Traiter le multilinguisme dans les discours politiques

Possibilités et limites d'une analyse lexicométrique dans un corpus composé des textes de différents langues

Ronny Scholz,
Université Paris XII
Université de Magdebourg (Allemagne)

Comment peut-on analyser un corpus multilingue avec des outils lexicométriques?

Pourquoi faire?

- pour analyser un champs discursif international. Présupposé que le champs discursif ne s'arrête pas à la frontière langagière.
- Comparaison directe des textes appartenant au même discours dans différents pays.
- Mesurer les caractéristiques lexicales d'un corpus multilingue représentant un certain discours indépendamment des langues contenues.

Le corpus soit établit à partir des textes représentants un discours qui concerne un sujet international. Les caractéristiques lexicales du corpus sont représentatif pour le corpus et permettent des conclusions concernant un discours international.

3 démarches possibles pour un traitement lexicométrique du multilinguisme

I. Approche multilinguistique

- Composer un seul corpus avec des textes de différentes langues pour que l'analyse lexicométrique soit effectuée dans l'ensemble des textes multilingues

II. Approche morphosyntaxique à l'aide du logiciel *TreeTagger.*

- Composer un corpus à la base des étiquètes représentants la structure morphosyntaxique du corpus. - Analyser la structure morphosyntaxique dans les textes des différentes langues. - Unifier l'étiquetage morphosyntaxique des différentes langues.

III. Approche de mesure croisée dans des corpus des différents langues

- Composer plusieurs corpus chacun contenant des textes d'une langue particulière. - Comparer les résultats d'analyse lexicométrique des différents corpus dans la mesure possible.

Ensemble des textes qui forment le corpus pour la période 1979 – 2004

corpus allemand:

REP*, CDU, CSU, FDP, Die Grünen, SPD, PDS*

corpus français:

FN, MPF, RPF, UMP (RPR, DL), UDF*, Les Verts, PS, PRG*, PCF, LO*

corpus britannique:

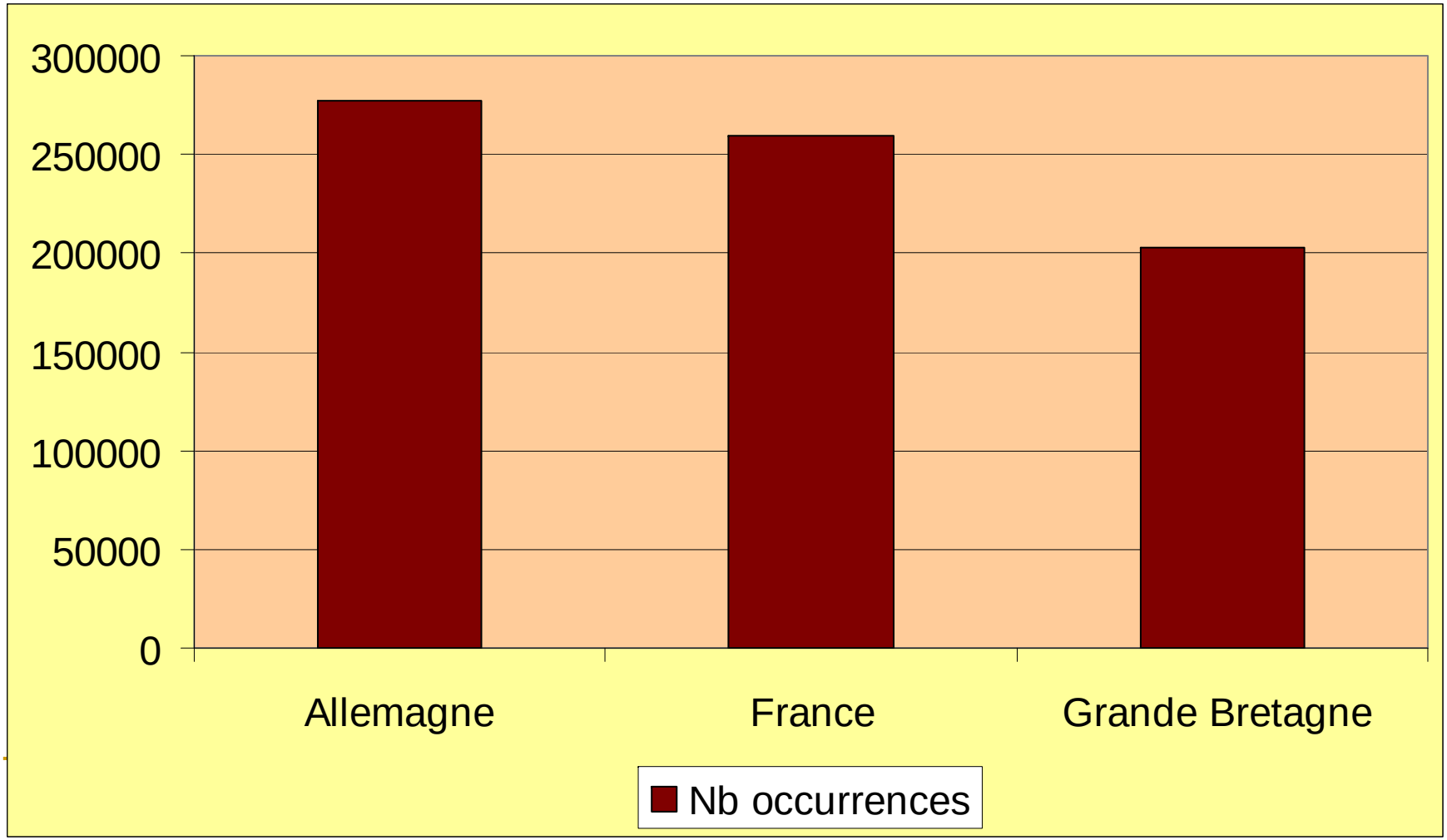
UKIP, CONS, LDP, LAB, GREENS, PC*, SNP*

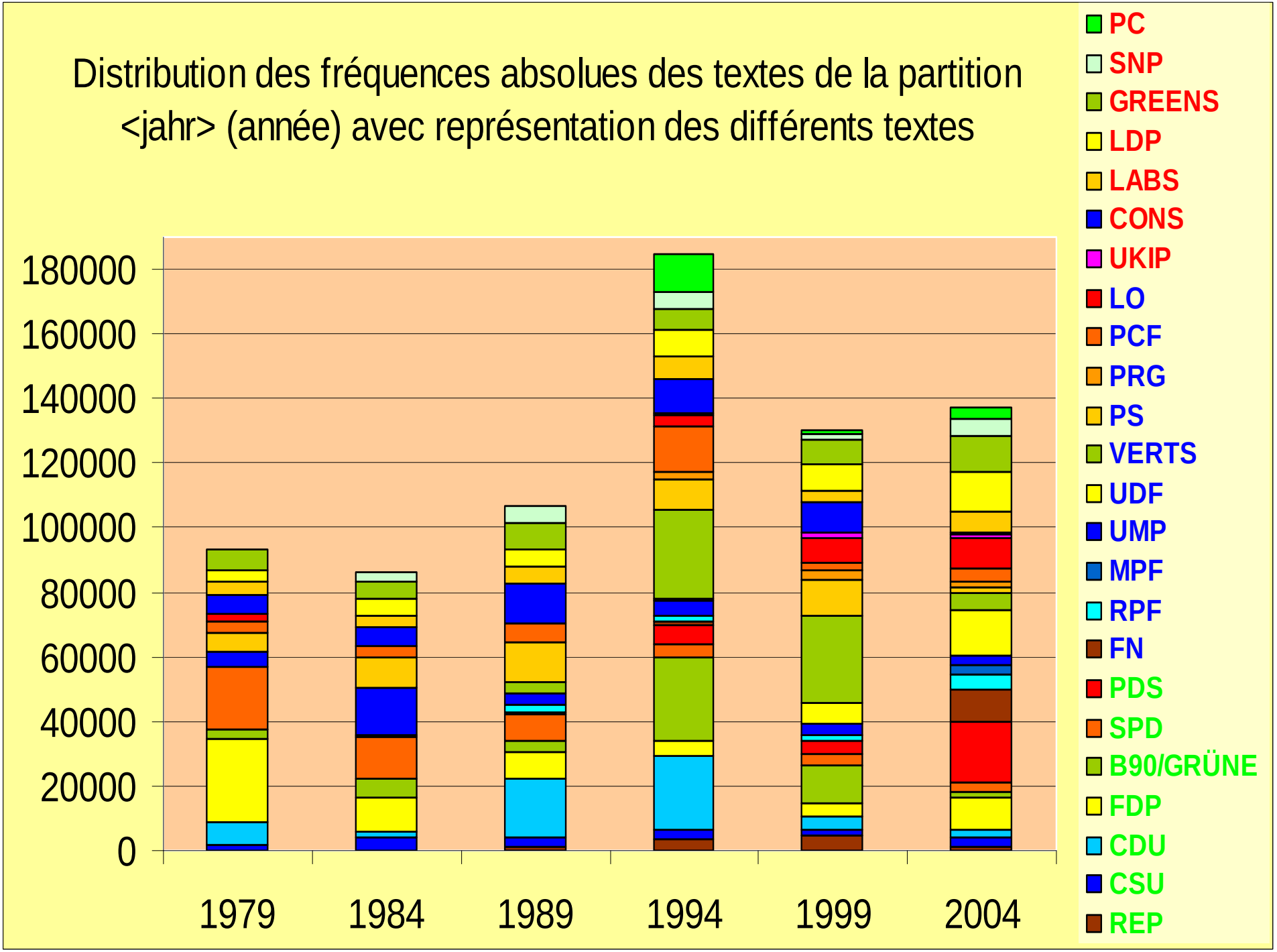
Caractéristiques du corpus

Principales caractéristiques lexicométriques

Nombre des occurrences	**805195**
Nombre des formes	**47468**
Fréquence maximale	**16391**
Nombre des hapax	**21976**
Nombre des textes	**121**

Distribution des fréquences absolues des textes de la partition <land> dans le corpus de programmes des Élections Européennes




Distribution des fréquences absolues des textes de la partition <jahr> (année) avec représentation des différents textes
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
1979
1984
1989
1994
1999
2004
PC
SNP
GREENS
LDP
LABS
CONS
UKIP
LO
PCF
PRG
PS
VERTS
UDF
UMP
MPF
RPF
FN
PDS
SPD
B90/GRÜNE
FDP
CDU
CSU
REP

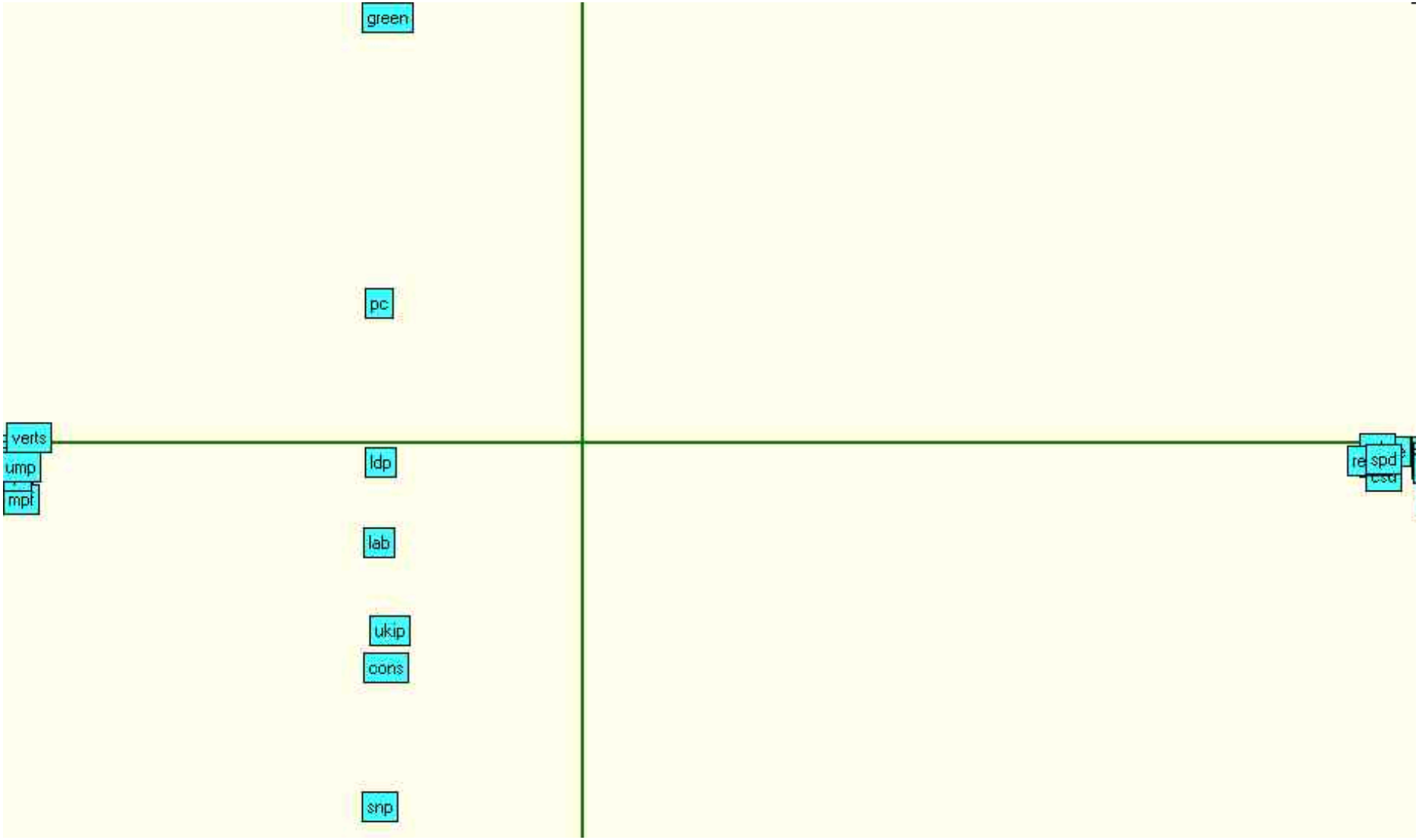
Mélanger les langues à condition que :

- le nombre d'occurrences soit équilibré entre les différentes langues
- la comparaison porte sur les caractéristiques sémantiques et pas sur les caractéristiques grammaticaux

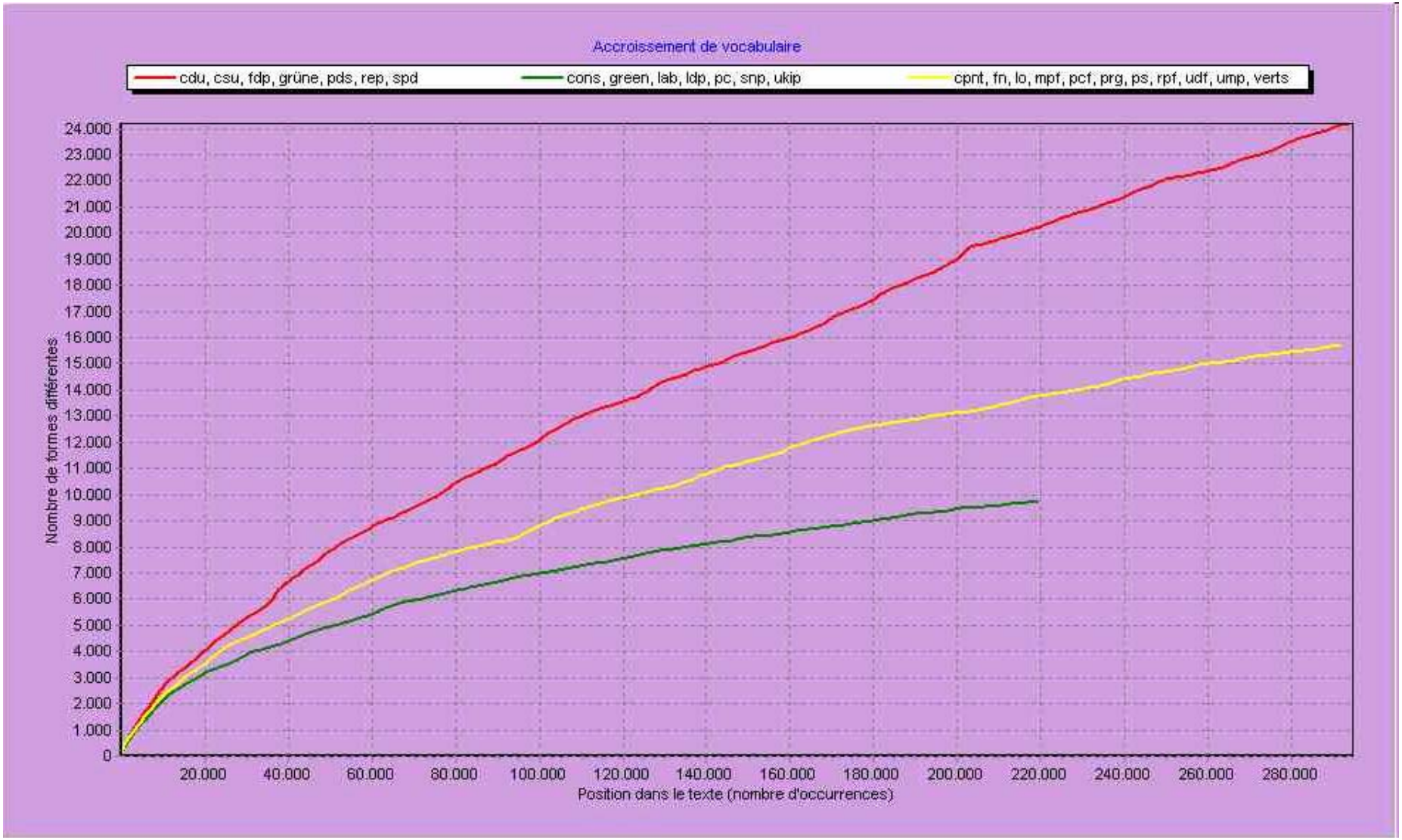
  (Une différence des caractéristique entre des différentes langues est évident et une analyse lexicométrique n'apporterait pas des résultats révélateurs. Par exemple Europe de l'est vs. Eastern Europe vs. Osteuropa)

Solution: Groupe de formes qui porte sur une caractéristiques sémantique – par exemple: Tgen *europ*+

Analyse Factorielle des Correspondances (AFC) dans la partition *parti*



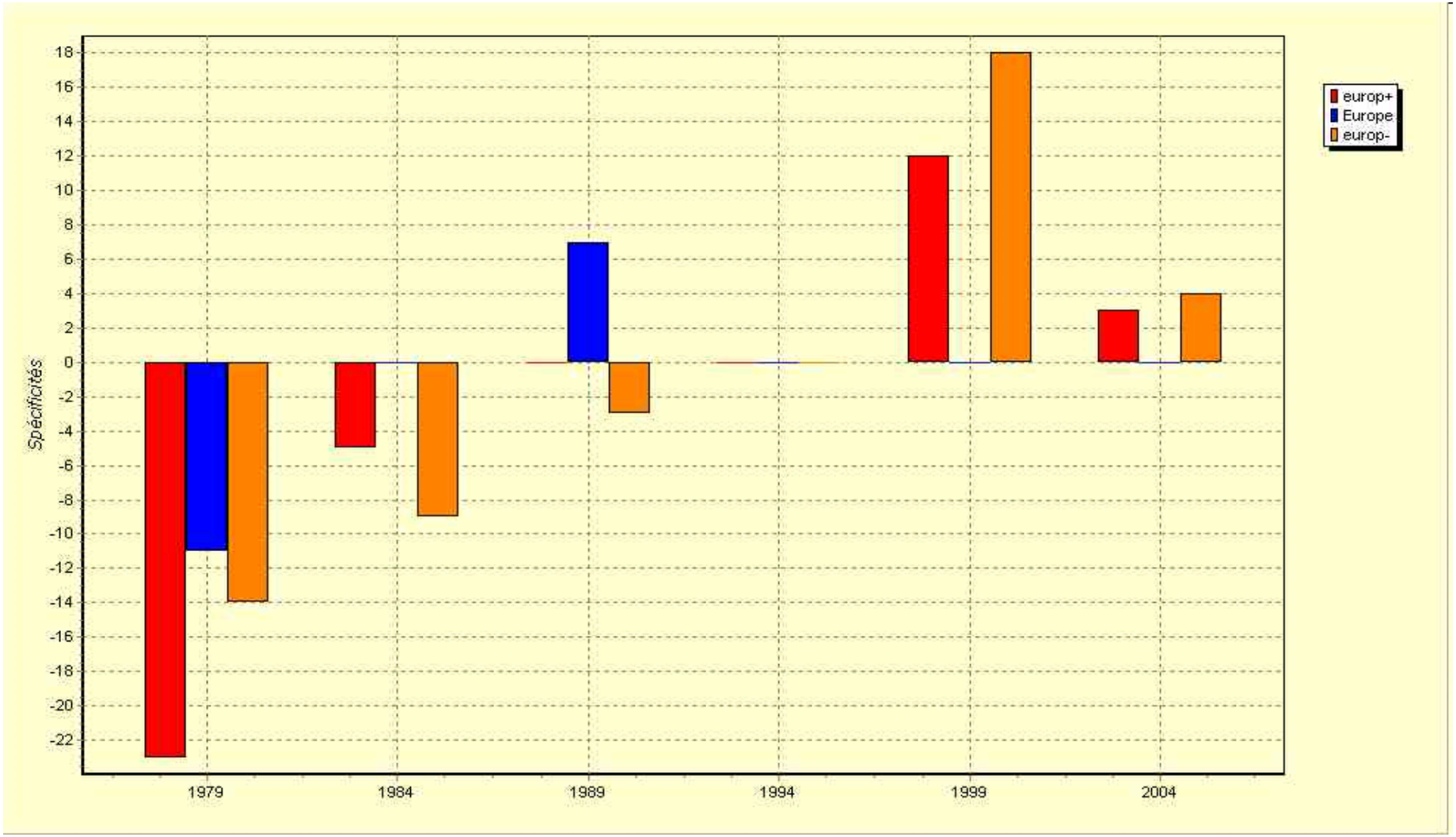
Accroissement vocabulaire
Comparaison des différents langues



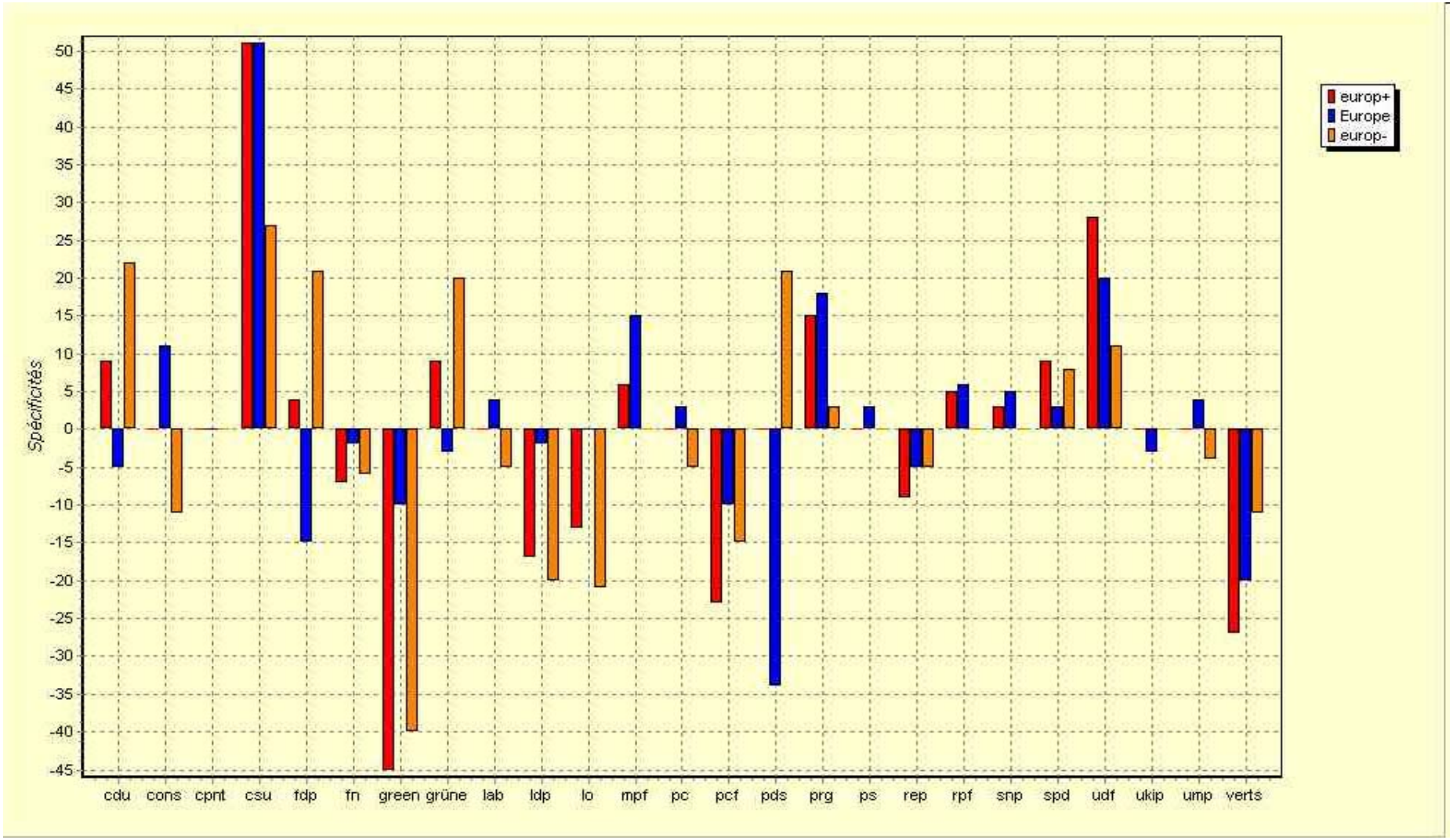
La Tgen europ+ (les premières formes)

Tgen *europ+*	
Forme	**Fréquence absolute**
europe	4080
europäischen	2362
european	1945
europa	1533
europäische	1371
européenne	1334
européen	903
européens	544
europas	473
européennes	371
europäischer	201
europol	73
europäisches	72
europäer	69
osteuropa	56
osteuropas	52
westeuropäischen	52
europapolitik	47
osteuropäischen	40
gesamteuropäischen	36
westeuropa	36
gesamteuropäische	31

Spécificités des Tgen ***europ+***; ***europ-***; ***Europa/Europe*** dans la partition ***année***



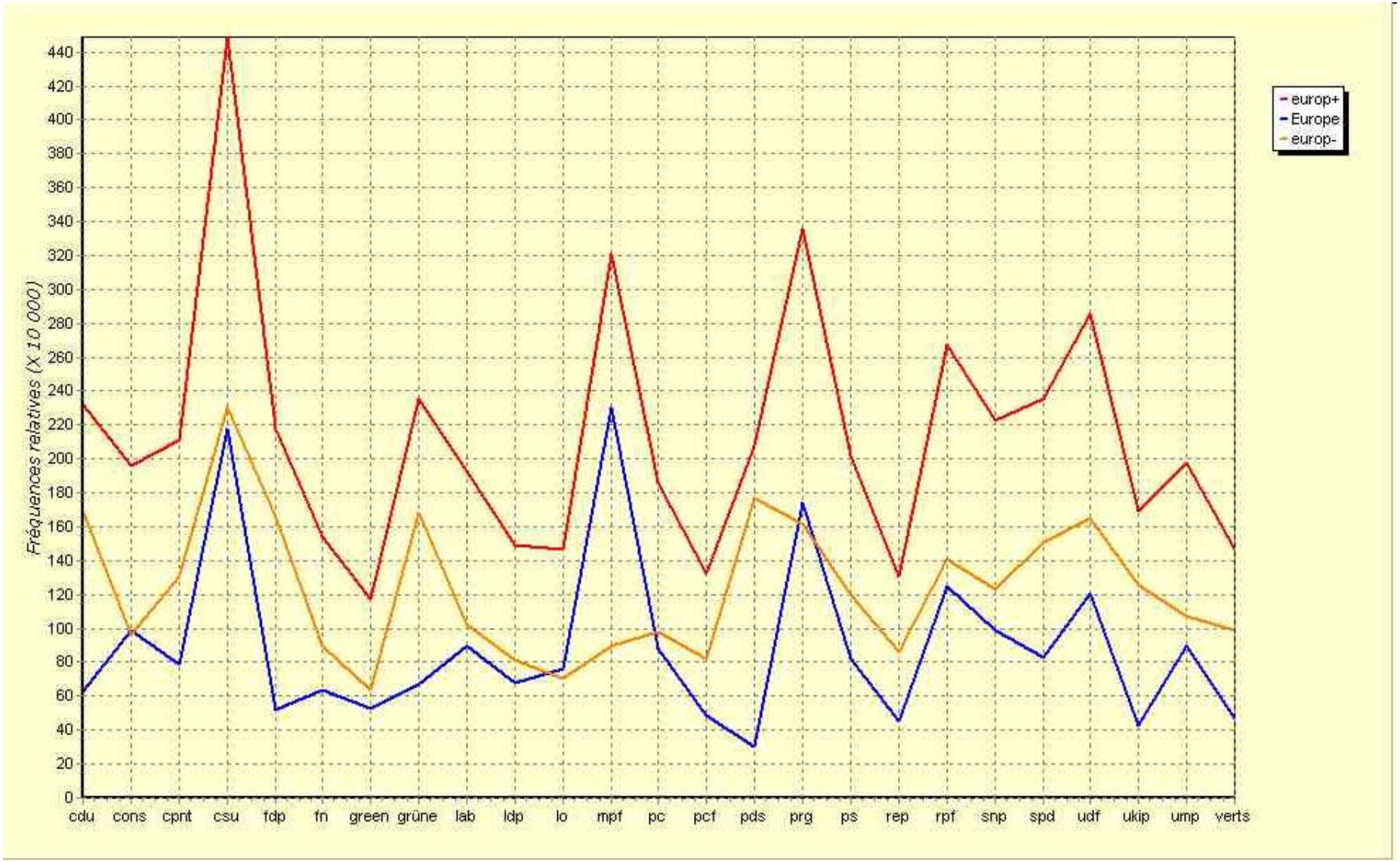
Spécificités des Tgen ***europ+***; ***europ-***; ***Europa/Europe*** dans la partition ***parti***



Cooccurrences de Tgen *europ+ (liste élaguée)*

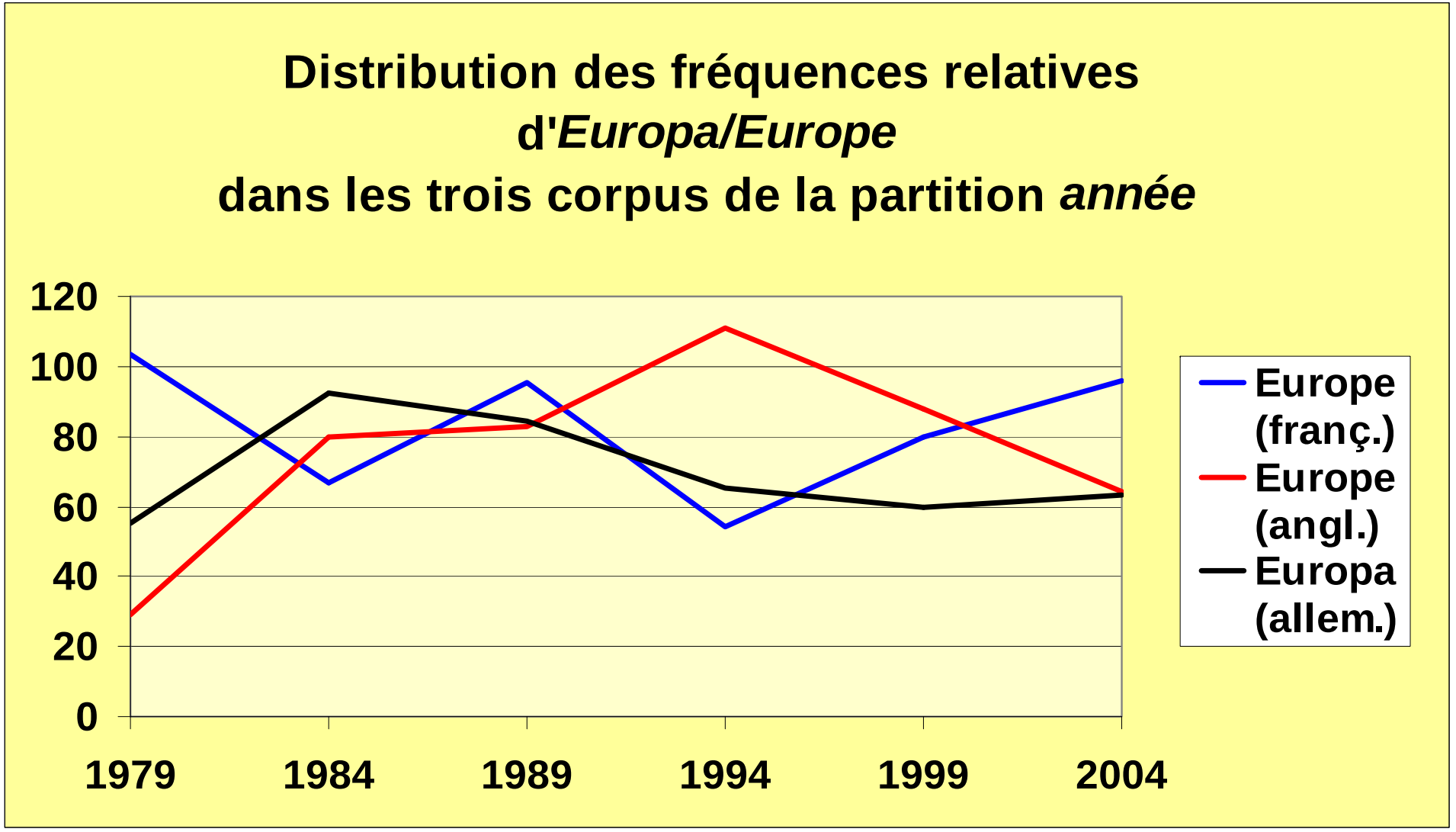
Forme	Frq. Tot.	Fréquence	Coeff.
union	2445	2054	***
parliament	560	406	49
parlement	294	243	47
une	3707	1944	39
construction	227	189	38
et	7978	3841	31
la	10431	4929	30
d	5209	2574	29
einigung	111	99	26
pour	2924	1494	25
ein	1955	1040	25
parlaments	112	97	23
élections	116	97	21
peuples	227	161	19
eastern	59	55	17
un	2926	1440	17
starkes	46	45	17
voulons	160	117	16
dimension	146	109	16
ebene	233	156	15
gemeinschaft	1131	602	15
staaten	470	276	14
au	1905	956	14
espace	134	96	13
nations	294	183	13
institutions	345	210	13
centrale	72	60	13
modèle	107	79	12
citoyens	245	156	12
conseil	280	172	12
bürger	419	242	12
geeintes	30	30	12
doit	886	471	12
zentralbank	35	33	11
sécurité	296	177	11

Fréquence relative des Tgen ***europ+***; ***europ-***; ***Europa/Europe*** dans la partition ***parti***



Fréquence relative des Tgen ***europ+***; ***europ-***; ***Europa/Europe*** dans la partition ***année***

Comparaison avec les Fréquences relatives dans trois corpus séparés



Limites d'une analyse lexicométriques dans un corpus composé des différents langues

- L'Analyse Factorielle des Correspondances n'est pas utile dans un corpus multilingue.
- Les mesures des spécificités et des cooccurrences peut être utile à condition que la distribution des occurrences soit équilibrés dans les partis comparé. Sinon le vocabulaire d'une partie avec une langue moins riche ou une longueur plus longues soit surreprésentée.
- En comparaison avec un corpus unilingue les spécificités des formes particuliers dans un corpus multilingues sont plus extrêmes du fait qu'on trait un corpus multilingue.

  Une forme qui est spécifique chez un certain locuteur ne peut pas apparaitre chez un autre locuteurs à cause de la différence langagière. Donc forcement la forme doit être surreprésenté plus extrêmement que dans un corpus unilingue.

→ Solution: Tgen multilingue

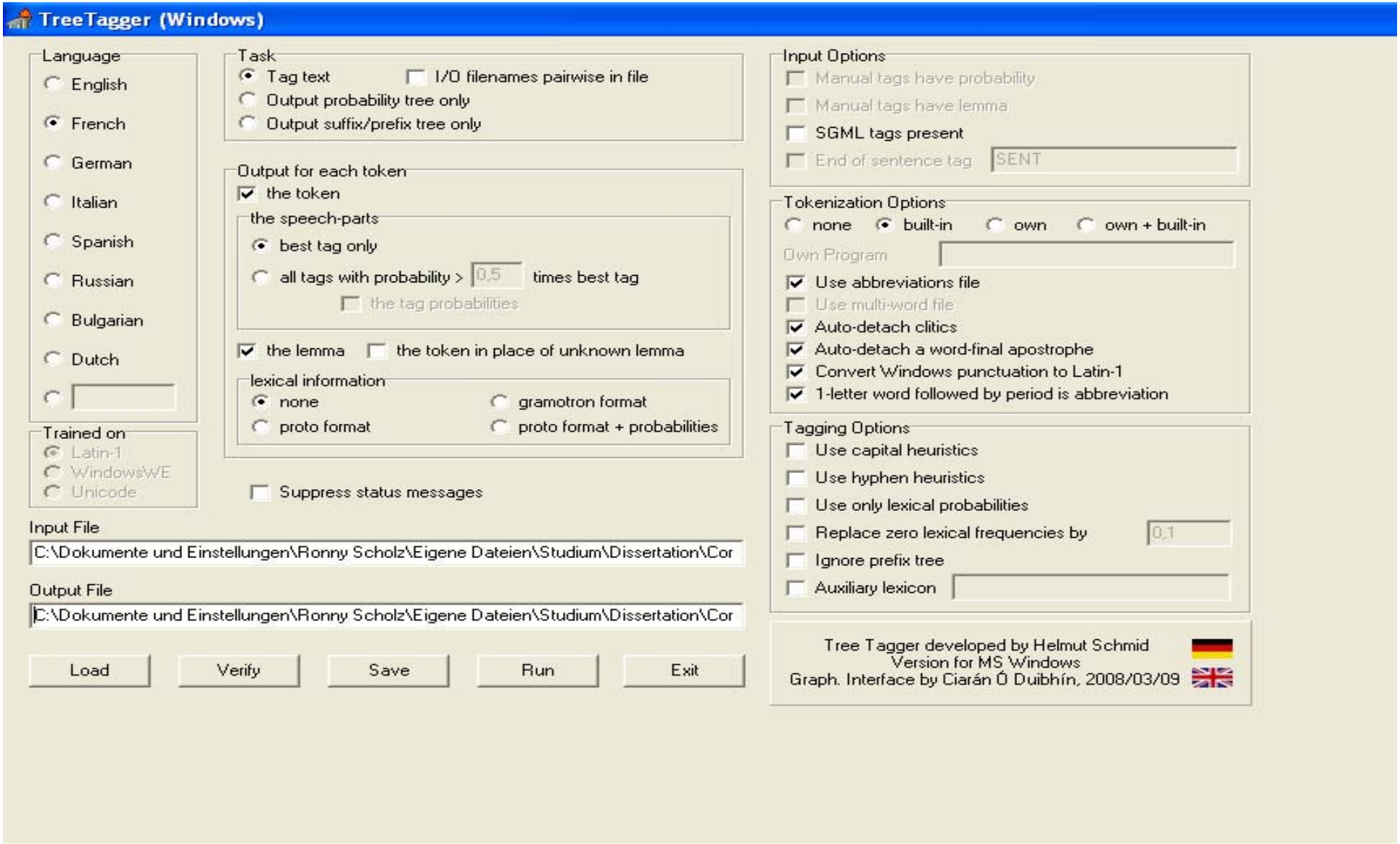
- Des mesures des Fréquences relatives dans des corpus séparés peut relever des résultats plus détaillés.
- L'analyse lexicométrique dans un corpus multilingue est utile pour une analyse d'un discours international mais elle n'est pas utile pour une comparaison des discours des différents pays.

Merci pour votre attention!

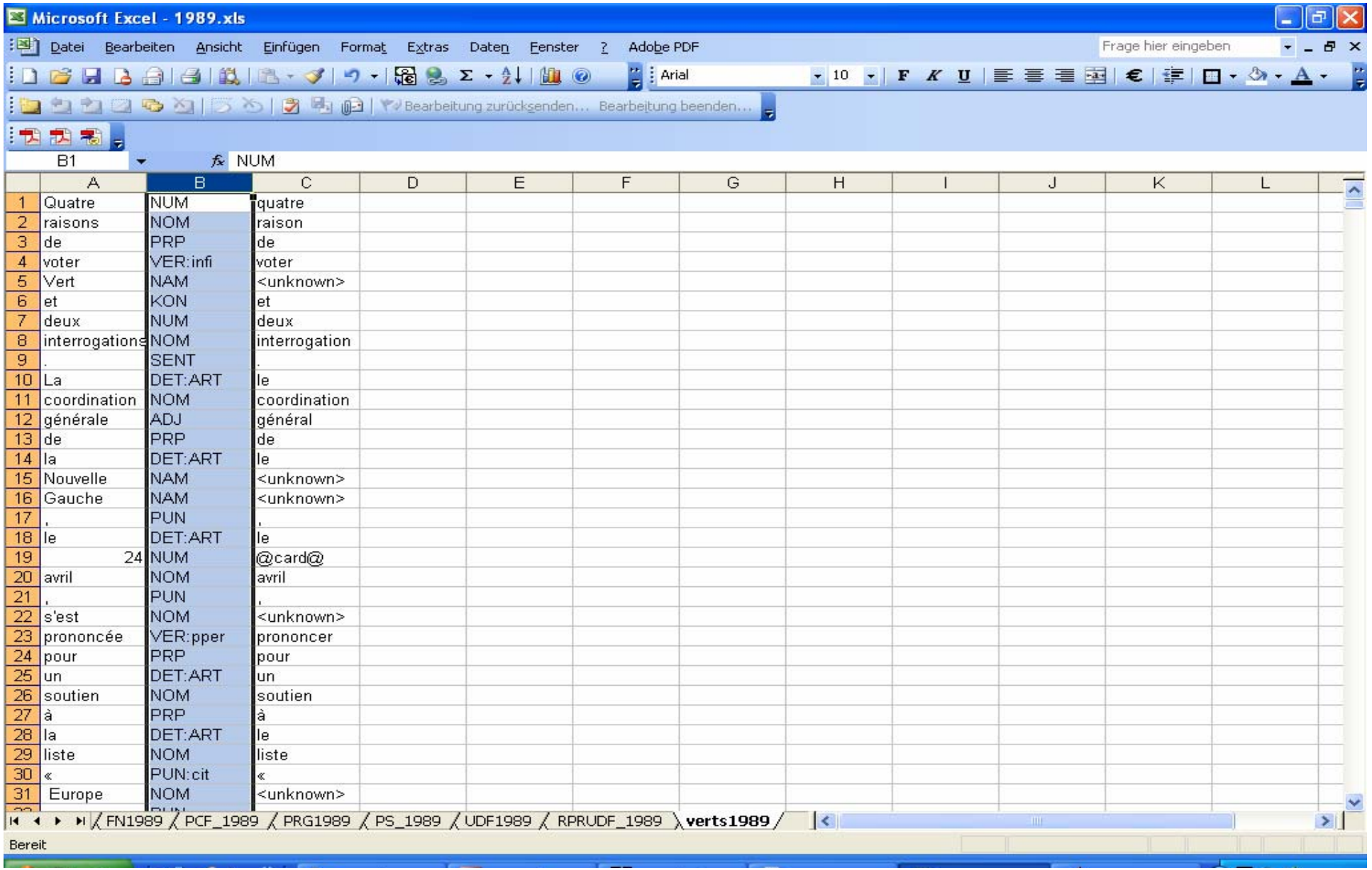
Démarches pour former un corpus multilingue expérimental

- Tree-Tagger
- Microsoft Excel
- Unification des tags et ses problèmes
- Lexico3

TreeTagger

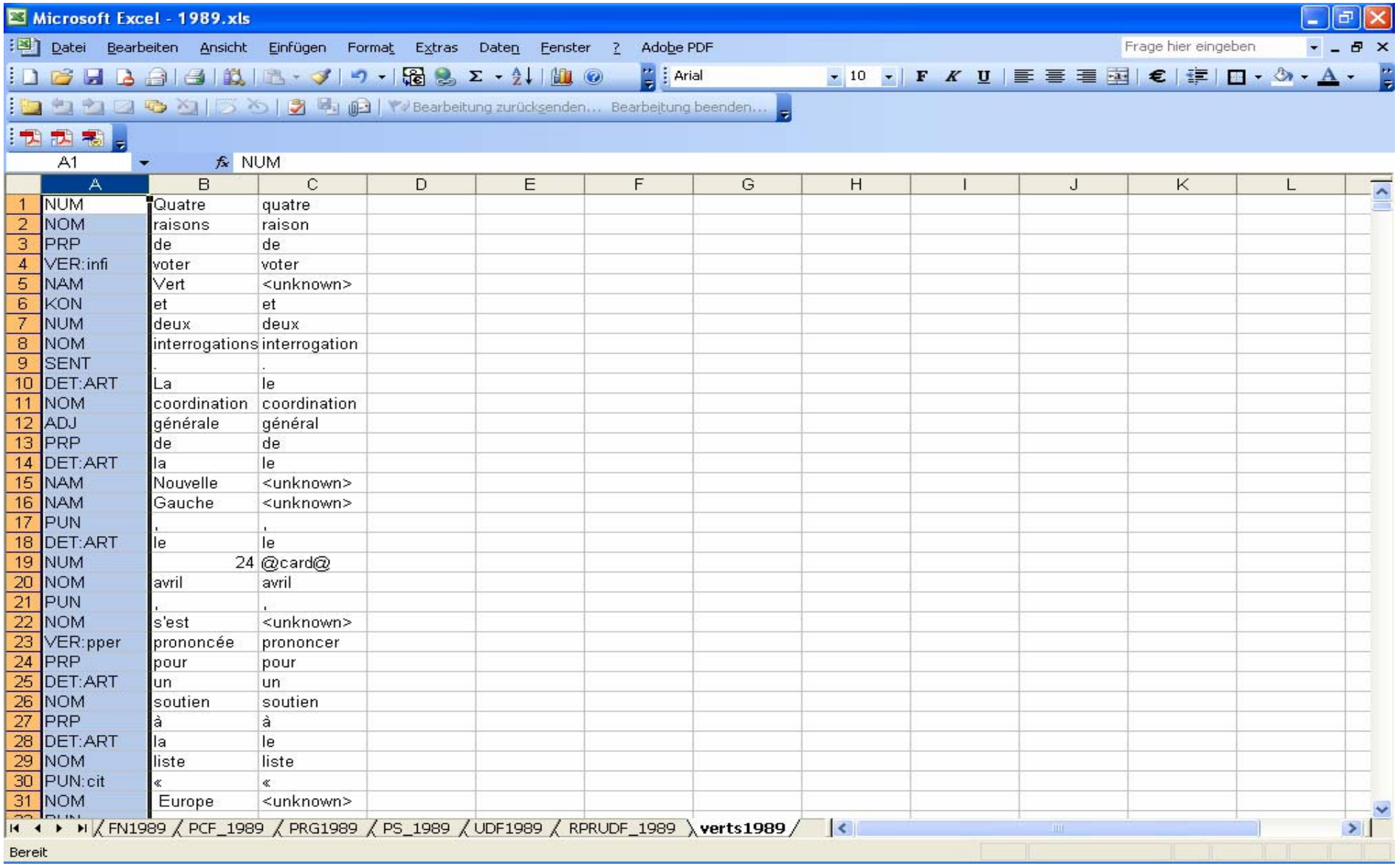

TreeTagger (Windows)
Language
English
French
German
Italian
Spanish
Russian
Bulgarian
Dutch
Trained on
Latin-1
WindowsWE
Unicode
Task
Tag text
I/O filenames pairwise in file
Output probability tree only
Output suffix/prefix tree only
Output for each token
the token
the speech-parts
best tag only
all tags with probability > 0,5 times best tag
the tag probabilities
the lemma
the token in place of unknown lemma
lexical information
none
gramotron format
proto format
proto format + probabilities
Suppress status messages
Input File
C:\Dokumente und Einstellungen\Ronny Scholz\Eigene Dateien\Studium\Dissertation\Cor
Output File
C:\Dokumente und Einstellungen\Ronny Scholz\Eigene Dateien\Studium\Dissertation\Cor
Load
Verify
Save
Run
Exit
Input Options
Manual tags have probability
Manual tags have lemma
SGML tags present
End of sentence tag
SENT
Tokenization Options
none
built-in
own
own + built-in
Own Program
Use abbreviations file
Use multi-word file
Auto-detach clitics
Auto-detach a word-final apostrophe
Convert Windows punctuation to Latin-1
1-letter word followed by period is abbreviation
Tagging Options
Use capital heuristics
Use hyphen heuristics
Use only lexical probabilities
Replace zero lexical frequencies by 0,1
Ignore prefix tree
Auxiliary lexicon
Tree Tagger developed by Helmut Schmid
Version for MS Windows
Graph. Interface by Ciarán Ó Duibhín, 2008/03/09

Corpus de vérification des tags



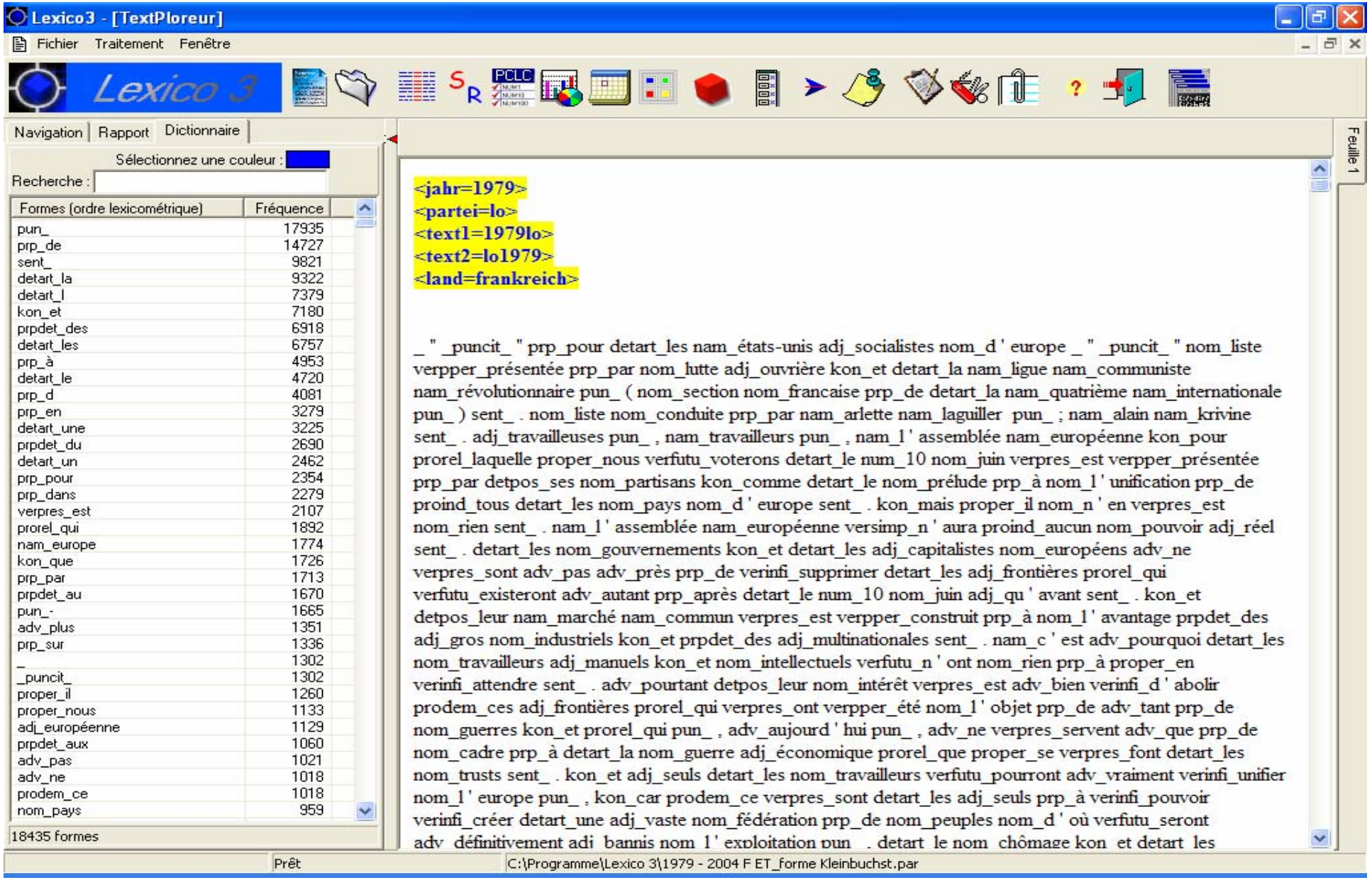
	A	B	C
1	Quatre	NUM	quatre
2	raisons	NOM	raison
3	de	PRP	de
4	voter	VER:infi	voter
5	Vert	NAM	<unknown>
6	et	KON	et
7	deux	NUM	deux
8	interrogations	NOM	interrogation
9	.	SENT	.
10	La	DET:ART	le
11	coordination	NOM	coordination
12	générale	ADJ	général
13	de	PRP	de
14	la	DET:ART	le
15	Nouvelle	NAM	<unknown>
16	Gauche	NAM	<unknown>
17	,	PUN	,
18	le	DET:ART	le
19	24	NUM	@card@
20	avril	NOM	avril
21	,	PUN	,
22	s'est	NOM	<unknown>
23	prononcée	VER:pper	prononcer
24	pour	PRP	pour
25	un	DET:ART	un
26	soutien	NOM	soutien
27	à	PRP	à
28	la	DET:ART	le
29	liste	NOM	liste
30	«	PUN:cit	«
31	Europe	NOM	<unknown>

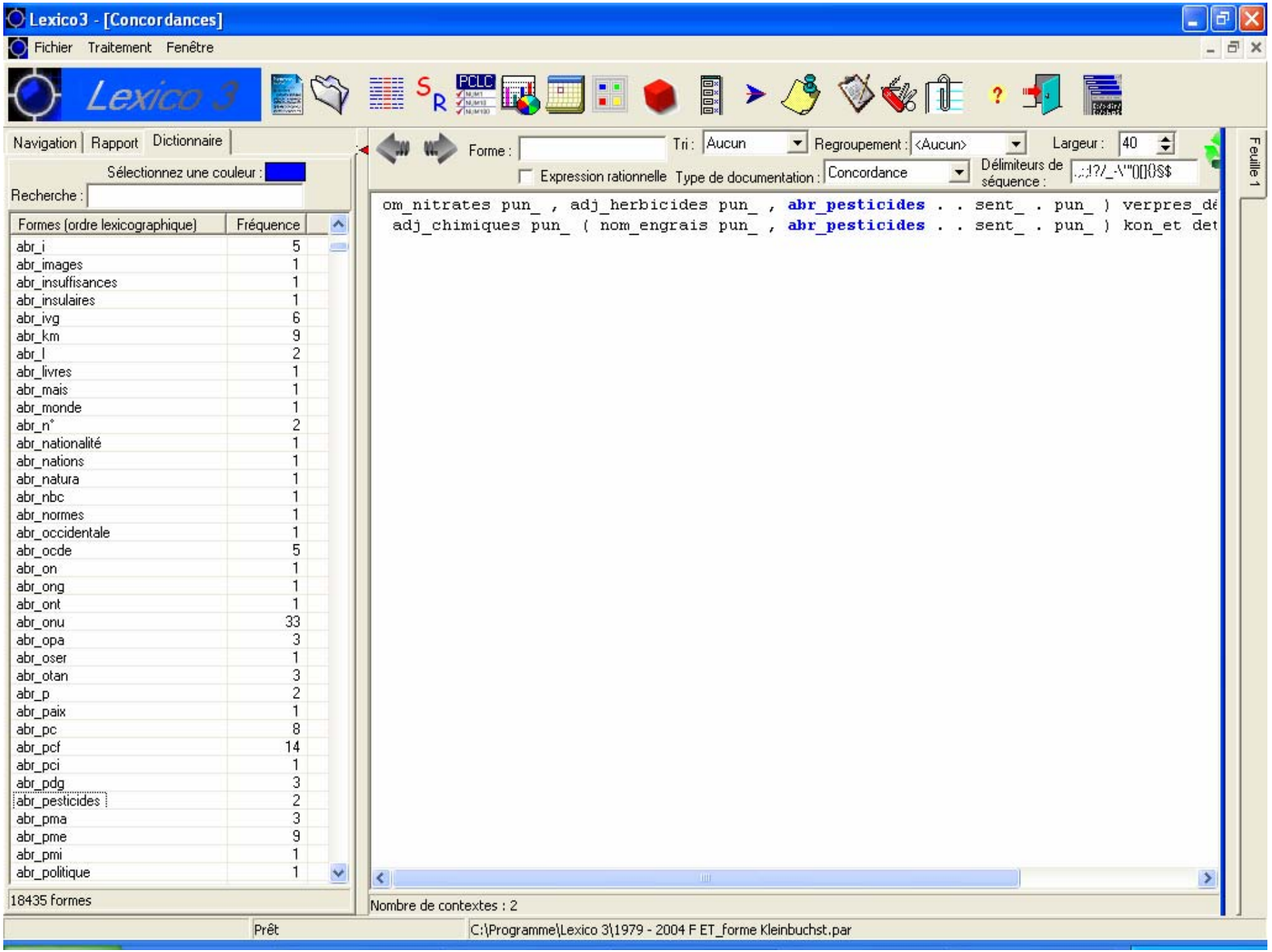
Corpus de vérification des tags



	A	B	C
1	NUM	Quatre	quatre
2	NOM	raisons	raison
3	PRP	de	de
4	VER:infi	voter	voter
5	NAM	Vert	<unknown>
6	KON	et	et
7	NUM	deux	deux
8	NOM	interrogations	interrogation
9	SENT	.	.
10	DET:ART	La	le
11	NOM	coordination	coordination
12	ADJ	générale	général
13	PRP	de	de
14	DET:ART	la	le
15	NAM	Nouvelle	<unknown>
16	NAM	Gauche	<unknown>
17	PUN	,	,
18	DET:ART	le	le
19	NUM	24	@card@
20	NOM	avril	avril
21	PUN	,	,
22	NOM	s'est	<unknown>
23	VER:pper	prononcée	prononcer
24	PRP	pour	pour
25	DET:ART	un	un
26	NOM	soutien	soutien
27	PRP	à	à
28	DET:ART	la	le
29	NOM	liste	liste
30	PUN:cit	«	«
31	NOM	Europe	<unknown>

Corpus de vérification des tags




Lexico3 - [Concordances]
Fichier Traitement Fenêtre
Lexico 3
Navigation Rapport Dictionnaire
Sélectionnez une couleur :
Recherche :
Formes (ordre lexicographique) Fréquence
abr_i 5
abr_images 1
abr_insuffisances 1
abr_insulaires 1
abr_ivg 6
abr_km 9
abr_l 2
abr_livres 1
abr_mais 1
abr_monde 1
abr_n° 2
abr_nationalité 1
abr_nations 1
abr_natura 1
abr_nbc 1
abr_normes 1
abr_occidentale 1
abr_ocde 5
abr_on 1
abr_ong 1
abr_ont 1
abr_onu 33
abr_opa 3
abr_oser 1
abr_otan 3
abr_p 2
abr_paix 1
abr_pc 8
abr_pcf 14
abr_pci 1
abr_pdg 3
abr_pesticides 2
abr_pma 3
abr_pme 9
abr_pmi 1
abr_politique 1
18435 formes
Forme :
Tri : Aucun
Regroupement : <Aucun>
Largeur : 40
Expression rationnelle
Type de documentation : Concordance
Délimiteurs de séquence : .,;:!?/_-\'"()[]{}§$
om_nitrates pun_ , adj_herbicides pun_ , abr_pesticides . . sent_ . pun_) verpres_dé
adj_chimiques pun_ (nom_engrais pun_ , abr_pesticides . . sent_ . pun_) kon_et det
Nombre de contextes : 2
Feuille 1
Prêt
C:\Programme\Lexico 3\1979 - 2004 F ET_forme Kleinbuchst.par

Unification des tags - quelques cas douteux

Langue taggée:	française	anglaise	allemande	Synthèse
Article	DET:ART	DT	ART	ARTPROIND
Pronom indéfini	PRO:IND	DT PDT	PIS PIAT PIDAT	
Pronom possessif	DET:POS (ma, leur) PRO:POS (mien, vôtre)	PPS (POS) WPS	PPOSS PPOSAT	PROPOS
Pronom démonstratif	PRO:DEM	(PDT)	PDS PDAT	PRODEM
Mot étranger		(FW)	(FM)	effacé
Parti verbale séparé du verbe [er kommt] an			(PTKVZ)	effacé
Particule d'infinitif avec zu [Er begann] zu [laufen].			(PTKZU)	effacé

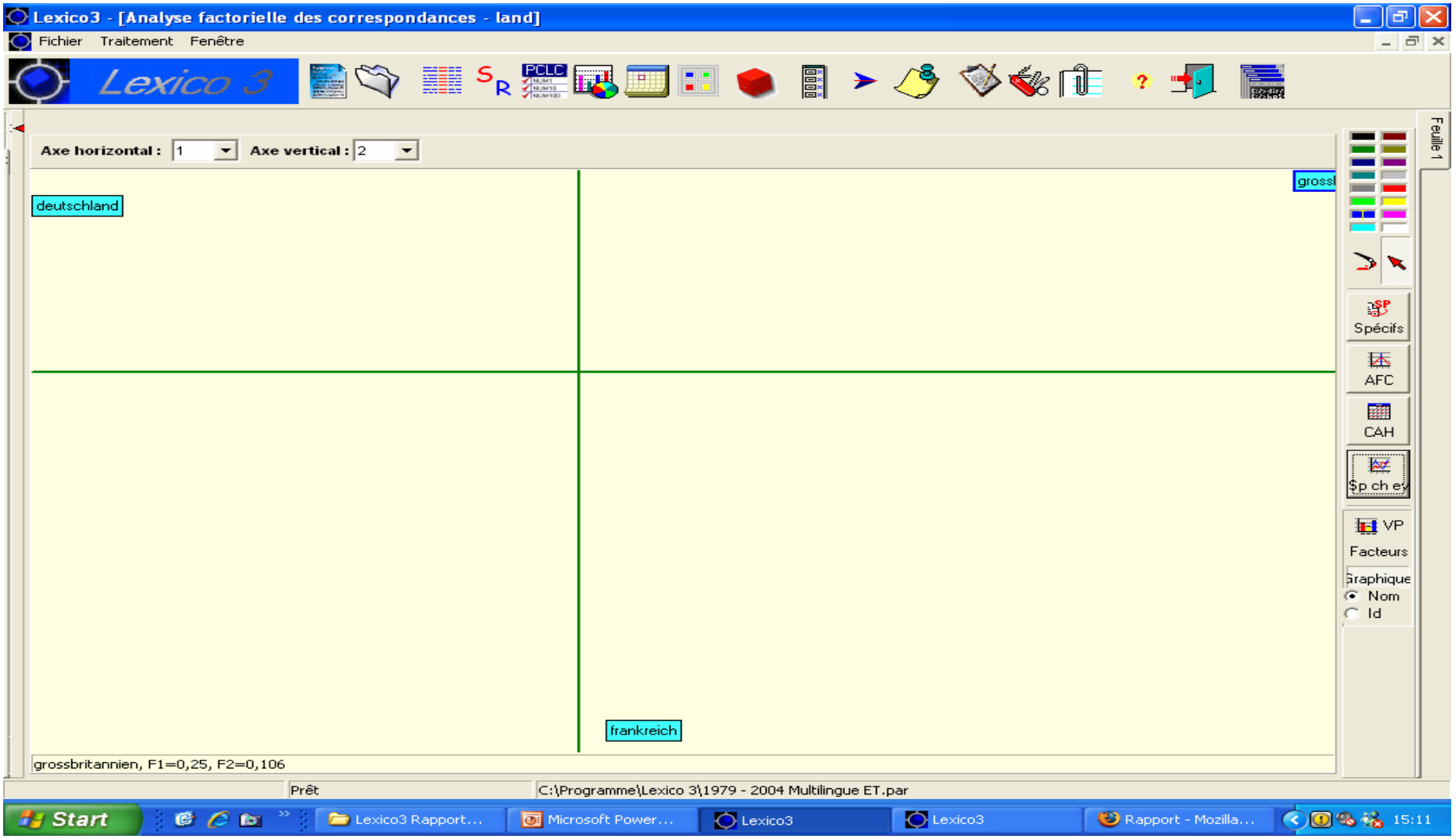
Verbe infinitif	VER:infi	VB (be) VH+ (have) VV	VVINF VVIZU VAINF VMINF	VERBINF
Verb présent	VER:pres VER:subp	VVP VVZ VBP VBZ VHP VHZ	VVFIN VAFIN VMFIN	VERBFIN
Verbe modal		MD		
Verbe futur	VER:futu			
Verbe passé	VER:impf VER:subi (imparfait) VER:simp	VBD VHD VVD		
Participe présent actif	VER:ppre	VBG VVG VHG	ADJD/ADJA	
Verbe impératif	VER:impe		VVIMP VAIMP	
Verbe conditional	VER:cond			

Index du corpus multilingue

Forme	Fréquence
NOM	182869
PRÄP	110775
ARTPROIND	107009
ADJ	83638
VERBFIN	54908
KON	40095
ADV	33800
NAM	31941
VERBINF	30698
PROPER	18077
PPP	17095
PROPOS	8391
PROREL	7259

Forme	Fréquence
NUM	5916
PRODEM	5843
PRON	478
SYM	452
INT	74
,	40025
.	35312
:	3763
(	3367
„	757
'	288
)	247

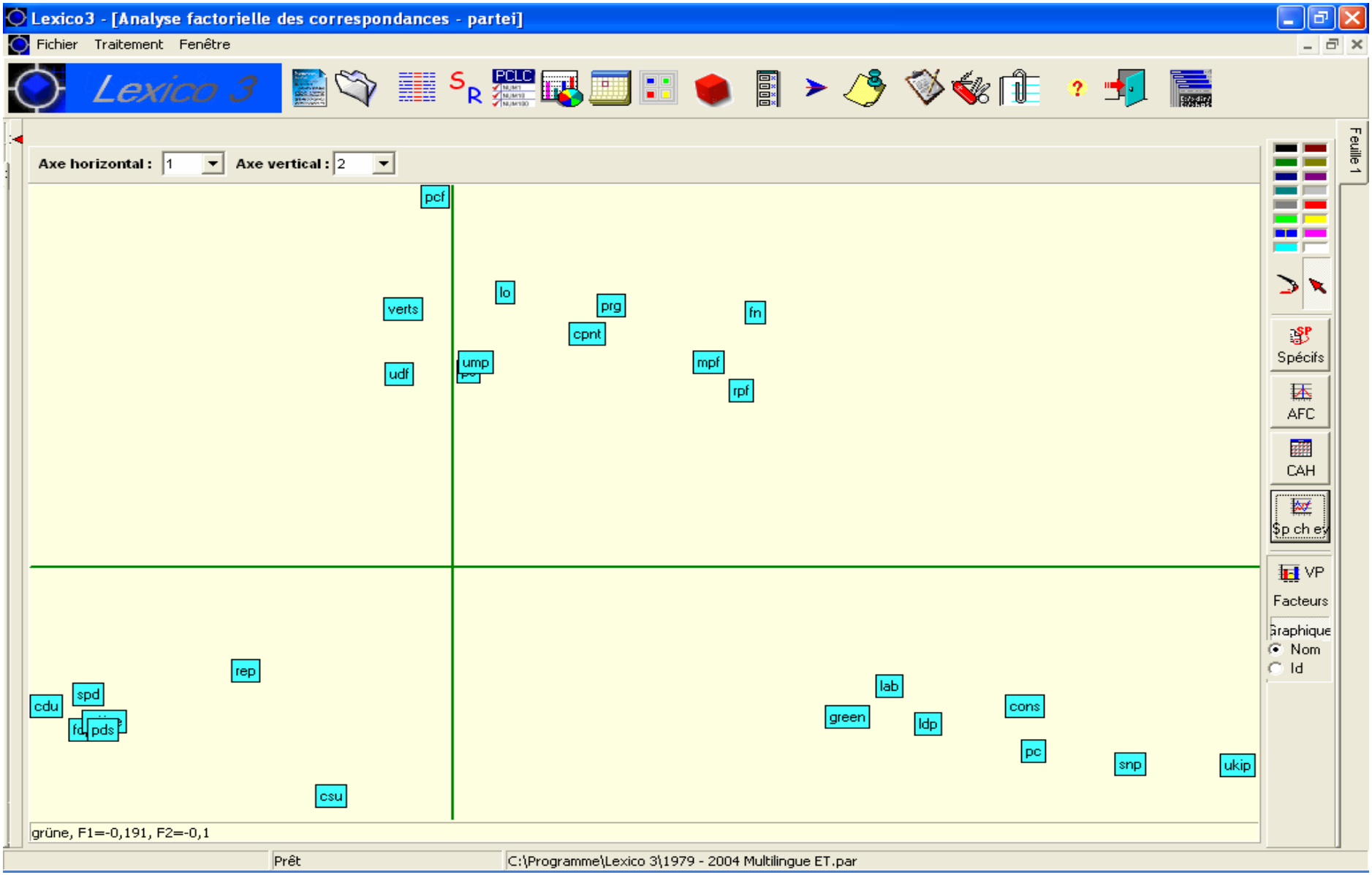
AFC de la partition <land> (pays)



Spécificités de la partition <land>

Formes	allemandes		français		britannique	
	Fréquence	*Spécificité*	*Fréquence*	*Spécificité*	*Fréquence*	*Spécificité*
NOM	74241	50	62901	-14	45727	-50
PRÄP	28364	-50	50387	50	32024	32
ARTPROIND	49374	50	35772	-35	21863	-50
ADJ	37420	50	24848	-50	21370	-41
VERBFIN	18988	-49	16836	-50	19084	50
KON	18159	50	12207	-50	9729	-50
ADV	12593	0	11863	0	9344	0
NAM	6226	-50	8459	-50	17256	50
VERBINF	11020	-9	9124	-50	10554	50
PROPER	6035	-31	7253	47	4789	-3
PPP	5824	-21	6124	2	5147	16
PROPOS	2772	-18	2885	0	2734	26
PROREL	2637	-2	3292	50	1330	-50
NUM	1196	-50	2879	50	1841	11
PRODEM	2222	0	3621	50	0	0
„	0	0	757	50	0	0
PRON	69	-29	88	-16	321	50
SYM	68	-26	286	34	98	-3
INT	24	0	39	3	11	-3

AFC de la partition <partei> (parti)



Spécificités de la partition <partei> des partis les plus extrêmes

Formes	CDU		CSU		PCF		UKIP	
	Fréquence	*Spécificité*	*Fréquence*	*Spécificité*	*Fréquence*	*Spécificité*	*Fréquence*	*Spécificité*
"		0		0	254	50		0
ADJ	7905	50	2132	22	3141	-36	329	-8
ADV	2200	-18		0	1349	-8	221	4
ARTPROIND	10566	50	2672	23	4619	-6	431	-9
INT		0	7	4	8	2		0
KON	3624	25	944	4	1592	-10	141	-7
NAM	955	-50	792	7	890	-50	438	50
NOM	15474	45	12207	-50		0	736	-16
NUM	337	-10	48	-15		0	51	4
PPP	1182	-5	302	-3		0	105	2
PRÄP	5754	-50	1549	-50	7035	50		0
PRODEM		0	328	-3	577	50		0
PRON	12	-7	85	-4	14	-2	9	3
PROPER	1184	-10	278	13	979	8		0
PROPOS	701	2	130	-2		0	78	6
PROREL	624	3	1	-4	387	3		0
SYM		0		0		0		0
VERBFIN	3642	-24		0	2020	-28	390	10
VERBINF		0	480	-12	1317	-3	197	3

Conclusions générales à partir de l'AFC

- Notre hypothèse:

  « L'échelle de la ressemblance morphosyntaxique entre les partis politiques dépend de la ressemblance dans l'orientation politique de ces derniers, indépendamment du pays. » ne peut pas être confirmée. L'AFC montre bien que, au niveau morphosyntaxique, les partis se distinguent bien par rapport à la langue dans laquelle ils sont écrits. Mais sur le niveau morphosyntaxique il est possible de distinguer des partis « plus allemand », « plus français », « plus britannique » que les autres.
- Par rapport aux textes français et britanniques, la spécificité des textes **allemands** se situe au niveau des noms qui sont liés avec les articles, les adjectifs et les conjonctions.
- La spécificité des textes **français** se situe au niveau des différents pronoms. On peut dire que le discours français est pronominal, c'est-à-dire qu'il fonctionne sur des rapports entre les parties de phrase et les phrases mêmes. La grande spécificité des pronoms personnels indique un discours éthique dans le sens de la rhétorique aristotélicienne. La spécificité pronominale des textes français demande un plus grand effort de réflexion du lecteur que les textes anglais. Les pronoms créent en effet un espace virtuel tandis que les noms créent un espace plus direct et plus réel.
- La spécificité des textes **anglais** se situe au niveau des verbes conjugués et des verbes à l'infinitif, c'est-à-dire que le discours britannique suggère une plus grande activité que les textes français et allemand. Par ailleurs, les textes anglais contiennent beaucoup plus de noms propres que les textes allemands et français.

Les oppositions linguistiques ?

Noms:	+A	–F(-14)	–GB	= A ⇔ GB
Pronoms personnels	–A	+ F	–GB (-3)	= A ⇔ F
Chiffres cardinaux	–A	+F	+GB (+11)	= A ⇔ F
Pronoms relatifs	–A (-2)	+F	–GB	= F ⇔ GB
Articles et pronoms indéfinis:	+A	–F	–GB	= A ⇔ F et GB
Adjectifs:	+A	–F	–GB	= A ⇔ F et GB
Conjonctions:	+A	–F	–GB	= A ⇔ F et GB
Prépositions:	–A	+F	+GB	= A ⇔ F et GB
Noms propres:	–A	–F	+GB	= A et F ⇔ GB
Verbes conjugués:	–A	–F	+GB	= A et F ⇔ GB
Verbes à l'infinitif	–A (-9)	–F	+GB	= A et F ⇔ GB