « parabole » dessinée sur le plan (Deroubaix et Gobin, 1987).

Ainsi, si l'ensemble des GOPE comprennent à la fois des recommandations générales (européennes) et des recommandations par pays, la manière dont sont réparties ces deux types de recommandations dans les textes et le poids qui leur est accordé ont varié dans le temps. Dans une première période de 1993 à 1998, l'organisation des recommandations est thématique (marché du travail, politique monétaire, …) et dans chaque thème sont reprises les recommandations générales et particulières.



À partir de 1999, le document comporte deux parties distinctes : l'une consacrée aux recommandations générales et l'autre aux recommandations détaillées pays par pays. Le poids des recommandations particulières est croissant. L'aspect normatif, et disciplinaire prend plus d'importance. Ce qui explique la rupture visible sur le plan factoriel entre le groupe 1999-1998 et celui de 1999 à 2002.

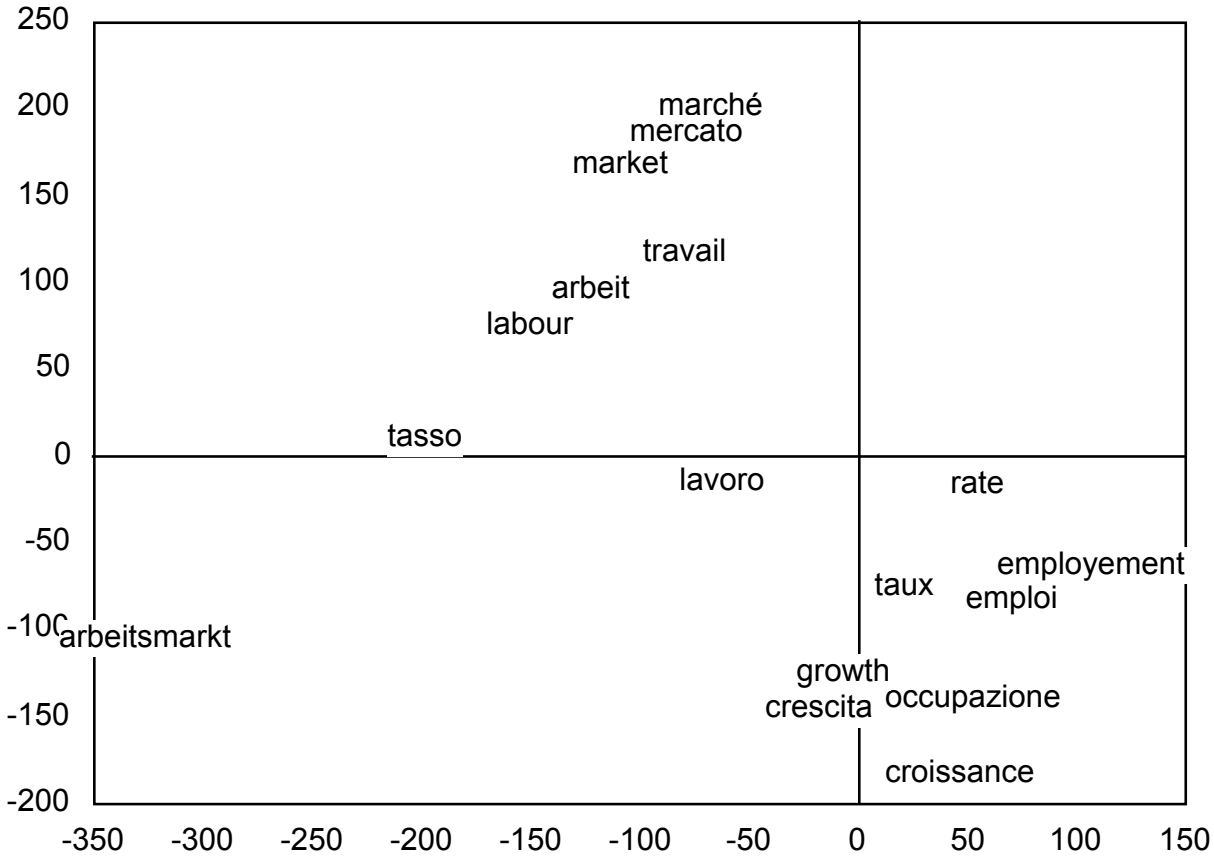
Avec cette AFC nous disposons d'un référentiel commun dans lequel il va être possible de projeter aussi les quatre versions linguistiques du corpus. Nous avons à cet effet ajouté aux 6017 lignes du tableau lexical, quatre lignes supplémentaires correspondant à la somme des occurrences des formes dans chaque langue (réduite de la même manière aux formes appa-

raissant au moins dix fois). La projection des profils « langue » nous montre que globalement les versions anglaise, française et italienne se démarquent peu du profil moyen.

Il n'en est pas de même du profil allemand, laissant à penser qu'au moins un découpage des formes composées devrait être effectué avant d'élaborer une comparaison.

Une exploration du lexique des GOPE en français (Gobin, 2003) a mis en évidence l'aspect central que ces documents donnent au « marché du travail », au « taux d'emploi » et à la « croissance », alors que l'on aurait pu s'attendre dans un texte intitulé « Grandes orientations de politique économique » à voir mis en valeur l'investissement, l'entreprise, etc. Nous nous sommes dès lors intéressé à la représentation des formes *marché*, *emploi*, *travail*, *croissance* et *taux,* et de leur traduction courante dans le corpus multilingue.

On constate tout d'abord que ces formes restent sensiblement groupées, d'une part le groupe *croissance*, *taux* et *emploi*, de l'autre le groupe *marché* et *travail*. On voit aussi qu'il y a eu un déplacement de l'accent mis sur ces deux groupes de formes entre la première période 93-98 et la seconde. Ceci ne signifie pas un désintérêt pour la croissance du taux d'emploi mais une montée relative de l'intérêt pour la réforme du « marché du travail ».



Cependant, cette homogénéité des représentations des formes dans les différentes langues est loin d'être absolue, *lavoro* mais plus encore *tasso* s'écartent du modèle moyen laissant supposer une utilisation différente.

5. Conclusions provisoires

Cette communication est une première approche, elle est partie d'un projet plus vaste d'étude de la circulation lexicale entre institutions internationales et nationales. Le corpus GOPE a été

choisi comme exemple de texte fondamental de politique économique et sociale de l'UE. En tant que tel il fait partie d'un ensemble de textes à partir desquels nous tenterons d'établir un glossaire de la protection sociale en Europe. Cela nécessite une exploration approfondie de ce corpus. La réalisation d'une typologie multilingue des formes et celle d'une décomposition de l'effet chronologique, selon une méthode mise au point dans (Deroubaix, 1997), sont en cours.

D'une certaine manière ces résultats encore préliminaires pourraient conforter le choix souvent opéré de travailler sur les versions unilingues de textes internationaux en vue d'explorer leur vocabulaire puisque les images des parties du corpus semblent relativement bien coïncider dans l'espace factoriel de référence. Pourtant les écarts observés entre les représentations des formes (pourtant parmi les plus fréquentes) selon la langue impose d'être prudent dans l'exportation pure et simple des conclusions tirées de l'étude d'une seule version linguistique. Les capacités informatiques actuelles et les outils statistiques permettant de traiter les corpus multilingues, il nous semble important de sauter le pas et travailler lorsque c'est possible sur les corpus les plus complets possibles. Nous appliquons ainsi d'ailleurs un des principes de l'étude sur corpus : l'exhaustivité.

Annexe

Pour la lisibilité des plans factoriels certains points ont été légèrement déplacés. Voici donc les coordonnées exactes des 11 GOPE.

J1	QLT	POID	INR	1#F	COR	CTR	2#F	CTR
1993	365	16	71	603	262	75	-178	2
1994	423	17	76	648	311	95	-190	1
1995	498	24	79	592	339	108	-172	1
1996	411	21	68	545	302	83	-174	1
1997	420	34	90	549	367	133	-86	5
1998	418	51	84	447	391	133	-55	3
1999	933	98	109	260	196	87	150	3
2000	991	137	93	5	0	0	201	147
2001	995	197	89	-154	170	61	203	533
2002	992	209	97	-203	288	112	89	296
2003	999	196	145	-211	194	113	-414	8

Références

Deroubaix J.-Cl. (1998). Deux langues pour une même politique : étude d'un corpus bilingue parallèle de textes politiques. In *Actes des JADT 1998.*

Deroubaix J.-Cl. (1997). *Les déclarations gouvernementales en Belgique (1944-1992).* Étude de lexicométrie politique. Thèse en sciences du langage, Sorbonne nouvelle Paris 3.

Friot B. (2003). Resource regime reforms and worker status. In Clasquin B. et Moncel N. (Eds), *Social Rights over Financial Resources : Issues for the Future of Employment and Social Protection in Europe*, Publication of the TSER European Network "Social Construction of Employment". Editions PIE-Peter Lang.

Gobin C. et Deroubaix J.-Cl. (1989). Les temps sociaux et le discours politique. Repérage de la notion de temps dans les Déclarations gouvernementales belges. *Histoire et Mesure*, vol. (3-4). Éd. du CNRS : 147-171.

Gobin C., Coron G. et Dufresne A. (2003). The European Union and resources restructuration : employment, pension and wage. In Clasquin B. et Moncel N. (Eds), *Social Rights over Financial Resources : Issues for the Future of Employment and Social Protection in Europe*. Publication of the TSER European Network "Social Construction of Employment". Editions PIE-Peter Lang.

Gobin C. (2003). L'Union européenne : l'institution politique est évanescente, le syndicat est un partenaire, le travailleur un problème, où est passé l'acteur ? Communication au colloque *The Economic's Representation of Actor at Work*. Université des Sciences et Techniques de Lille.

Kraif O. (2001). *Constitution et exploitation de bi-textes pour l'Aide à la traduction.* Thèse en sciences du langage. Université de Nice.

Martinez W. et Zimina M. (2002). Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. In *Actes des JADT 2002.*

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.