ATHIS. L'historien, le texte et l'ordinateur. Les logiciels de traitement informatique du texte.

Alain Dallo
Université Paris I. LAMOP

1. Des logiciels qui ont une histoire. L'utilisation de corpus. 4
1.1. Plusieurs centres de recherche 4
1.2. Des avancées récentes 4
1.3. Corpus, logiciels, approches 4
2. L'approche thématique 5
2.1. La forme et sa distribution dans le corpus 5
2.1.1. Index avec Hyperbase 6
2.1.2. Dictionnaire des formes avec Lexico 3 7
2.1.3. Dictionnaire des lemmes avec Weblex 8
2.1.4. Nooj : liste des formes et dictionnaire 9
2.1.5. Nooj : Liste des formes et dictionnaire 10
2.1.6. Calcul des rafales dans weblex 11
2.1.7. Graphique de la répartition de je 12
2.1.8. Rafale, topologie : Europe (Vème République) 13
2.1.9. Rafale, topologie : Europe chez Mitterand 14
2.1.10. Rafale, carte des sections : Europe 15
2.2. Les concordances 16
2.2.1. Concordances de la forme je chez Mitterand 17
2.2.2. Concordances du lemme croire 18
2.2.3. Concordances d'une liste de quelques formes avec Lexico 3 19
2.2.4. Les concordances de deux lemmes éloignés avec Weblex. Utilisation de CQP = Corpus Query Processor 20
2.3. Les segments répétés 22
2.3.1. Les segments répétés avec Lexico 3 23
2.3.2. Les segments répétés du corpus voeux. Weblex 24
2.4. Les cooccurrences 25
2.4.1. Les cooccurrences : Unité France 26
2.4.2. Les contextes courts de je 27
2.4.3. L'environnement de la forme je 28
2.4.4. Extrait de la liste de toutes les cooccurrences significatives 29
2.4.5. Cooccurrences droites et gauches de France 30
2.4.6. Graphique des cooccurrences de je 31
2.4.7. Résumé du graphique des cooccurrences de je 32

3. L'approche statistique ou lexicométrique ____ *33*

3.1. Les caractéristiques d'un corpus ____ **33**

3.1.1. L'étude des formes en fonction de leur fréquence ____ 33

3.1.2. Le nombre d'occurrences du corpus et de ses composantes ____ 33

3.1.3. Fréquence de nous ____ 34

3.1.4. Les mots outils dans le tableau lexical entier ____ 35

3.1.5. Les vrais et les faux hapax ____ 36

3.2. Les spécificités d'une partie ____ **37**

3.2.1. J.C. Spécificités positives ____ 38

3.2.2. J.C. Spécificités négatives ____ 39

3.2.3. J.C. Spécificités phrastiques ____ 40

3.3. Rapprochement et éloignement des textes d'un corpus ____ **41**

3.3.1. L'analyse factorielle du dictionnaire ____ 42

3.3.2. AFC sur une liste de mots ____ 43

3.3.3. Le classement automatique hiérarchisé ____ 44

4. Une autre approche en gestation ____ *45*

4.1. Lemmes et catégories grammaticales ____ **45**

4.1.1. formes et lemmes ____ 46

4.1.2. formes et catégories ____ 47

4.1.3. concordance dnav n=France ____ 48

4.1.4. concordance (N, V, A) nom = France, verbe = être, ____ 49

5. Quelques liens ____ *50*

5.1. Logiciels ____ **50**

5.2. Corpus ____ **50**

1. Des logiciels qui ont une histoire. L'utilisation de corpus.

1.1. Plusieurs centres de recherche

- Charles Muller
- Étienne Brunet. Université de Nice
- ENS de Saint-Cloud
- Maurice Gross, Paris VII. Max Silberztein

1.2. Des avancées récentes

- Corpus étiquetés
- Préparation des corpus et balisage morpho-syntaxique
- Les moteurs de recherche, la fouille de textes et l'indexation fine de documents

1.3. Corpus, logiciels, approches

Partir du mot pour aller jusqu'au positionnement des textes d'un corpus, en passant par des études thématiques

- Un corpus : les voeux des présidents de la V ème République.
  - Corpus mis en place sur Weblex dans le cadre de la thèse de Jean-Marc Leblanc
  - Corpus intégré dans les autres logiciels à partir des textes fournis par Damon Mayaffre
- Les logiciels : Hyperbase, Lexico3, Weblex et Nooj
- L'approche thématique ou documentaire
- L'approche statistique

2. L'approche thématique

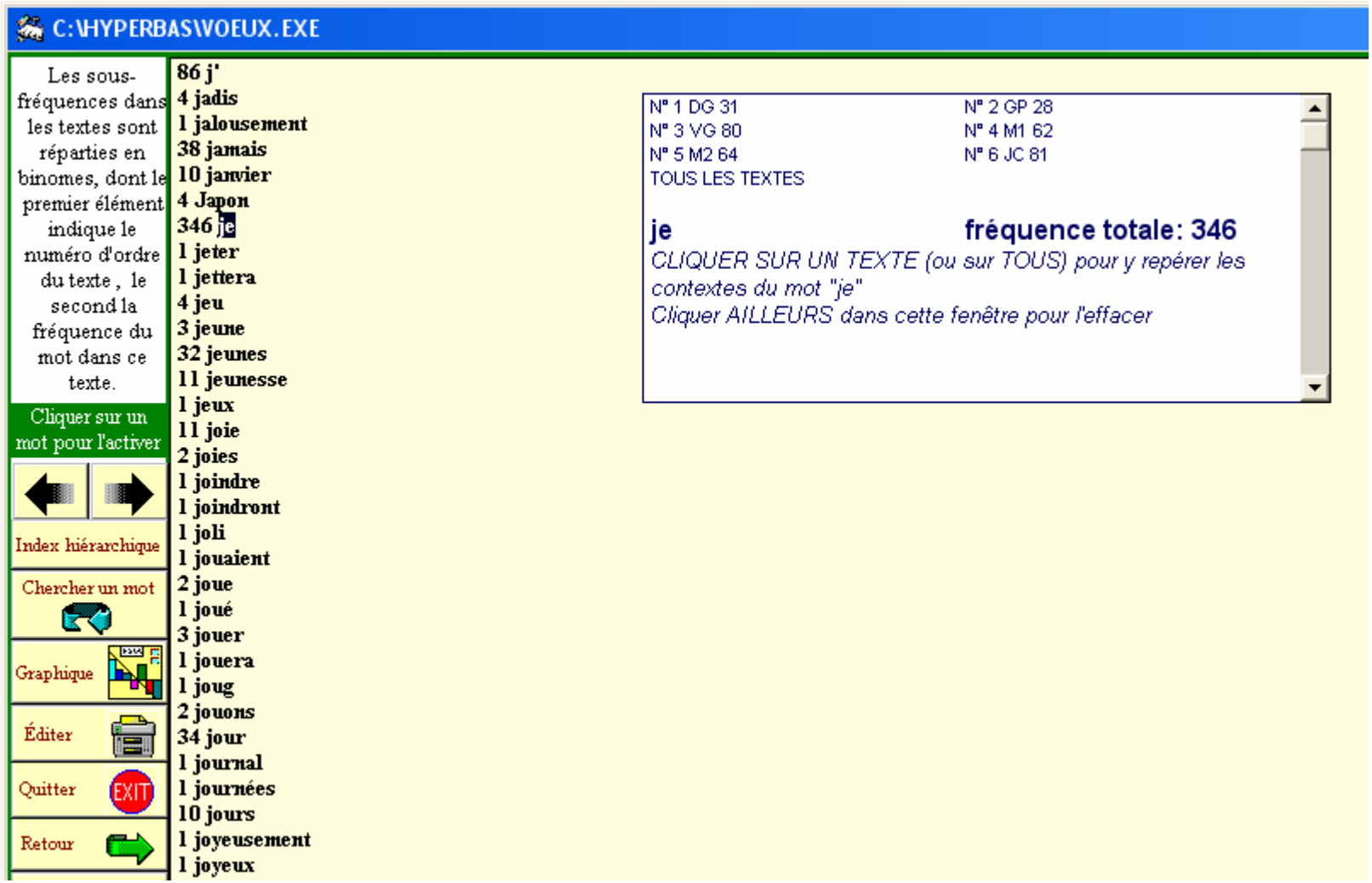
De la forme isolée à son environnement

- L'étude du dictionnaire
- Les concordances d'une forme
- Les segments
- Les cooccurrences

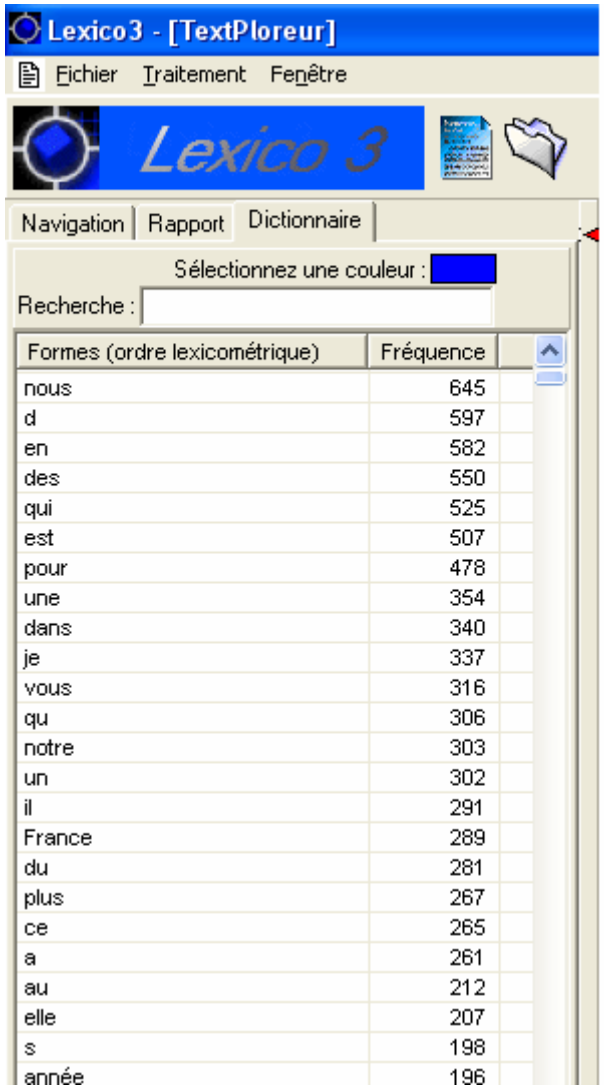
2.1. La forme et sa distribution dans le corpus

- Un index avec Hyperbase
- Un dictionnaire des formes avec Lexico 3
- Dictionnaire des lemmes avec avec Weblex
- Le dictionnaire des lemmes repérés : Nooj. Images 1 et 2.
- Calcul des rafales des voeux dans Weblex et graphique de la répartition de je
- Rafales et distribution d'une forme avec Hyperbase. Images : 1 et 2
- Rafales et distribution d'une forme avec Lexico3

2.1.1. Index avec Hyperbase



2.1.2. Dictionnaire des formes avec Lexico 3



Formes (ordre lexicométrique)	Fréquence
nous	645
d	597
en	582
des	550
qui	525
est	507
pour	478
une	354
dans	340
je	337
vous	316
qu	306
notre	303
un	302
il	291
France	289
du	281
plus	267
ce	265
a	261
au	212
elle	207
s	198
année	196

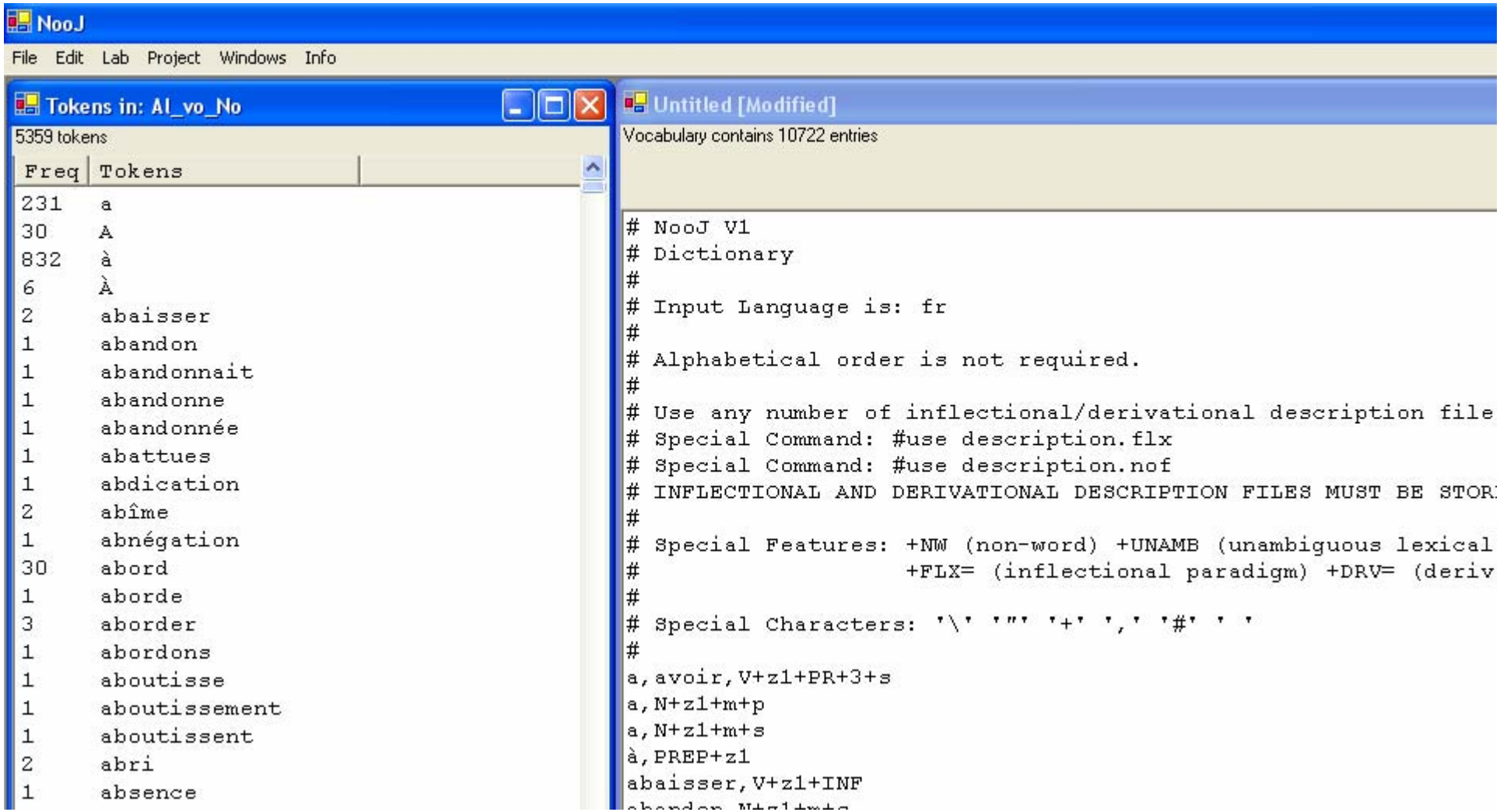
2.1.3. Dictionnaire des lemmes avec Weblex

Vocabulaire «élagué» du

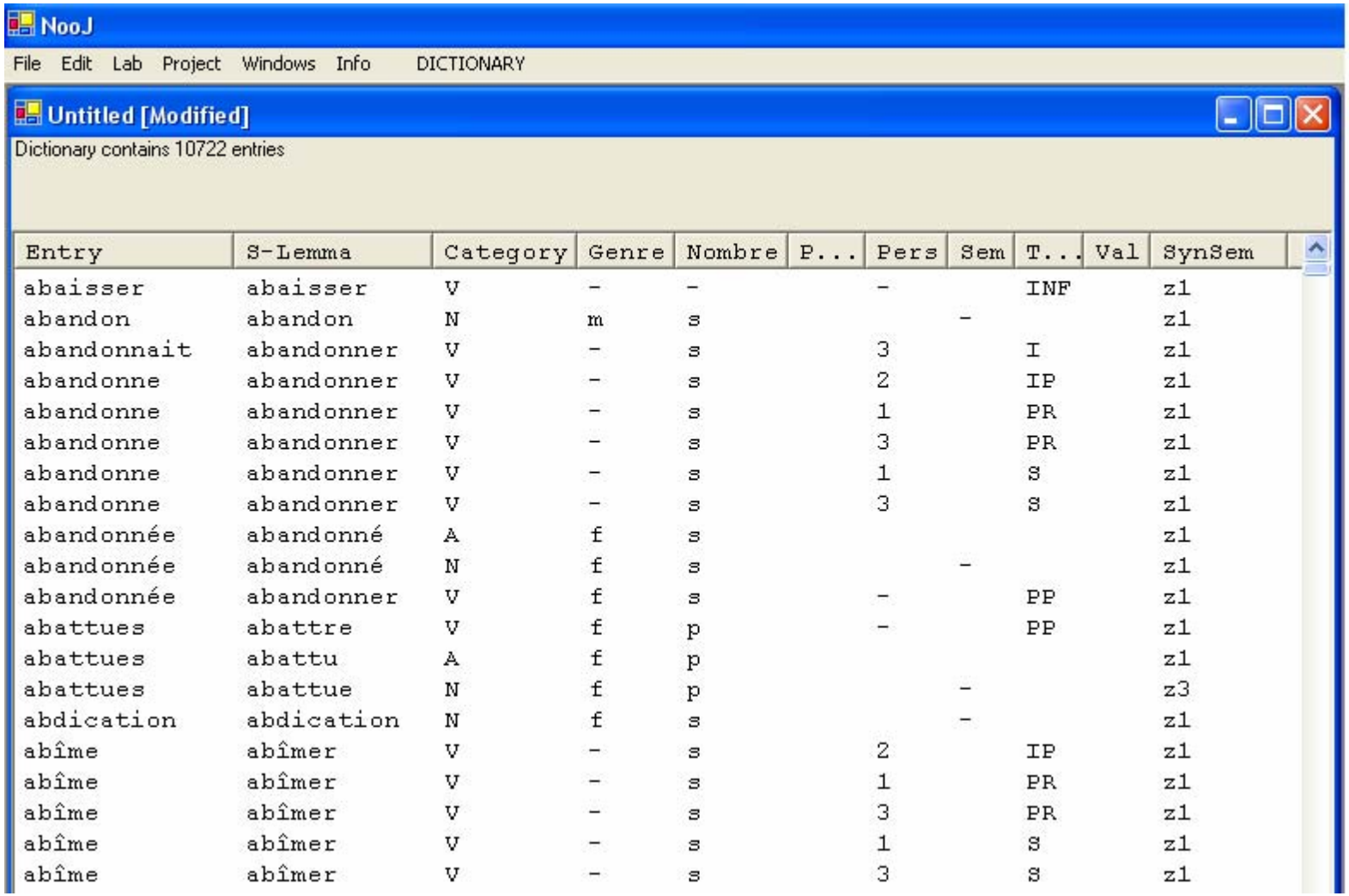
p3s de fréquence supérieure ou égale à 3.

Liste alphabétique			Liste hiérarchique	
ord	p3	fréq.	p3	fréq.
1	%	18	nous	654
2	Afrique	14	je	426
3	Airbus	4	France	300
4	Algérie	21	français	219
5	Allemagne	12	année	202
6	Amérique	15	faire	183
7	Angleterre	3	pouvoir	149
8	Asie	6	devoir	125
9	Berlin	3	pays	122
10	Bosnie	3	on	113
11	Canada	3	monde	102
12	Chine	4	Europe	99
13	Douze	3	vouloir	92
14	Europe	99	voeu	88
15	France	300	bon	87
16	Gaulle	3	grand	85
17	Gorbatchev	3	nouveau	85
18	Grande-Bretagne	3	dire	82
19	Guerre	3	peuple	82
20	Irak	4	an	80
21	Israël	3	souhaiter	79
22	Italie	4	savoir	77
23	Japon	4	aller	75
24	Koweït	5	falloir	68
25	Liban	4	vivre	68
26	Moyen-Orient	4	état	67
27	Nations-Unies	5	cher	66
28	Nouvel	12	vie	66
29	Noël	3	paix	64
30	Outre-Mer	9	compatriote	63
31	Paris	8	politique	61
32	Pologne	4	République	60
33	Russie	5	social	59

2.1.4. Nooj : liste des formes et dictionnaire



2.1.5. Nooj : Liste des formes et dictionnaire



Entry	S-Lemma	Category	Genre	Nombre	P...	Pers	Sem	T...	Val	SynSem
abaisser	abaisser	V	-	-		-		INF		z1
abandon	abandon	N	m	s			-			z1
abandonnait	abandonner	V	-	s		3		I		z1
abandonne	abandonner	V	-	s		2		IP		z1
abandonne	abandonner	V	-	s		1		PR		z1
abandonne	abandonner	V	-	s		3		PR		z1
abandonne	abandonner	V	-	s		1		S		z1
abandonne	abandonner	V	-	s		3		S		z1
abandonnée	abandonné	A	f	s						z1
abandonnée	abandonné	N	f	s			-			z1
abandonnée	abandonner	V	f	s		-		PP		z1
abattues	abattre	V	f	p		-		PP		z1
abattues	abattu	A	f	p						z1
abattues	abattue	N	f	p			-			z3
abdication	abdication	N	f	s			-			z1
abîme	abîmer	V	-	s		2		IP		z1
abîme	abîmer	V	-	s		1		PR		z1
abîme	abîmer	V	-	s		3		PR		z1
abîme	abîmer	V	-	s		1		S		z1
abîme	abîmer	V	-	s		3		S		z1

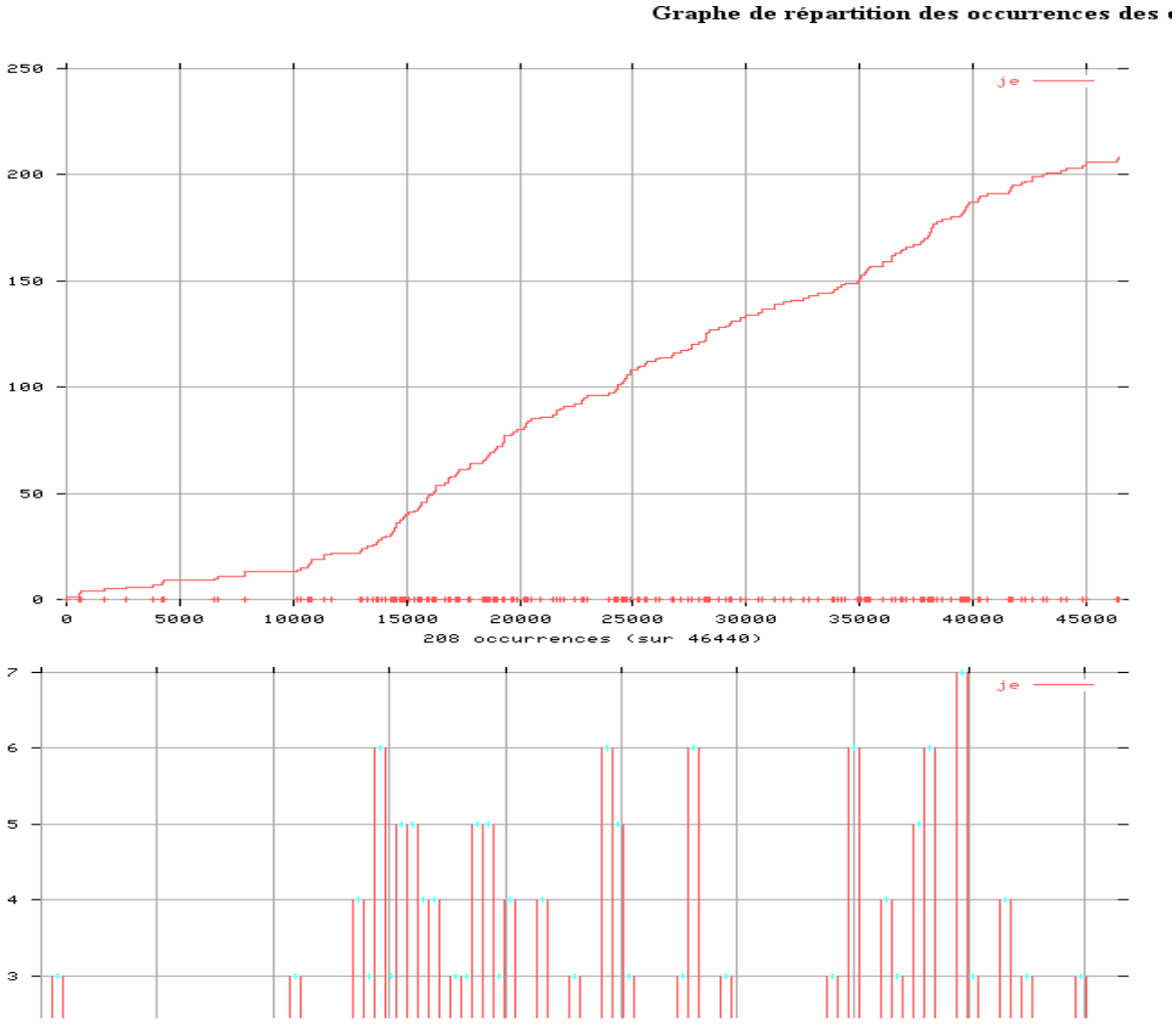
2.1.6. Calcul des rafales dans weblex

Répartition du vocal

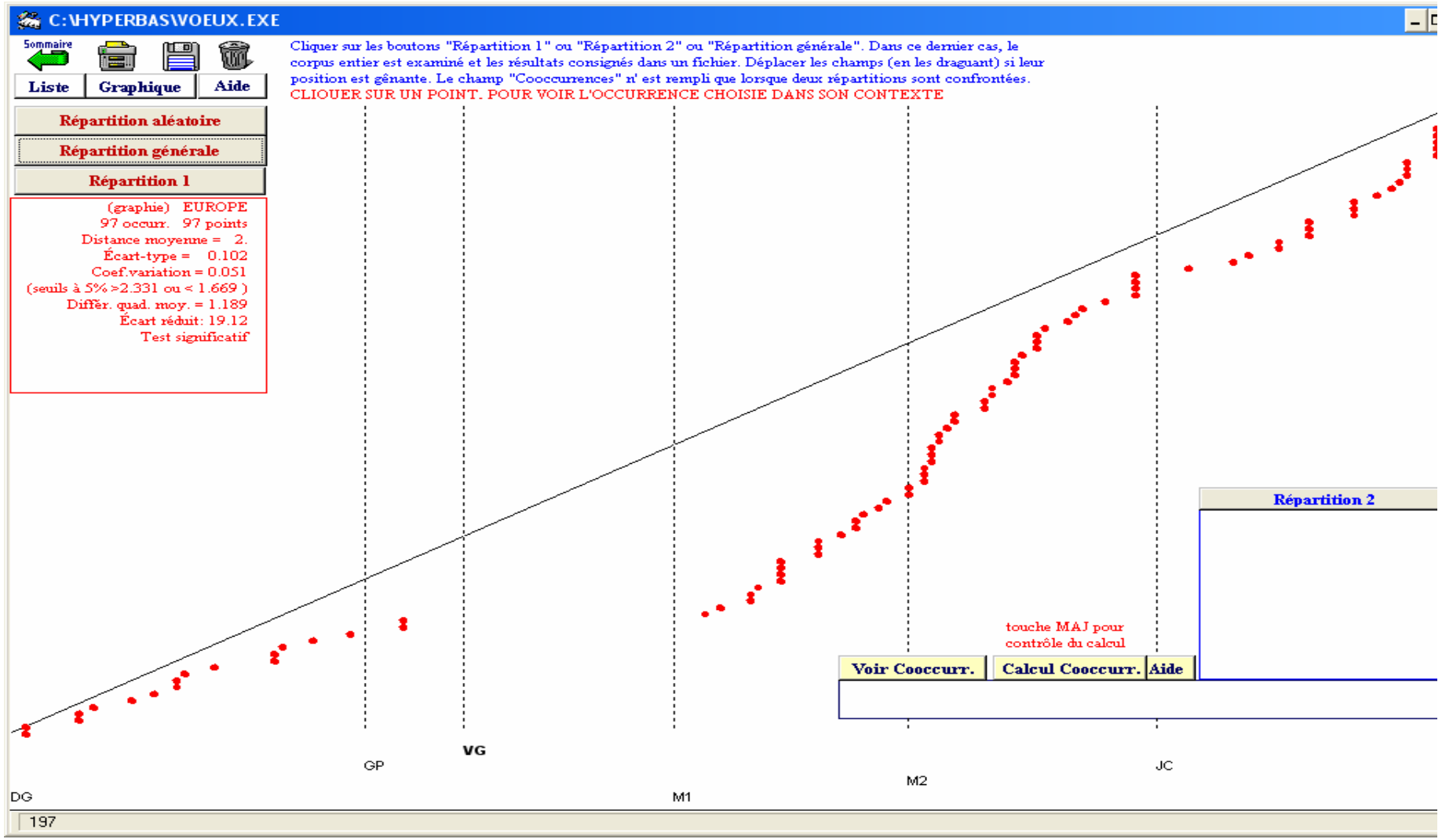
Z décroissant

	forme	Z	fp
1	compatriote	20.8582	63
2	euro	20.145	13
3	je	17.1565	376
4	Europe	14.4225	91
5	Algérie	14.1149	19
6	nier	13.8835	10
7	peur	13.8149	8
8	cher	13.0681	65
9	scientifique	12.8823	10
10	nous	12.0972	483
11	but	11.8479	10
12	progrès	11.6406	51
13	ami	11.5533	11
14	développement	10.6219	22
15	difficultés	10.6133	19
16	soldat	9.92321	10
17	vive	9.28211	53
18	apporter	9.22639	23
19	hausse	8.88236	6
20	au-dehors	8.38362	7
21	%	8.18859	10
22	bonheur	8.11869	20
23	valeurs	8.08423	11
24	démocratie	8.01422	16
25	fête	7.25487	7

2.1.7. Graphique de la répartition de je



2.1.8. Rafale, topologie : Europe (Vème République)



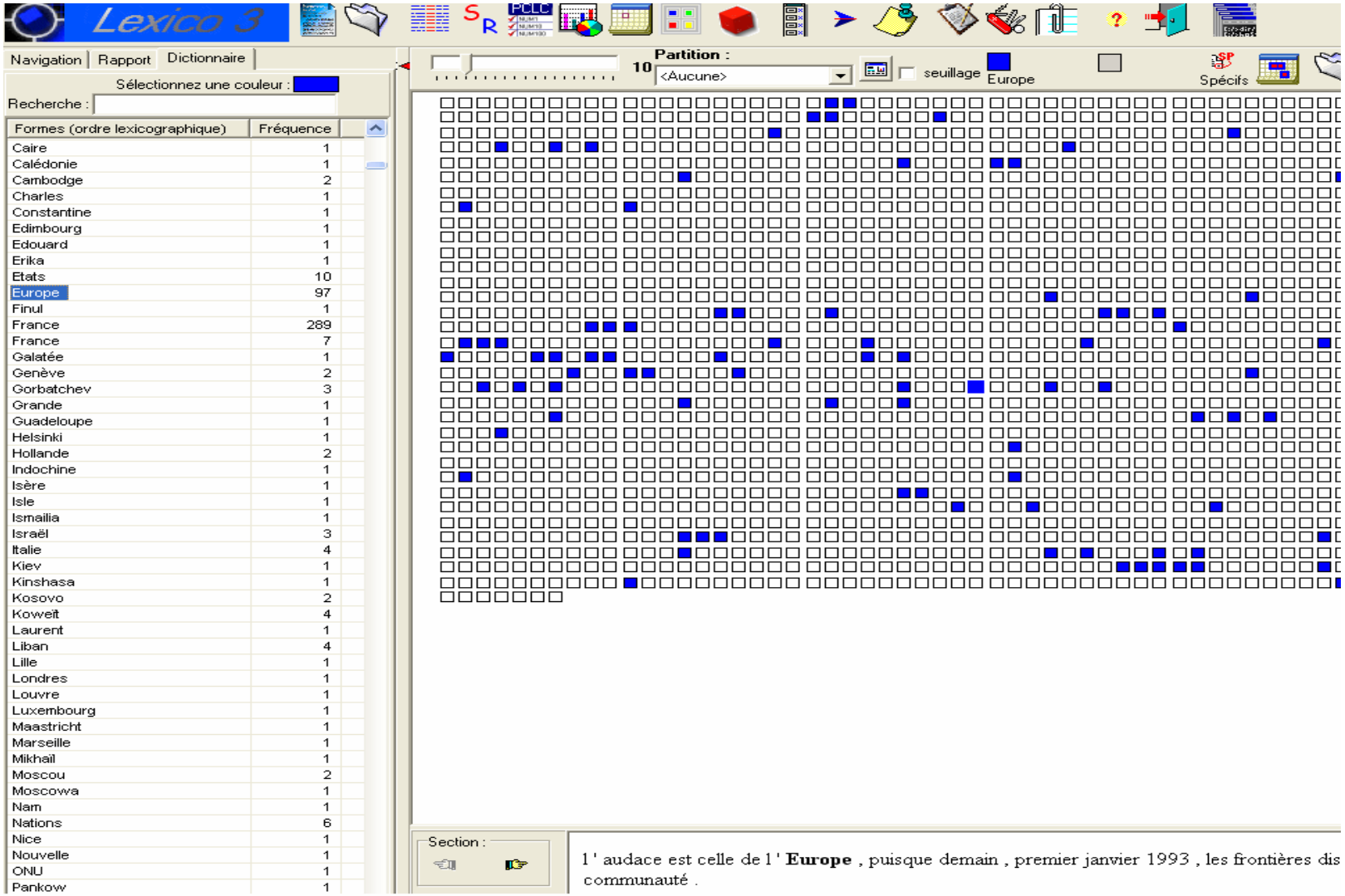
2.1.9. Rafale, topologie : Europe chez Mitterand

Depuis quelques mois , nous avons dans les yeux les images terribles des combats que se livrent les peuples d Yougoslavie , hier encore associés sous un même drapeau . Comment arrêter cette guerre ?

La France soutient les efforts de négociation et d' arbitrage de la Communauté et des Nations - Unies . Ell reconnaît le principe de l' autodétermination . Mais il lui paraît urgent que soient mises en place des structure inter - européennes , où le droit à l' indépendance ne se confondra pas avec l' anarchie des tribus d' autrefois . C sera , je le pense , l' un des enjeux majeurs de 1992 .

Car l' inquiétude gagne l' Europe de l' Est où l' on redoute la contagion . Comment cette inquiétude nou épargnerait - elle , nous qui , à l' Ouest , avons pourtant la chance de vivre en paix et d' avoir dépassé nos propre divisions ? Raison de plus de se réjouir des récents accords de Maastricht . Une monnaie commune , l' amorce d une diplomatie , d' une défense et d' une armée communes à l' Europe des Douze , une charte sociale , l' exempl de stabilité offert aux peuples qui se déchirent , bientôt 350 à 360 millions d' Européens solidaires sur la scène du monde - en attendant les autres - , bref , l' Europe qui se fait , voilà un grand dessein , capable d' enthousiasmer , d rassembler et de justifier l' espérance .

2.1.10. Rafale, carte des sections : Europe

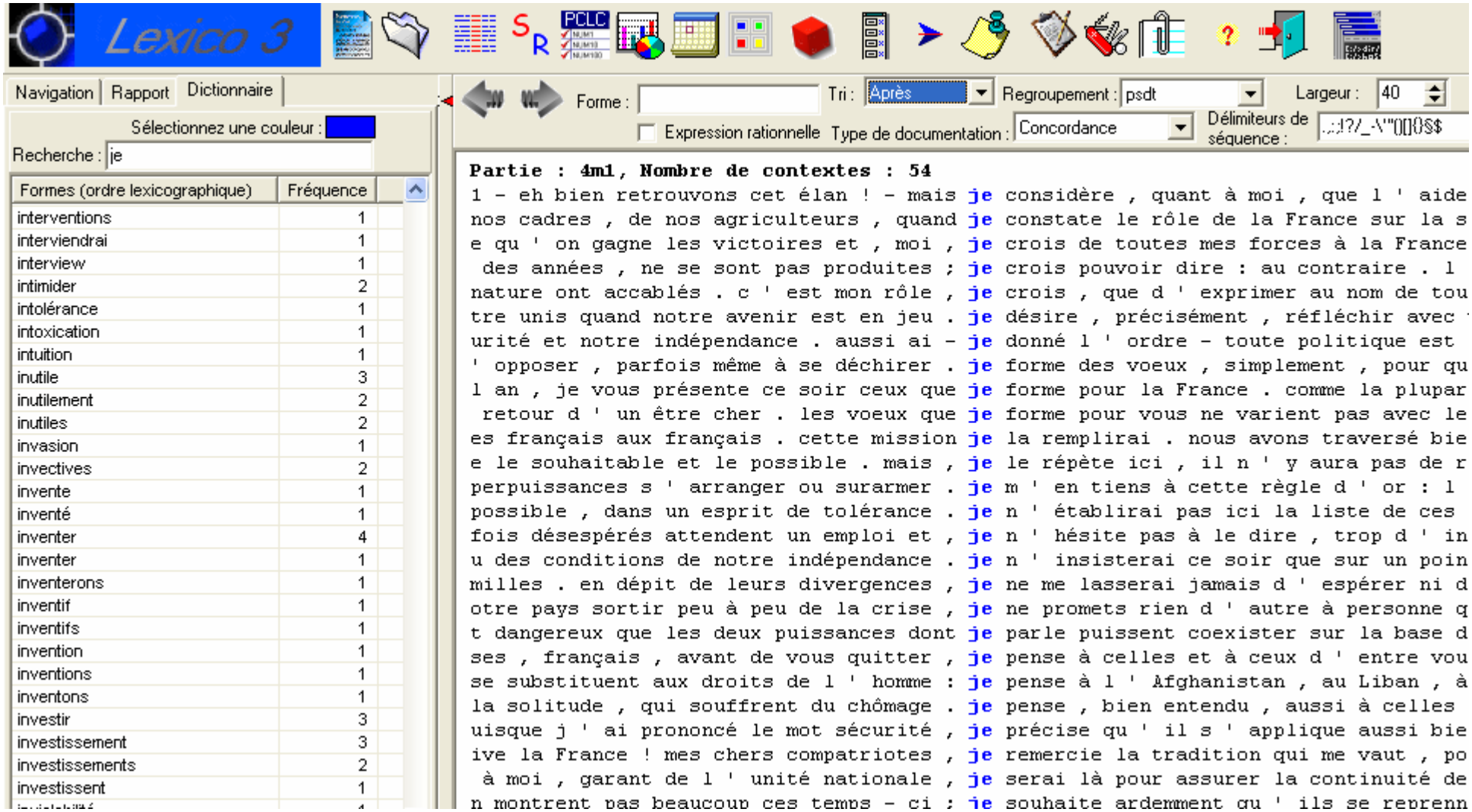


2.2. Les concordances

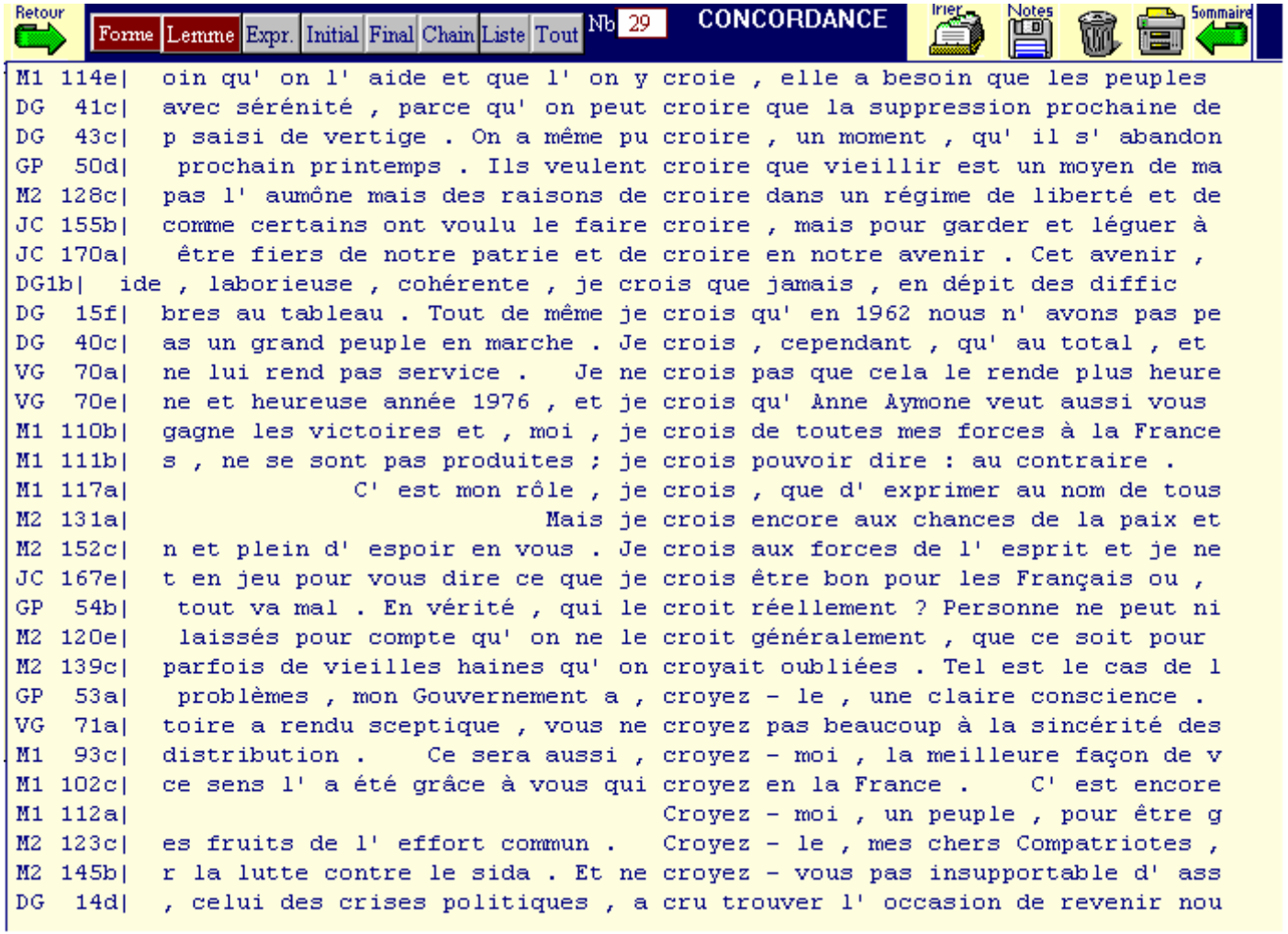
L'environnement proche d'une forme

- **Les concordances d'une forme : Lexico 3**
- **Les concordances d'un lemme avec Hyperbase**
- **Les concordances d'une liste de formes**
- **Les concordances de deux lemmes éloignés avec Weblex. Utilisation de CQP = Corpus Query Processor**

2.2.1. Concordances de la forme je chez Mitterand

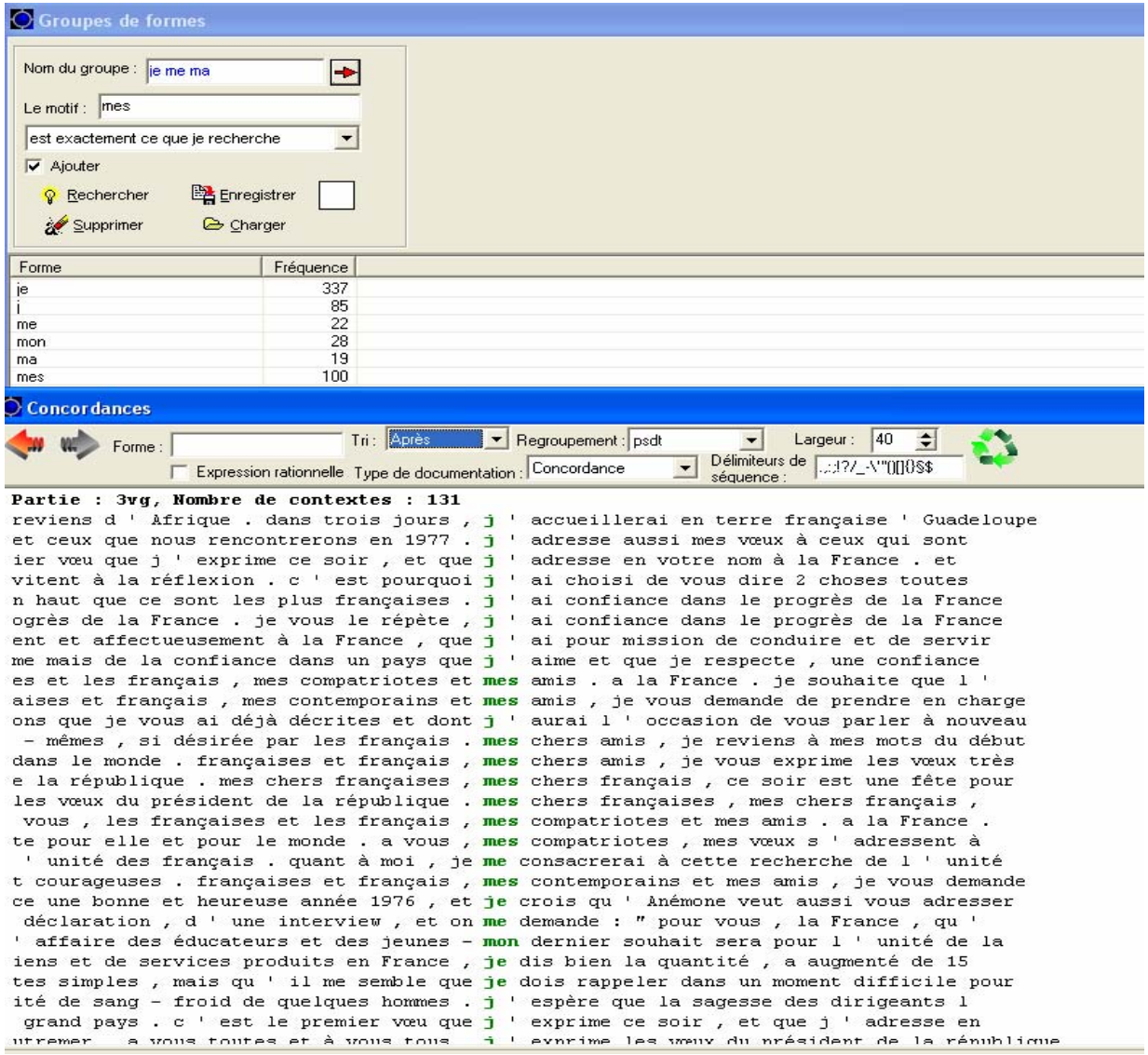


2.2.2. Concordances du lemme croire



M1 114e| oin qu' on l' aide et que l' on y croie , elle a besoin que les peuples
DG 41c| avec sérénité , parce qu' on peut croire que la suppression prochaine de
DG 43c| p saisi de vertige . On a même pu croire , un moment , qu' il s' abandon
GP 50d| prochain printemps . Ils veulent croire que vieillir est un moyen de ma
M2 128c| pas l' aumône mais des raisons de croire dans un régime de liberté et de
JC 155b| comme certains ont voulu le faire croire , mais pour garder et léguer à
JC 170a| être fiers de notre patrie et de croire en notre avenir . Cet avenir ,
DG1b| ide , laborieuse , cohérente , je crois que jamais , en dépit des diffic
DG 15f| bres au tableau . Tout de même je crois qu' en 1962 nous n' avons pas pe
DG 40c| as un grand peuple en marche . Je crois , cependant , qu' au total , et
VG 70a| ne lui rend pas service . Je ne crois pas que cela le rende plus heure
VG 70e| ne et heureuse année 1976 , et je crois qu' Anne Aymone veut aussi vous
M1 110b| gagne les victoires et , moi , je crois de toutes mes forces à la France
M1 111b| s , ne se sont pas produites ; je crois pouvoir dire : au contraire .
M1 117a| C' est mon rôle , je crois , que d' exprimer au nom de tous
M2 131a| Mais je crois encore aux chances de la paix et
M2 152c| n et plein d' espoir en vous . Je crois aux forces de l' esprit et je ne
JC 167e| t en jeu pour vous dire ce que je crois être bon pour les Français ou ,
GP 54b| tout va mal . En vérité , qui le croit réellement ? Personne ne peut ni
M2 120e| laissés pour compte qu' on ne le croit généralement , que ce soit pour
M2 139c| parfois de vieilles haines qu' on croyait oubliées . Tel est le cas de l
GP 53a| problèmes , mon Gouvernement a , croyez - le , une claire conscience .
VG 71a| toire a rendu sceptique , vous ne croyez pas beaucoup à la sincérité des
M1 93c| distribution . Ce sera aussi , croyez - moi , la meilleure façon de v
M1 102c| ce sens l' a été grâce à vous qui croyez en la France . C' est encore
M1 112a| Croyez - moi , un peuple , pour être g
M2 123c| es fruits de l' effort commun . Croyez - le , mes chers Compatriotes ,
M2 145b| r la lutte contre le sida . Et ne croyez - vous pas insupportable d' ass
DG 14d| , celui des crises politiques , a cru trouver l' occasion de revenir nou

2.2.3. Concordances d'une liste de quelques formes avec Lexico 3



2.2.4. Les concordances de deux lemmes éloignés avec Weblex. Utilisation de CQP = Corpus Query Processor

Chercher dans la page :

Il y a 7 occurrences de *"France"[]*"enfant"* dans le corpus voeux2-author-Degaulle (propr

1	Degaulle, 1960	ur et à son énergie . Car l' Algérie a besoin de la communauté française , et la	**France , pour son oeuvre , a besoin d' elle en Algérie . Bien entendu , et , quoi qu' il arrive , la enfants**
2	Degaulle, 1963	ustes envers nous-mêmes . Car le bilan est positif . Pendant ces douze mois , la	**France a continué de monter . En 1963 , notre population s' est augmentée de presque six cen est né environ neuf cent mille bébés . La proportion croissante des jeunes devant progressiv processus , on peut penser que , parmi les enfants**
3	Degaulle, 1963	ns , j' ai l' honneur et la charge de parler en notre nom à tous , j' offre à la	**France , cette fois encore , les voeux très ardents et très confiants de ses enf**
4	Degaulle, 1964	mmes sur la terre n' ont tant attendu de nous ni éprouvé tant d' attrait pour la	**France . Certes , la vie est la vie , autrement dit un combat , pour une nation comme pour un ho , toujours et partout , des difficultés à vaincre , des efforts à déployer , des peines à supporter , de dignité , de justice , de fraternité . Mais , tous ensemble , nous sommes un peuple , dont l' év les enfants**
5	Degaulle, 1965	ns nos têtes , dans nos coeurs et dans nos mains tout ce qu' il faut pour que la	**France parcoure une étape décisive de son progrès , pour qu' elle apporte plus de justice enco entre tous ses enfants**
6	Degaulle, 1965	eurs voeux pour 1966 et que , tous ensemble nous souhaitons une bonne année à la	**France . Vive la République ! Vive la France . Françaises , Français , Je vous souhaite une bon nom de nous tous , mes voeux sont l' expression de ceux que la France forme pour chacu**

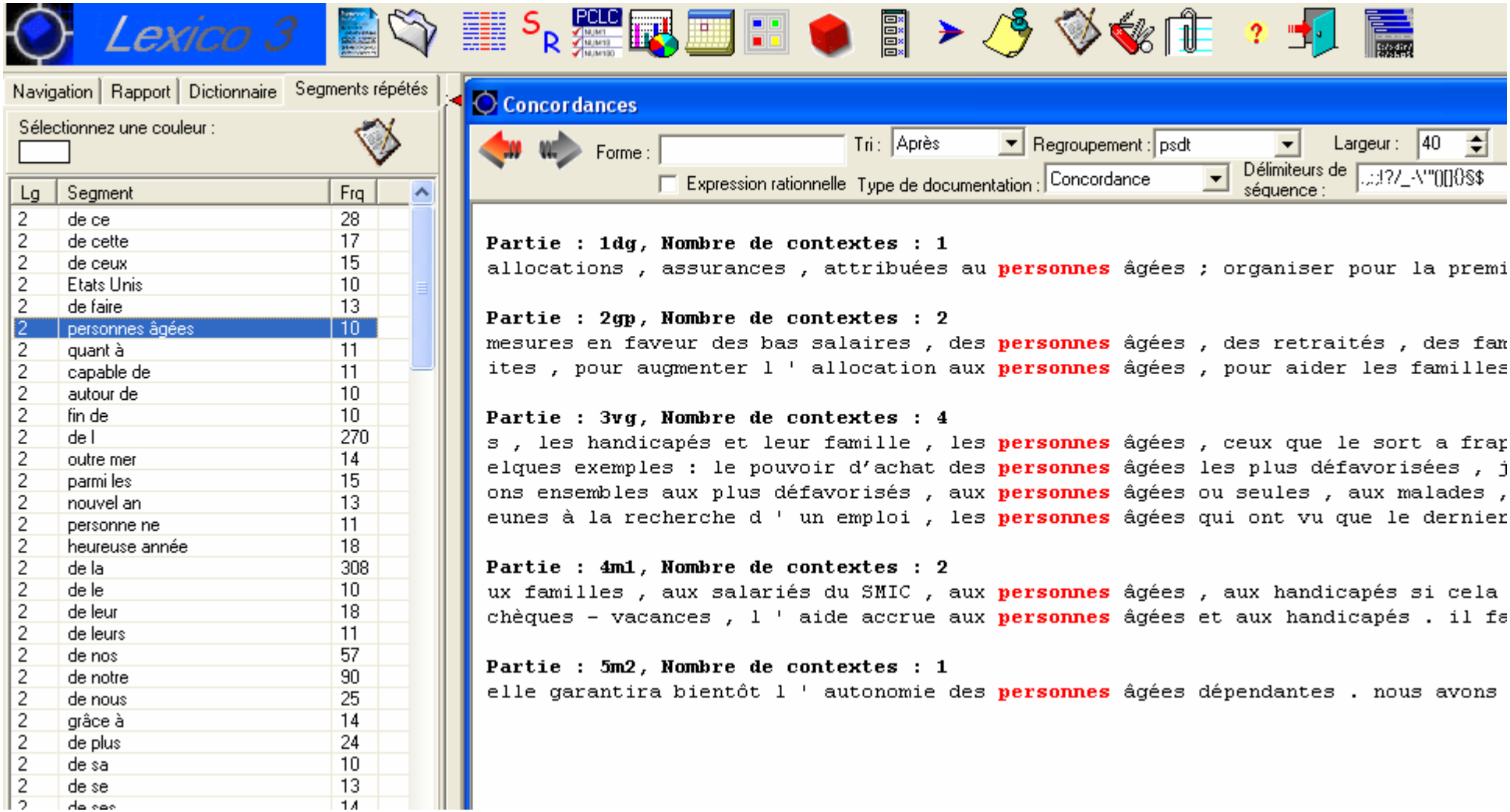
2.3. Les segments répétés

Les successions de formes contiguës du texte

- **Avec Lexico 3**

- **Avec Weblex**

2.3.1. Les segments répétés avec Lexico 3



2.3.2. Les segments répétés du corpus voeux. Weblex

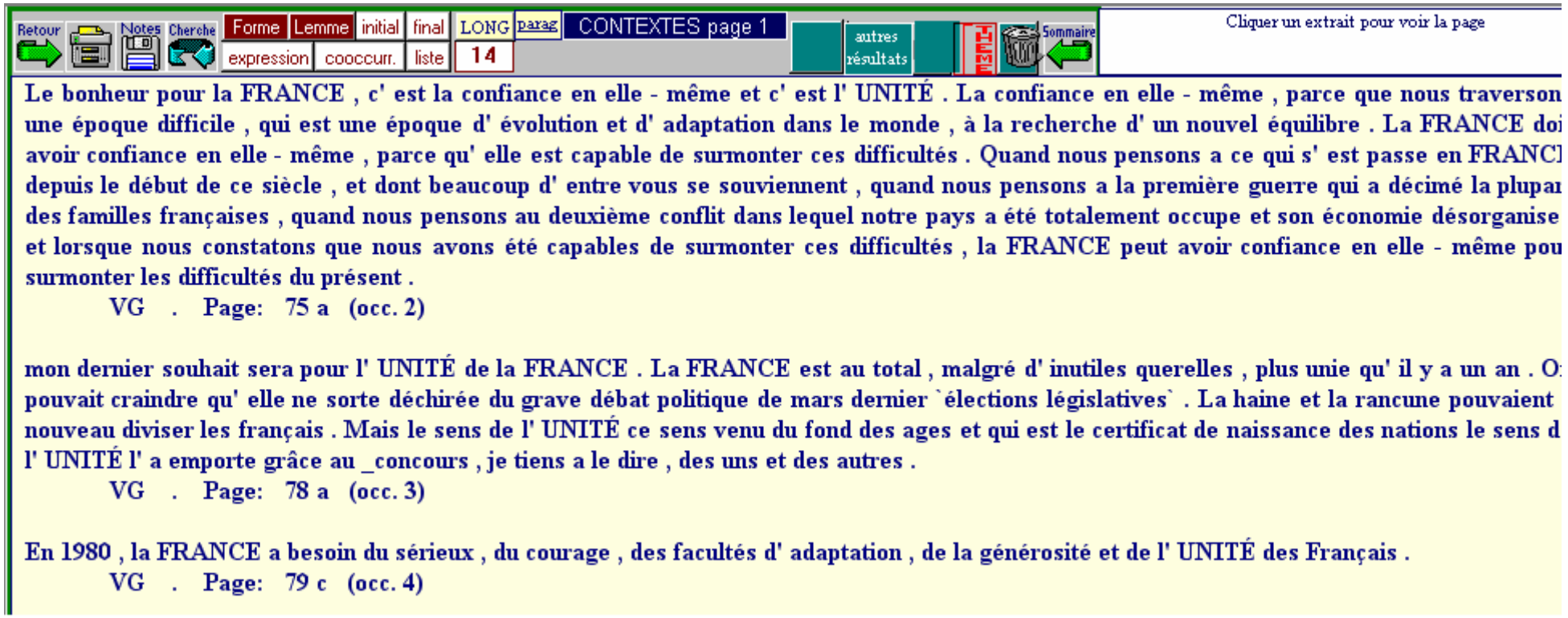
f	segment	l
8	chacune et à chacun d' entre vous	7
5	à chacune et à chacun d' entre	7
5	à chacune et à chacun de vous	7
3	C' est au nom de l' emploi	7
3	au nom de l' emploi que nous	7
3	est au nom de l' emploi que	7
3	À chacune et à chacun d' entre	7
8	chacune et à chacun d' entre	6
8	et à chacun d' entre vous	6
6	à chacune et à chacun de	6
5	chacune et à chacun de vous	6
5	à chacune et à chacun d'	6
4	Nous ne sommes pas les plus	6
4	À chacune et à chacun d'	6
3	C' est au nom de l'	6
3	Il n' y a pas d'	6
3	Personne ne peut nier que la	6
3	au nom de l' emploi que	6
3	ceux qui ont la chance d'	6

2.4. Les cooccurrences

Présence de deux formes non contiguës

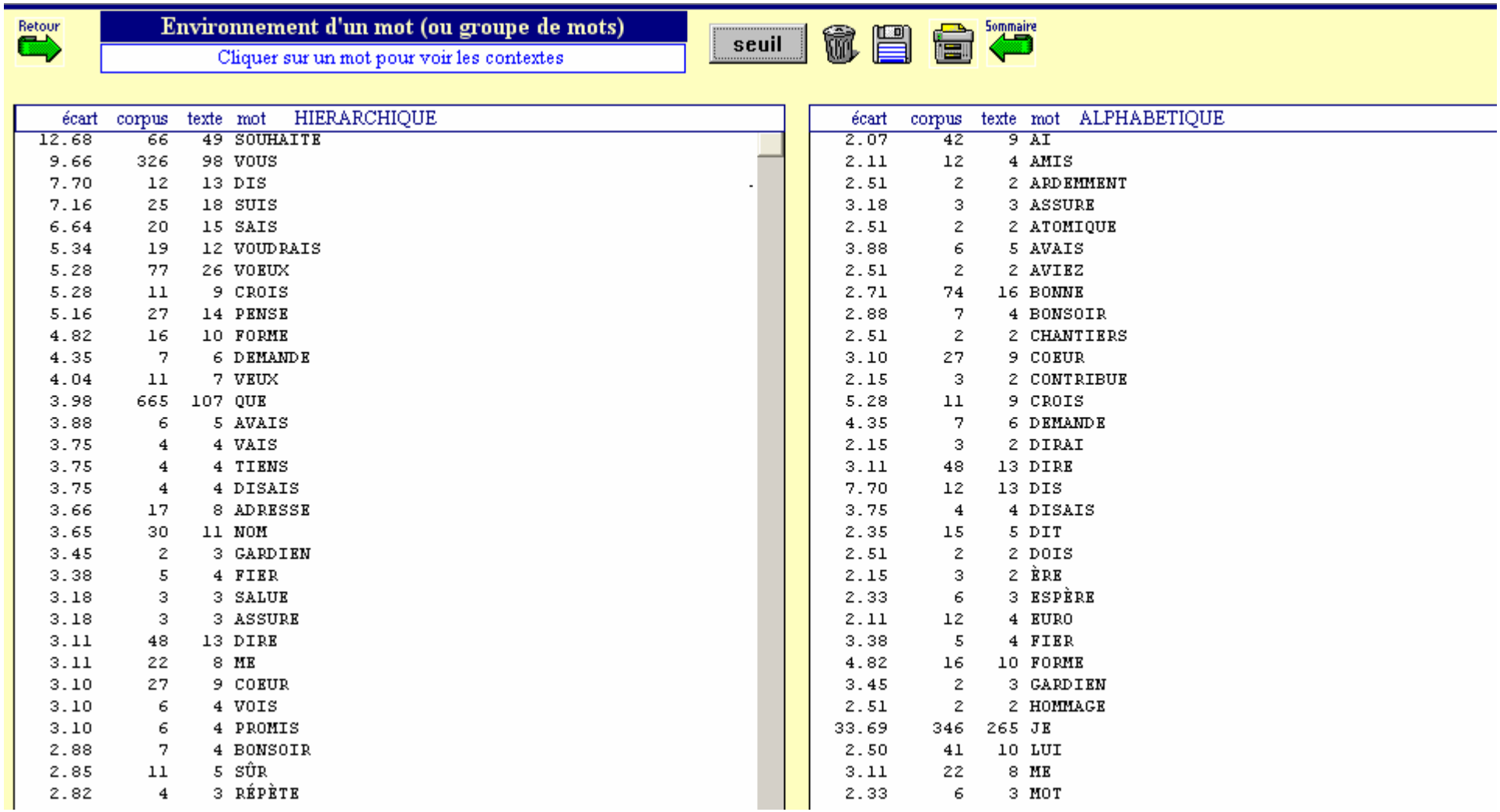
- **Les cooccurrences de deux formes dans hyperbase**
- **Utilisation de la fonction thème d'hperbase pour les découvrir. Images : 1 et 2.**
- **Le calcul des cooccurrences dans weblex**
- **Cooccurrences gauches et droites**
- **Graphique**
- **Résumé du graphique**

2.4.1. Les cooccurrences : Unité France



2.4.2. Les contextes courts de je

2.4.3. L'environnement de la forme je



écart	corpus	texte	mot	HIERARCHIQUE
12.68	66	49	SOUHAITE	
9.66	326	98	VOUS	
7.70	12	13	DIS	
7.16	25	18	SUIS	
6.64	20	15	SAIS	
5.34	19	12	VOUDRAIS	
5.28	77	26	VOEUX	
5.28	11	9	CROIS	
5.16	27	14	PENSE	
4.82	16	10	FORME	
4.35	7	6	DEMANDE	
4.04	11	7	VEUX	
3.98	665	107	QUE	
3.88	6	5	AVAIS	
3.75	4	4	VAIS	
3.75	4	4	TIENS	
3.75	4	4	DISAIS	
3.66	17	8	ADRESSE	
3.65	30	11	NOM	
3.45	2	3	GARDIEN	
3.38	5	4	FIER	
3.18	3	3	SALUE	
3.18	3	3	ASSURE	
3.11	48	13	DIRE	
3.11	22	8	ME	
3.10	27	9	COEUR	
3.10	6	4	VOIS	
3.10	6	4	PROMIS	
2.88	7	4	BONSOIR	
2.85	11	5	SÛR	
2.82	4	3	RÉPÈTE	

écart	corpus	texte	mot	ALPHABETIQUE
2.07	42	9	AI	
2.11	12	4	AMIS	
2.51	2	2	ARDEMMENT	
3.18	3	3	ASSURE	
2.51	2	2	ATOMIQUE	
3.88	6	5	AVAIS	
2.51	2	2	AVIEZ	
2.71	74	16	BONNE	
2.88	7	4	BONSOIR	
2.51	2	2	CHANTIERS	
3.10	27	9	COEUR	
2.15	3	2	CONTRIBUE	
5.28	11	9	CROIS	
4.35	7	6	DEMANDE	
2.15	3	2	DIRAI	
3.11	48	13	DIRE	
7.70	12	13	DIS	
3.75	4	4	DISAIS	
2.35	15	5	DIT	
2.51	2	2	DOIS	
2.15	3	2	ÈRE	
2.33	6	3	ESPÈRE	
2.11	12	4	EURO	
3.38	5	4	FIER	
4.82	16	10	FORME	
3.45	2	3	GARDIEN	
2.51	2	2	HOMMAGE	
33.69	346	265	JE	
2.50	41	10	LUI	
3.11	22	8	ME	
2.33	6	3	MOT	

2.4.4. Extrait de la liste de toutes les cooccurrences significatives

Cooccurrences du corpus "voeux2" (forme)

Seuils : **f** 3, **cf** 3, **p** 5.0E-2, $\mathbf{d_m}$ 1000.0

A	B	f_A	f_B	cf	p	d_m
Chers	Compatriotes	34	34	34	2e-75	0.0
Françaises	Français	38	124	34	8e-41	0.7
chers	compatriotes	21	28	18	1e-34	0.0
Vive	République	55	60	29	3e-34	1.0
bonne	année	51	199	37	4e-30	0.7
Nouvel	An	12	16	12	1e-28	0.0
Je	souhaite	136	65	34	2e-26	0.2
personnes	âgées	12	10	10	1e-25	0.0
pouvoir	achat	18	9	9	2e-20	1.0
intérêts	particuliers	15	10	9	2e-20	0.0
Bonne	année	24	199	21	2e-20	0.4
bonne	heureuse	51	22	14	6e-19	1.6
heureuse	année	22	199	19	3e-18	1.6
souhaite	bonne	65	51	18	3e-16	7.0
je	souhaite	208	65	30	5e-16	0.2
peut	nier	41	10	9	1e-15	0.0
cent	mille	12	10	7	3e-15	0.4
départements	territoires	6	6	5	1e-13	1.4

2.4.5. Cooccurrences droites et gauches de France

Lexicogramme du pôle "France" dans le corpus voeux2-author-Degaulle (p3)

Seuils : **f** 3, **cf** 3, **p** 5.0E-2, $\mathbf{d_m}$ 1000.0

	France (70)								
	cooccurrents gauches				**cooccurrents droits**				
	f	**cf**	**p**	$\mathbf{d_m}$		**f**	**cf**	**p**	$\mathbf{d_m}$
vive	18	9	1e-04	1.0	enfant	8	5	1e-03	8.6
souhaiter	12	6	2e-03	4.5					
nom	10	5	5e-03	13.2					
adresser	4	3	8e-03	14.7					
année	47	11	2e-02	6.5					
voeu	10	4	3e-02	13.0					
maintenant	6	3	3e-02	16.3					
terre	6	3	3e-02	16.7					

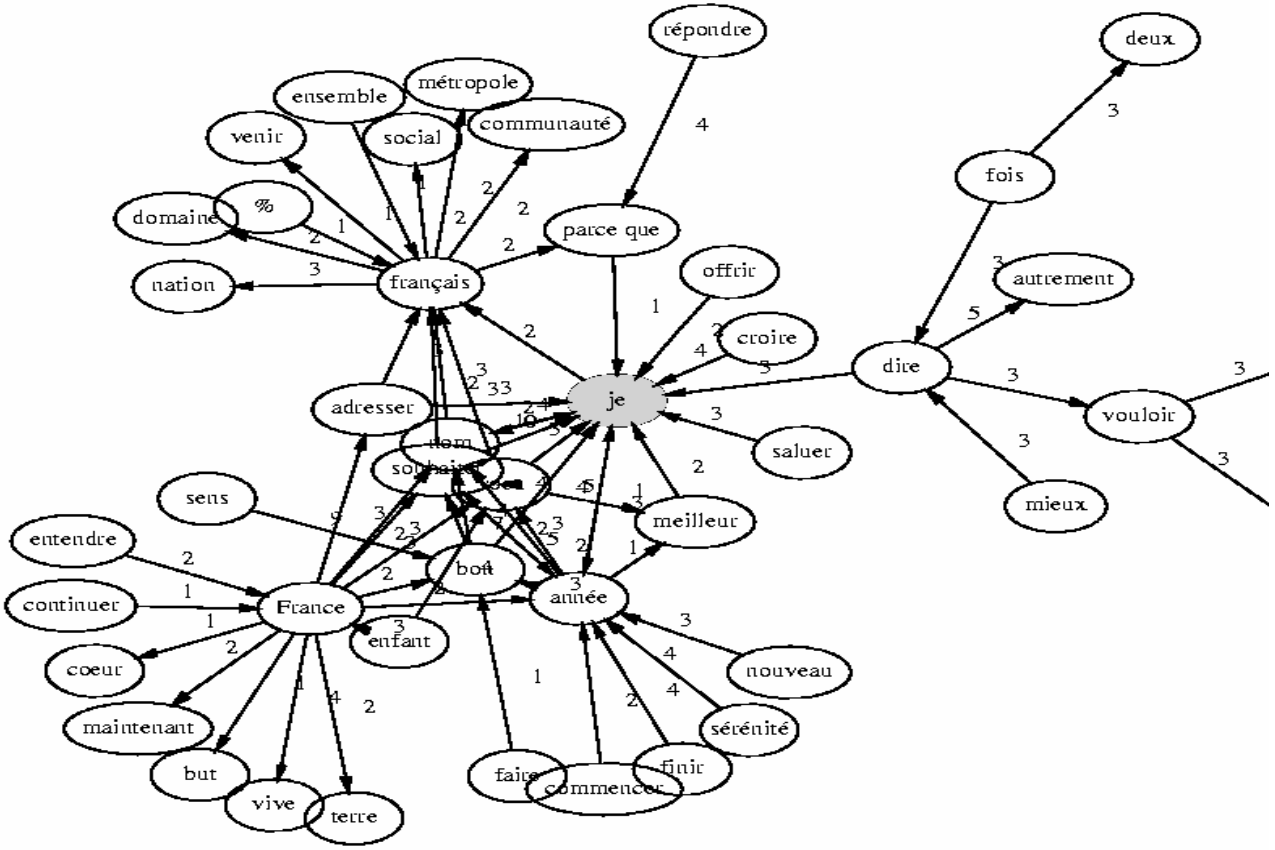
2.4.6. Graphique des cooccurrences de je

Lexicogramme récursif autour du pôle je dans le corpus `voeux2-author-Degaulle`

Seuils : **p** 2e-01, **r** 3, **f** 3, $\mathbf{d_n}$ 1000.0, **pl** 3

Synthèse

0 noeuds (0 arcs)



2.4.7. Résumé du graphique des cooccurrences de je

Calcul en cours (abandon)...

Synthèse des lexicogrammes récursifs du corpus `voeux2-author-Degaulle`

Seuils : **p** 5e-02, **r** 3, **f** 3, $\mathbf{d_m}$ 1000.0, **pl** 3

Tri décroissant par le nombre de noeuds par lexicogramme.

- 28 (nous avancer prospérité voie développement Amérique latin disposer lui-même part prendre propre évolution aide état Europe détente union coopération pui grand nombre peuple Afrique)
- 12 (communauté France finir nouveau sérénité bon je nom souhaiter année commencer Algérie)
- 11 (enfant meilleur voeu métropole nation catégories domaine économique social français adresser)
- 5 (d abord expansion niveau vie agir)
- 3 (commerce industrie agriculture)

---

3. L'approche statistique ou lexicométrique

De l'étude des fréquences des formes au positionnement des textes

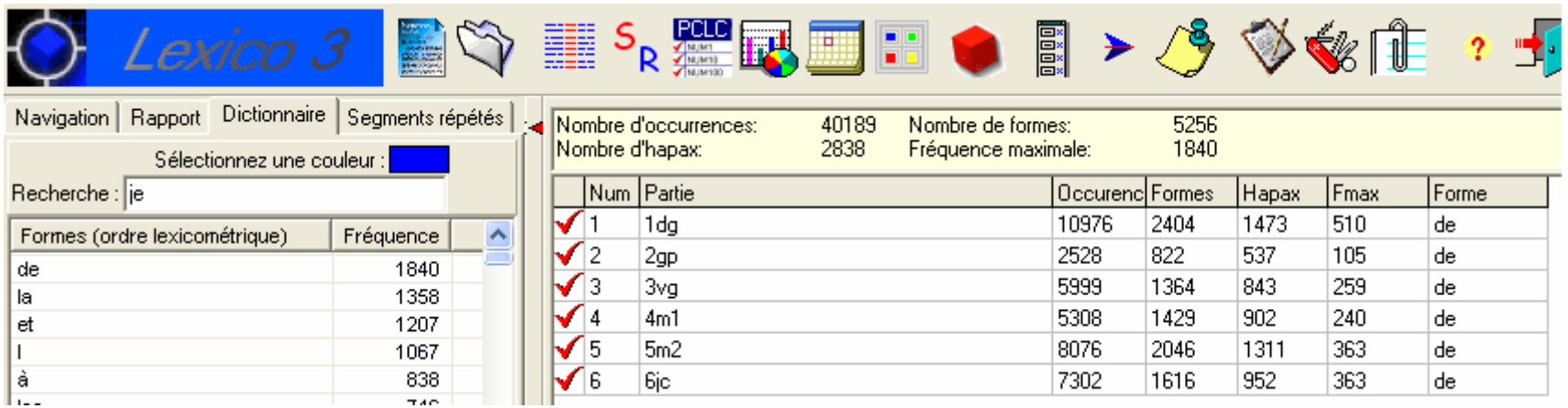
- **Les caractéristiques d'un corpus**
- **Les spécificités d'une partie de celui-ci**
- **Rapprochement et éloignement des textes le composant**

3.1. Les caractéristiques d'un corpus

3.1.1. L'étude des formes en fonction de leur fréquence

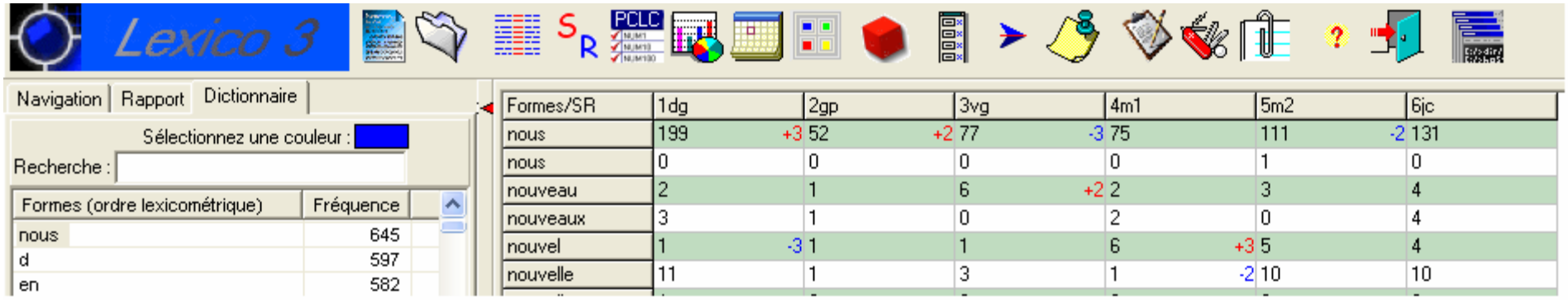
- Le nombre d'occurrences
- Les fréquences absolues et relatives
- Les mots outils
- Les hapax

3.1.2. Le nombre d'occurrences du corpus et de ses composantes

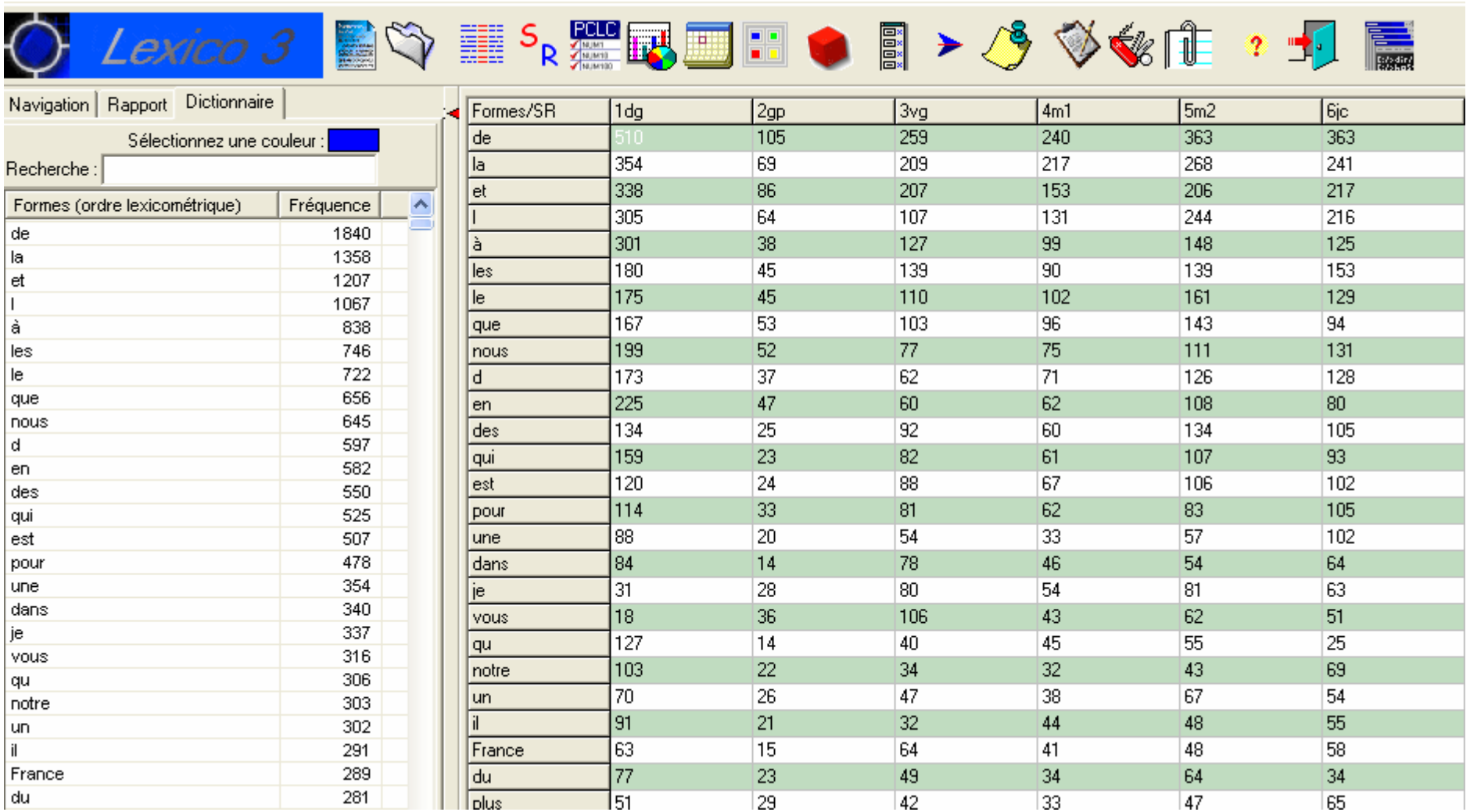


3.1.3. Fréquence de nous

```
Nb. d'occurrences du corpus :                              40189
Fréquence de nous :                                                645
Fréquence relative de nous :        645 / 40189     =      0.016
Cadence :                           40189 / 645     =      62
```



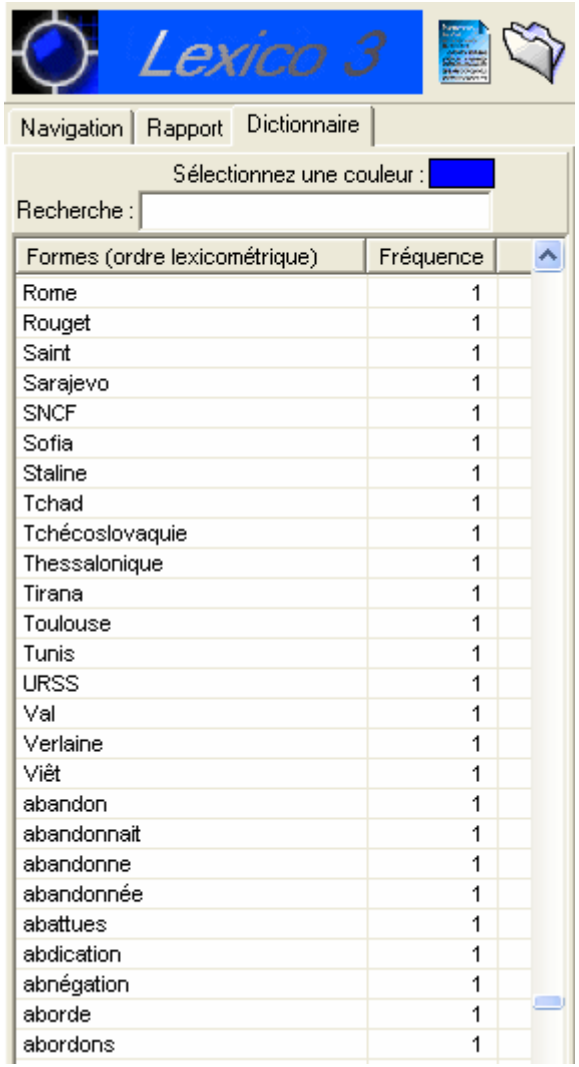
3.1.4. Les mots outils dans le tableau lexical entier



Formes (ordre lexicométrique)	Fréquence
de	1840
la	1358
et	1207
l	1067
à	838
les	746
le	722
que	656
nous	645
d	597
en	582
des	550
qui	525
est	507
pour	478
une	354
dans	340
je	337
vous	316
qu	306
notre	303
un	302
il	291
France	289
du	281

Formes/SR	1dg	2gp	3vg	4m1	5m2	6jc
de	510	105	259	240	363	363
la	354	69	209	217	268	241
et	338	86	207	153	206	217
l	305	64	107	131	244	216
à	301	38	127	99	148	125
les	180	45	139	90	139	153
le	175	45	110	102	161	129
que	167	53	103	96	143	94
nous	199	52	77	75	111	131
d	173	37	62	71	126	128
en	225	47	60	62	108	80
des	134	25	92	60	134	105
qui	159	23	82	61	107	93
est	120	24	88	67	106	102
pour	114	33	81	62	83	105
une	88	20	54	33	57	102
dans	84	14	78	46	54	64
je	31	28	80	54	81	63
vous	18	36	106	43	62	51
qu	127	14	40	45	55	25
notre	103	22	34	32	43	69
un	70	26	47	38	67	54
il	91	21	32	44	48	55
France	63	15	64	41	48	58
du	77	23	49	34	64	34
plus	51	29	42	33	47	65

3.1.5. Les vrais et les faux hapax



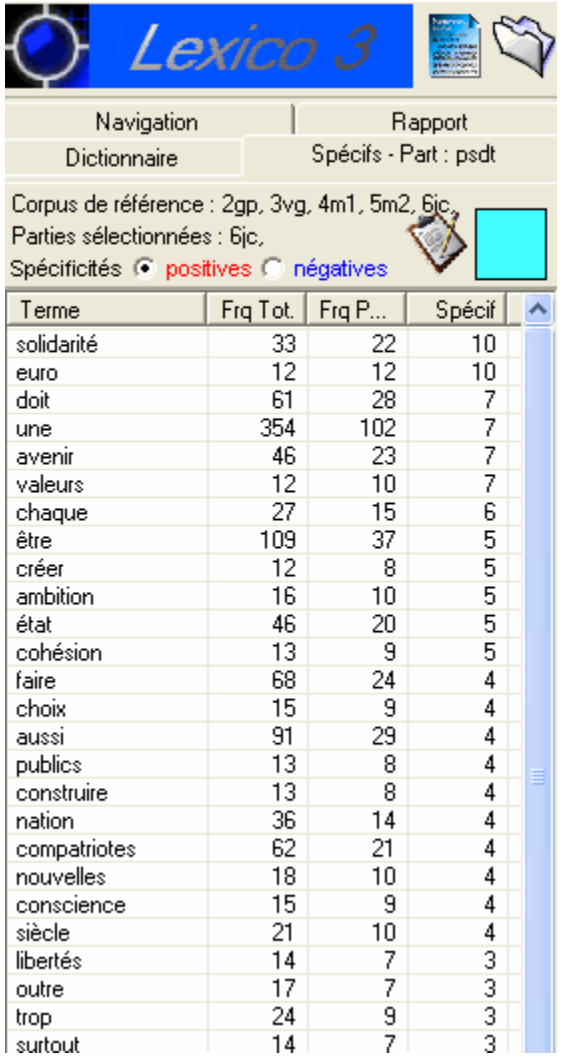
Formes (ordre lexicométrique)	Fréquence
Rome	1
Rouget	1
Saint	1
Sarajevo	1
SNCF	1
Sofia	1
Staline	1
Tchad	1
Tchécoslovaquie	1
Thessalonique	1
Tirana	1
Toulouse	1
Tunis	1
URSS	1
Val	1
Verlaine	1
Viêt	1
abandon	1
abandonnait	1
abandonne	1
abandonnée	1
abattues	1
abdication	1
abnégation	1
aborde	1
abordons	1

3.2. Les spécificités d'une partie

Les formes caractérisant une partie d'un corpus

- La loi hypergéométrique
- J.C. Spécificités positives et négatives
- J.C. Spécificités phrastiques

3.2.1. J.C. Spécificités positives

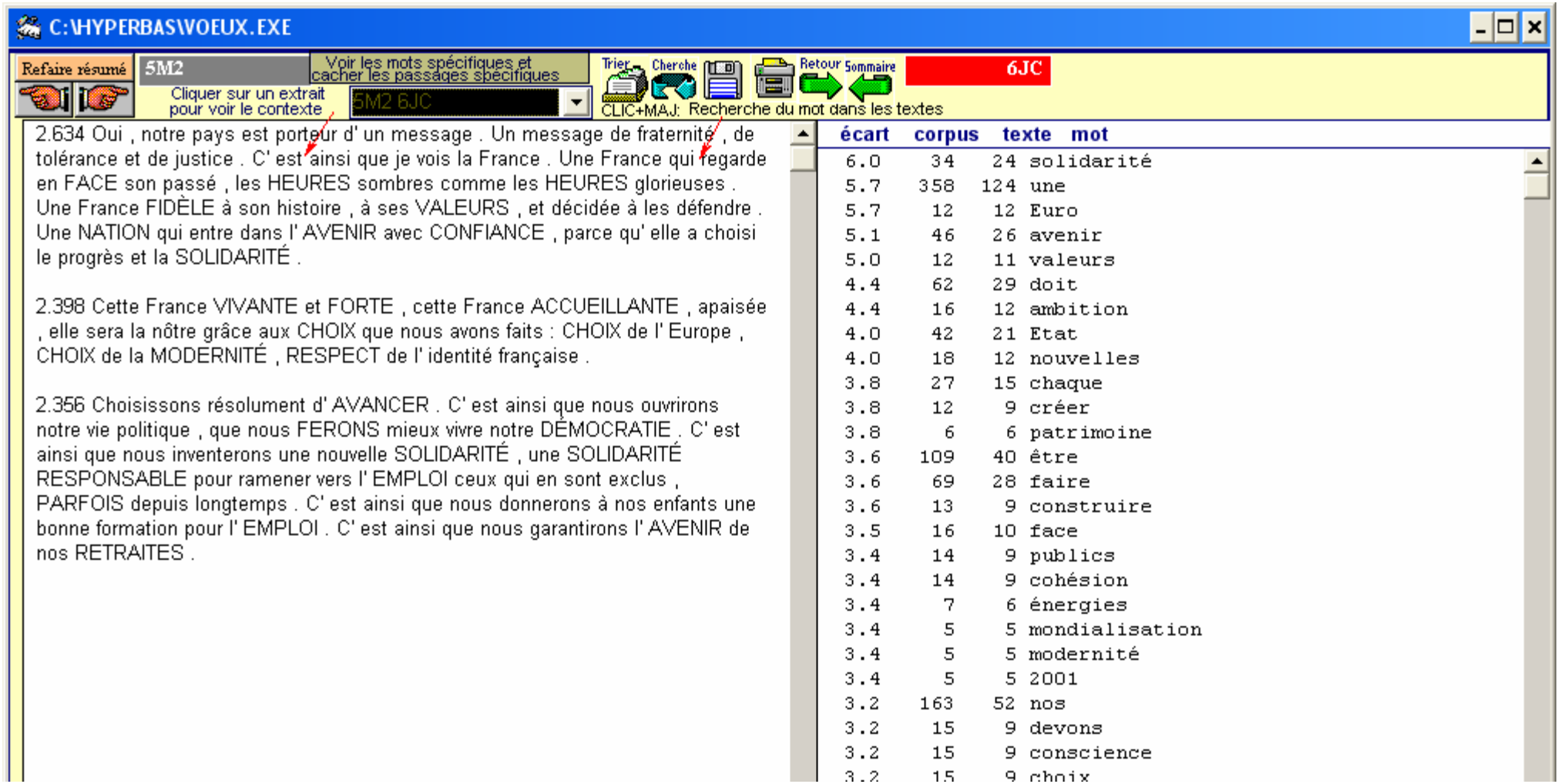


Terme	Frq Tot.	Frq P...	Spécif
solidarité	33	22	10
euro	12	12	10
doit	61	28	7
une	354	102	7
avenir	46	23	7
valeurs	12	10	7
chaque	27	15	6
être	109	37	5
créer	12	8	5
ambition	16	10	5
état	46	20	5
cohésion	13	9	5
faire	68	24	4
choix	15	9	4
aussi	91	29	4
publics	13	8	4
construire	13	8	4
nation	36	14	4
compatriotes	62	21	4
nouvelles	18	10	4
conscience	15	9	4
siècle	21	10	4
libertés	14	7	3
outre	17	7	3
trop	24	9	3
surtout	14	7	3

3.2.2. J.C. Spécificités négatives

Terme	Frq Tot.	Frq P...	Spécif
lui	41	3	-2
mois	34	2	-2
ans	42	3	-2
leurs	42	3	-2
on	112	10	-3
pays	118	11	-3
ils	66	5	-3
sans	85	8	-3
par	122	13	-3
que	656	94	-3
à	838	125	-3
a	261	34	-3
du	281	34	-3
depuis	38	2	-3
peuples	39	2	-3
celle	32	1	-3
ni	34	1	-3
françaises	39	2	-3
paix	62	5	-3
votre	58	4	-3
économique	45	3	-3
politique	52	3	-3
en	582	80	-4
qu	306	25	-7

3.2.3. J.C. Spécificités phrastiques

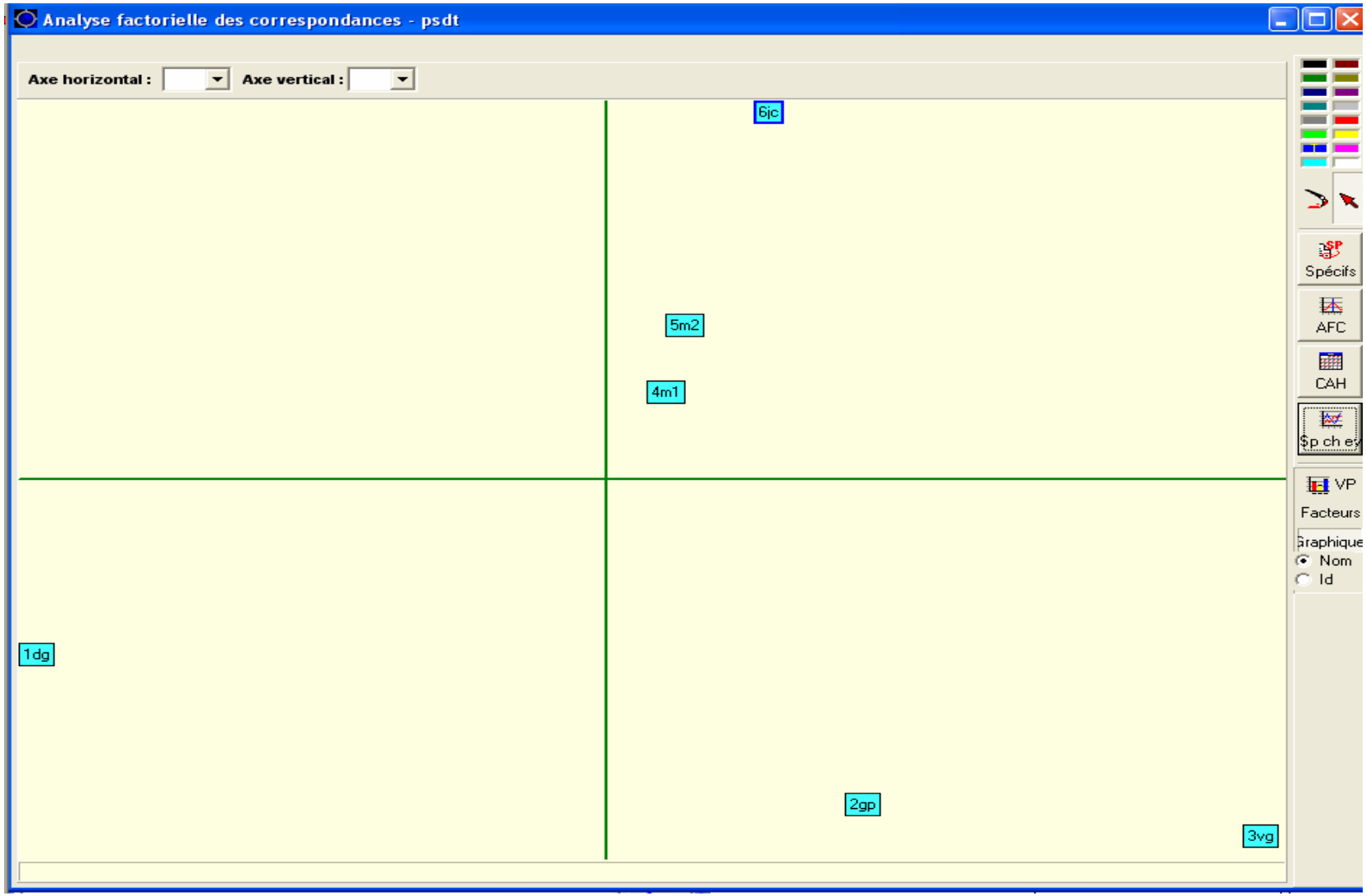


3.3. Rapprochement et éloignement des textes d'un corpus

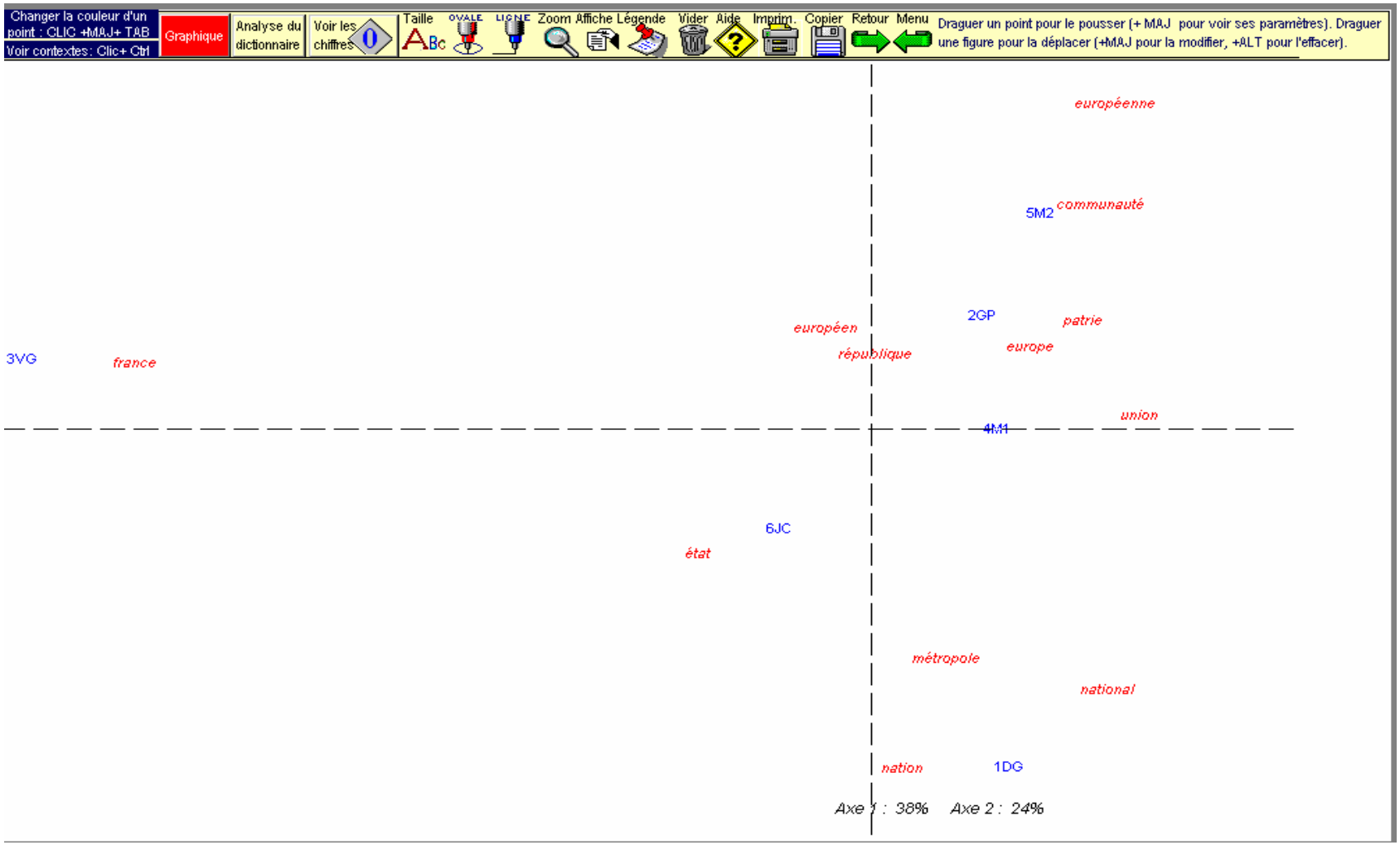
Oppositions, rapprochements et distances entre les textes d'un corpus

- **L'analyse factorielle du dictionnaire**
- **AFC sur une liste de mots**
- **Le classement automatique hiérarchisé**

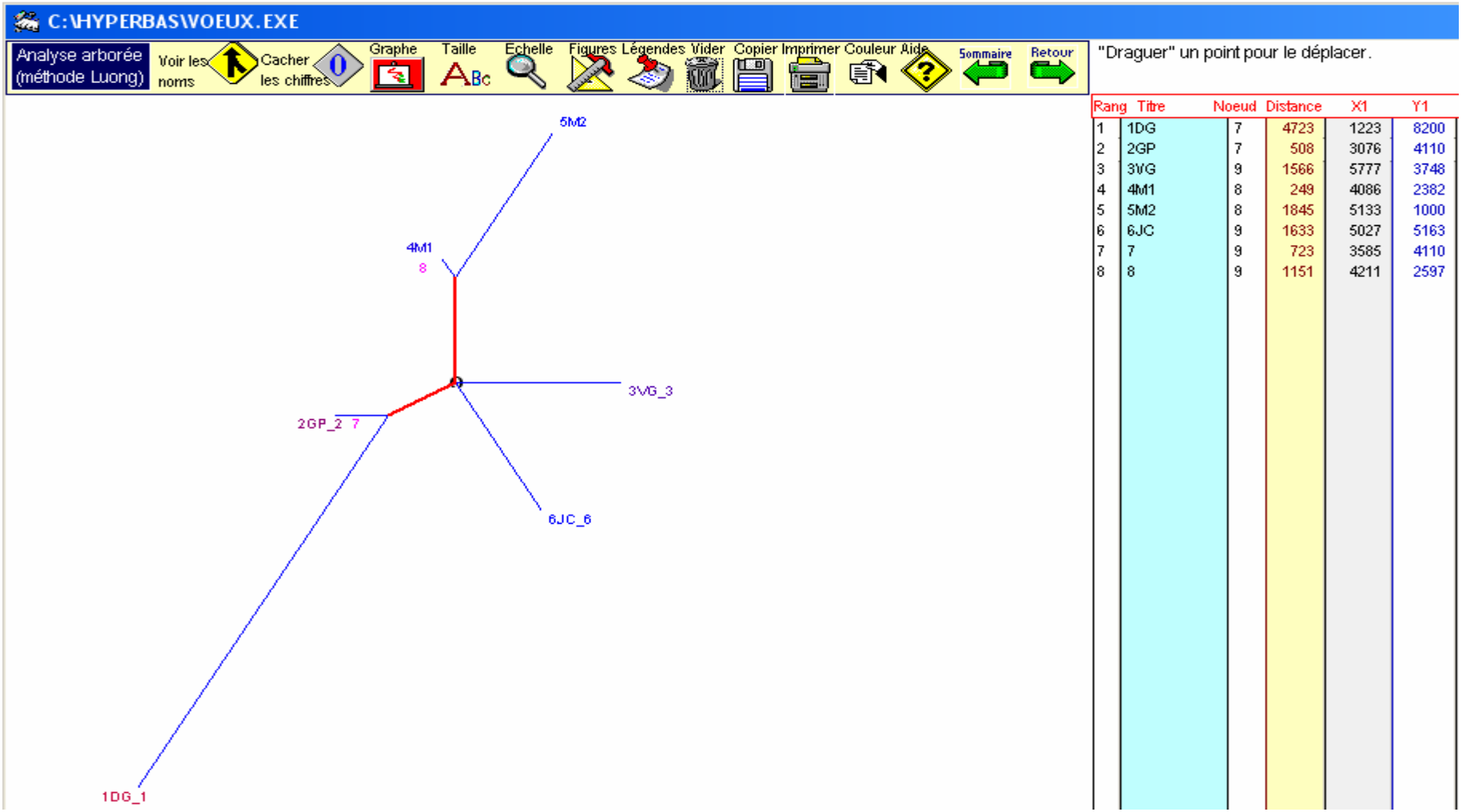
3.3.1. L'analyse factorielle du dictionnaire



3.3.2. AFC sur une liste de mots



3.3.3. Le classement automatique hiérarchisé



Rang	Titre	Noeud	Distance	X1	Y1
1	1DG	7	4723	1223	8200
2	2GP	7	508	3076	4110
3	3VG	9	1566	5777	3748
4	4M1	8	249	4086	2382
5	5M2	8	1845	5133	1000
6	6JC	9	1633	5027	5163
7	7	9	723	3585	4110
8	8	9	1151	4211	2597

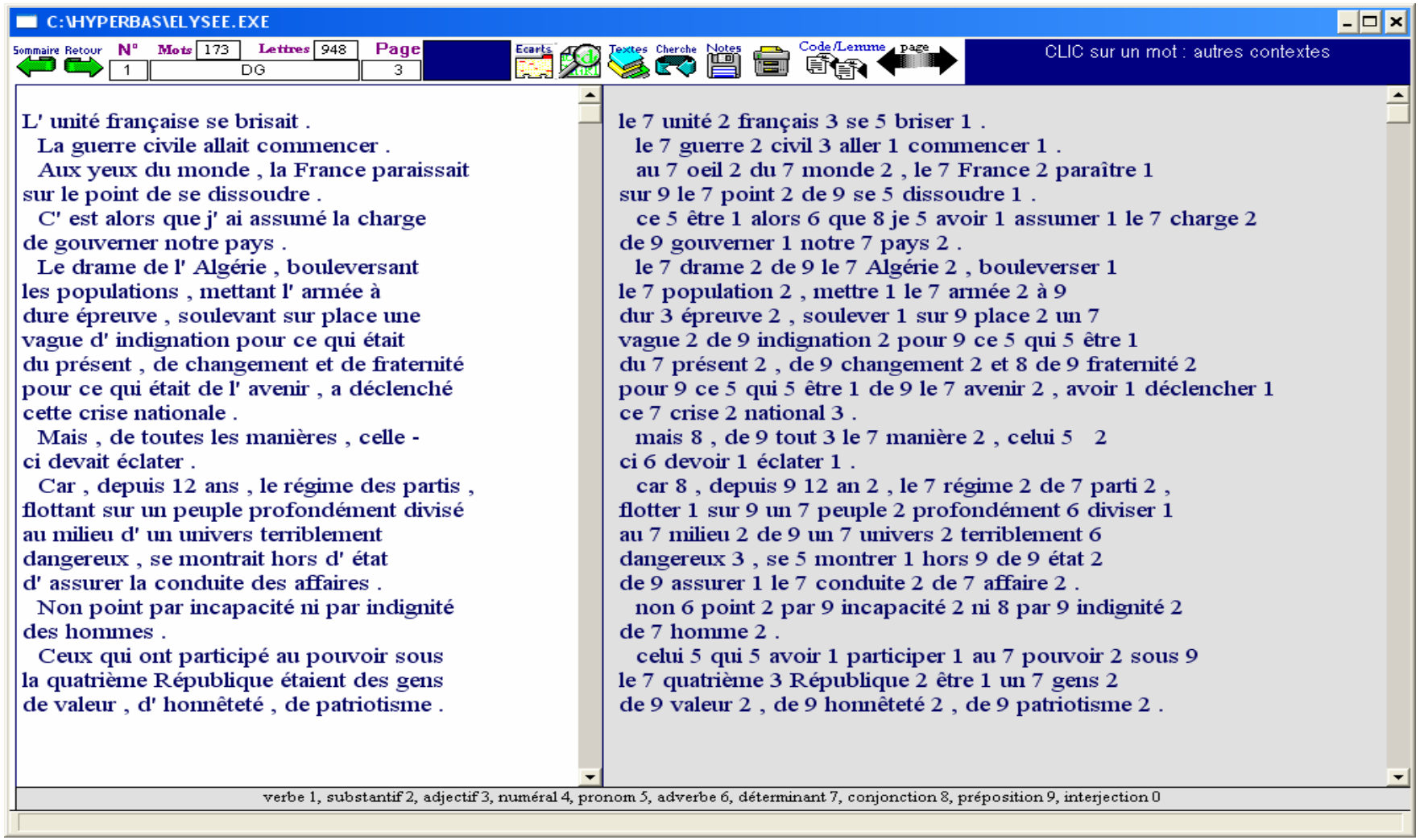
4. Une autre approche en gestation

- L'intégration de corpus et de textes formatés différemment dont ceux en xml (TEI)
- Les traitements préalables des textes : treetagger, cordial
- Un travail plus tourné sur des formes lemmatisées
- Des analyses plus fines des formes et des segments
- Un ou des logiciels exploitant un fond commun de petits outils

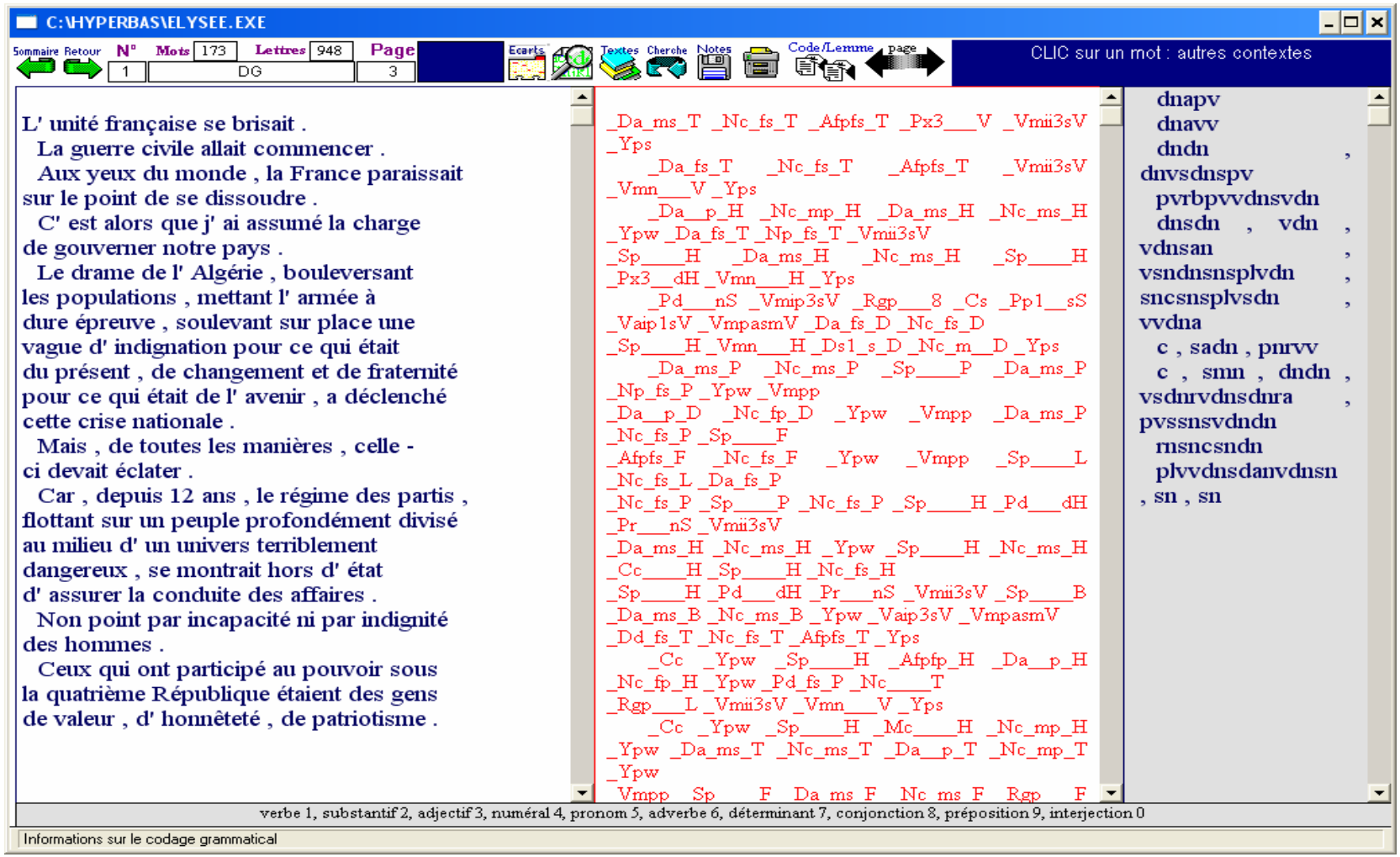
4.1. Lemmes et catégories grammaticales

- Hy : Formes non lemmatisées et lemmes
- Hy : Formes et catégories grammaticales
- Hy : concordance D N A V (det. nom. adj. ver.) nom = France
- Nooj : concordance (N, V, A) nom = France, verbe = être,

4.1.1. Formes et lemmes

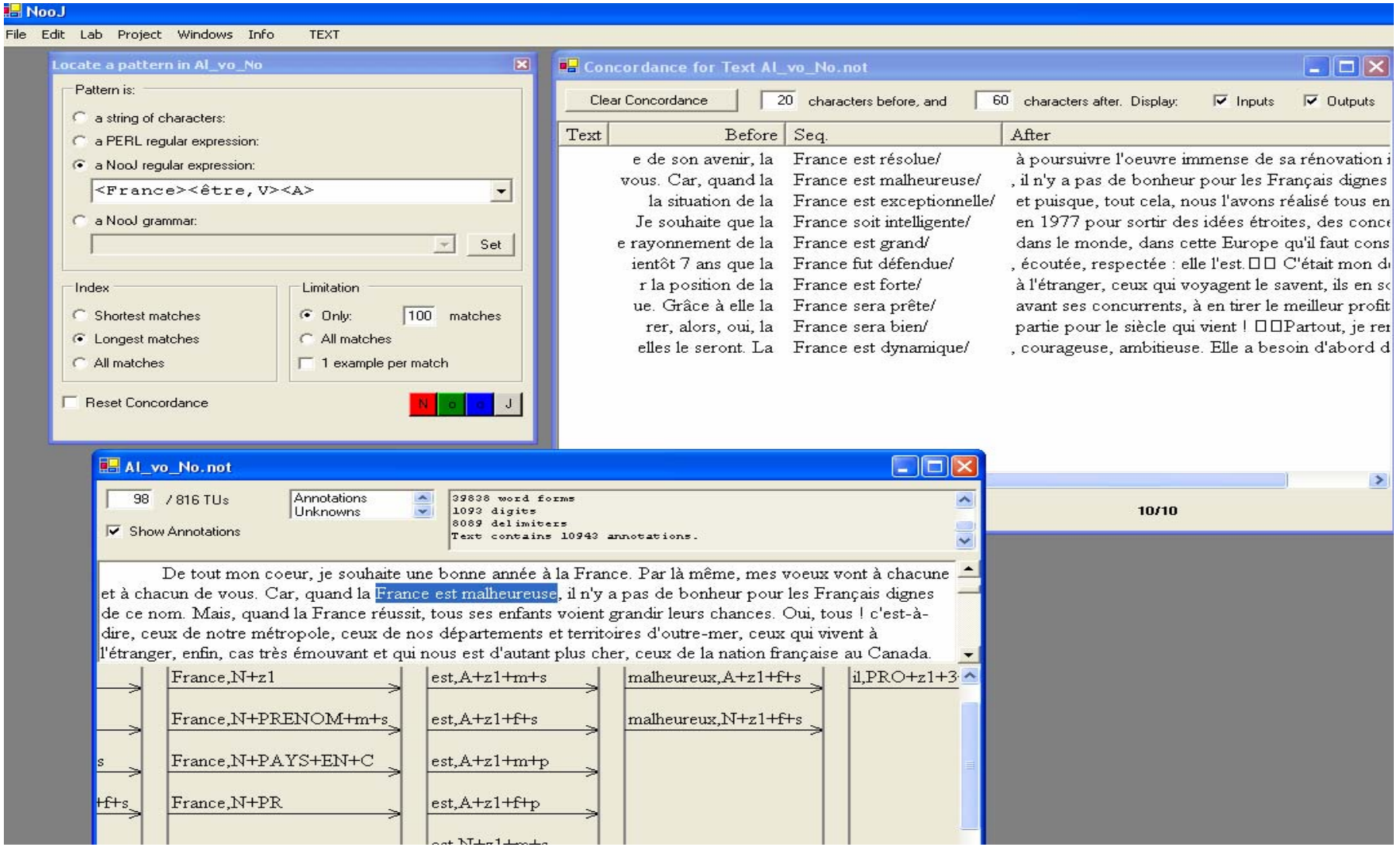


4.1.2. Formes et catégories



4.1.3. Concordance dnav n=France

4.1.4. Concordance (N, V, A) nom = France, verbe = être,



5. Quelques liens

- Logiciels
- Corpus

5.1. Logiciels

- Hyperbase : Etienne Brunet, UFR Lettres, 98 bd Herriot, 06204 Nice Fax : (33) 04 93 37 54 45 Courriel : brunet@unice.fr
- Lexico 3 http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/
- Weblex http://weblex.ens-lsh.fr/doc/weblex/index.html#sommaire
- Nooj http://perso.orange.fr/rosavram/pages/noojpag.html

5.2. Corpus

- Damon Mayaffre : Politext http://www.unice.fr/ILF-CNRS/politext/
- La documentation Française http://discours-publics.vie-publique.fr/rechlogos/servlet/RechServlet?_page=ACCUEIL&_type=NEW
- Textopol http://textopol.free.fr/
- Jean-Marc Leblanc http://www.univ-paris12.fr/www/labos/ceditec/leblanc.html/