

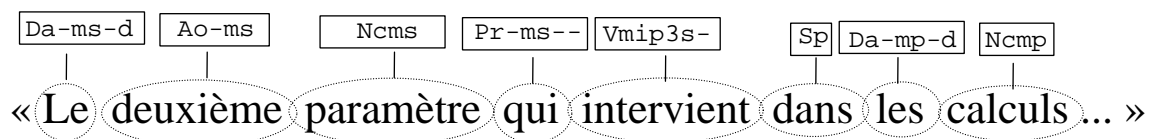
Lexicométrie sur corpus étiquetés

Bénédicte Pincemin
CNRS, LLI Univ. Paris 13

7es Journées internationales d'analyse statistique des données textuelles (JADT 2004)

Louvain-la-Neuve (Belgium), 10-12 mars 2004

Les corpus étiquetés



Le	le	Da-ms-d
deuxième	deuxième	Ao-ms
paramètre	paramètre	Ncms
qui	qui	Pr-ms--
intervient	intervenir	Vmip3s-
dans	dans	Sp
les	le	Da-mp-d
calculs	calcul	Ncmp

```
<w lemma='le' type='Da-ms-d'>Le</w>
<w lemma='deuxième' type='Ao-ms'>
deuxième</w>
<w lemma='paramètre' type='Ncms'>
paramètre</w>
<w lemma='qui' type='Pr-ms-'>qui</w>
<w lemma='intervenir' type='Vmip3s-'>
intervient</w>
<w lemma='dans' type='Sp'>dans</w>
<w lemma='le' type='Da-mp-d'>les</w>
<w lemma='calcul' type='Ncmp'>
calculs</w>
```

Quelle prise en compte proposent les logiciels lexicométriques ?

- Première solution : remplacer le texte
 - par le texte lemmatisé :
 - « *le deuxième paramètre qui intervenir dans le calcul ...* »
 - par les étiquettes :
 - « *Da-ms-d Ao-ms Ncms Pr-ms-- Vmip3s- Sp Da-mp-d Ncmp ...* »
- Deuxième solution : vues parallèles

Exemple : lecture parallèle graphie/lemme dans Hyperbase

The screenshot shows the Hyperbase software interface. The title bar reads "C:\HYPERBAS\FLAUCORR.EXE". The menu bar includes "Sommaire Retour", "N° Mots 179", "Lettres 910", "Page 3459", "Ecarté", "Textes Cherche", "Notes", and "Code/Lecture page". The main window is split into two panes. The left pane displays the original text, and the right pane displays the lemmatized text with morphological tags. At the bottom, a legend defines the tags: "verbe 1, substantif 2, adjectif 3, numéral 4, pronom 5, adverbe 6, déterminant 7, conjonction 8, préposition 9, interjection 0".

La lettre du père également est bonne .
Mais je ne vois pas de différence de caractère entre Mlle Lizel et Clémence .
On arrive à la proposition d' aller au bal masqué ;
très bien ;
et le lecteur s' attend à y suivre les personnages .
Pas du tout , on le mène à la campagne , et on le fait assister aux amours de deux personnages épisodiques !
Il y a là - dedans des détails gentils (bien que votre Frédéric parle tantôt comme un artiste : Quelle charmante courbe d' épaule et tantôt comme un notaire : Scellons ce pacte) .
Où diable avez - vous rencontré des gens qui disent :
Scellons ce pacte ?
Puis nous revenons au bal (juste au moment où l' on s' intéresse à vos deux enfants) et ce bal ne tient pas plus de place que le passage précédent .
Pourquoi n' avez - vous pas fait une description à fond de ce bal , puisqu' il a une importance décisive sur Jacqueline ?

le 7 lettre 2 du 7 père 2 également 6 être 1 bonne 3 .
mais 8 je 5 ne 6 voir 1 pas 6 de 9 différence 2 de 9 caractère 2 entre 9 Mlle 2 Lizel 2 et 8 Clémence 2 .
on 5 arriver 1 à 9 le 7 proposition 2 de 9 aller 1 au 7 bal 2 masqué 3 ;
très 6 bien 6 ;
et 8 le 7 lecteur 2 se 5 attendre 1 à 9 y 5 suivre 1 le 7 personnage 2 .
pas 6 du 7 tout 2 , on 5 le 5 mener 1 à 9 le 7 campagne 2 ,
et 8 on 5 le 5 faire 1 assister 1 au 7 amour 2 de 9 deux 4 personnage 2 épisodique 3 !
il 5 y 5 avoir 1 là 6 - dedans 6 de 7 détail 2 gentil 3 (bien 6 que 5 votre 7 Frédéric 2 parler 1 tantôt 6 comme 8 un 7 artiste 2 : quel 3 charmant 3 courbe 2 de 9 épaule 2 et 8 tantôt 6 comme 8 un 7 notaire 2 : sceller 1 ce 7 pacte 2) .
où 5 diable 2 avoir 1 - vous 5 rencontrer 1 un 7 gens 2 qui 5 dire 1 :
sceller 1 ce 7 pacte 2 ?
puis 8 nous 5 revenir 1 au 7 bal 2 (juste 6 au 7 moment 2 où 5 l' 5 on 5 se 5 intéresser 1 à 9 votre 7 deux 4 enfant 2) et 8 ce 7 bal 2 ne 6 tenir 1 pas 6 plus 6 de 9 place 2 que 8 le 7 passage 2 précédent 3 .
pourquoi 6 ne 6 avoir 1 - vous 5 pas 6 fait 3 un 7 description 2 à 9 fond 2 de 9 ce 7 bal 2 , puisque 8 il 5 avoir 1 un 7 importance 2 décisif 3 sur 9 Jacqueline 2 ?

verbe 1, substantif 2, adjectif 3, numéral 4, pronom 5, adverbe 6, déterminant 7, conjonction 8, préposition 9, interjection 0

Nombre de mots dans la page

Exemple : lecture parallèle graphie/code dans Hyperbase

The screenshot shows the Hyperbase application window titled "C:\HYPERBAS\FLAUCORR.EXE". The top menu bar includes "Sommaire Retour", "N° Mots 160", "Lettres 759", "Page 2236", and "CLIC sur un mot: autres contextes". Below the menu are icons for "Ecartis", "Textes", "Cherhe", "Notes", and "Code/Lemme page".

The main window is split into two panes. The left pane displays a text document with the following content:

À LOUISE COLET .
Entièrement inédite en 1927 .
Mercredi soir .
Janvier 1854] .
Qu' est - ce que Bouilhet me conte ?
Je n' y comprends goutte !
Il me dit que tu te plains de n' avoir
pas de lettres de moi , que je t' oublie ,
etc ...
Si je n' avais la tête vissée d' aplomb
sur les épaules , voilà de ces choses
qui me la feraient tourner .
En fait de lettres , celle - ci est la
troisième depuis vendredi .
Or , à moins que de s' écrire tous les
jours , je ne vois guère moyen de s' écrire
plus souvent .
Tu as dû avoir une lettre de moi samedi .
Dimanche le paquet du Crocodile , dont
tu ne m' as pas même fait la gracieuseté
de m' accuser réception , et ce matin
tu as dû avoir encore une lettre écrite
avant - hier .

The right pane displays a list of morphological codes in red text, such as `_Sp_Np Js _Np_s _Yps`, `_Rgp_AfpS _Sp_Nc_m _Yps`, etc. Below the list, a legend identifies the codes: "verbe 1, substantif 2, adjectif 3, numéral 4, pronom 5, adverbe 6, déterminant 7, conjonction 8, préposition 9, interjection 0". At the bottom, it says "Pour aller à un autre endroit du corpus".

Exemple : spécificités parallèles graphie/lemme dans Hyperbase

The screenshot shows the Hyperbase application window titled "C:\HYPERBAS\FLAUCORR.EXE". The top menu bar includes "Refaire résumé", "1846(forme)", "Mots Phrases Codes Syntaxe", "Cherhe", "Trier", "Sommaire", and "1846(lemme)". Below the menu are icons for "Cherhe", "Trier", "Sommaire", and "CLIC+MAJ: Recherche du mot dans les textes".

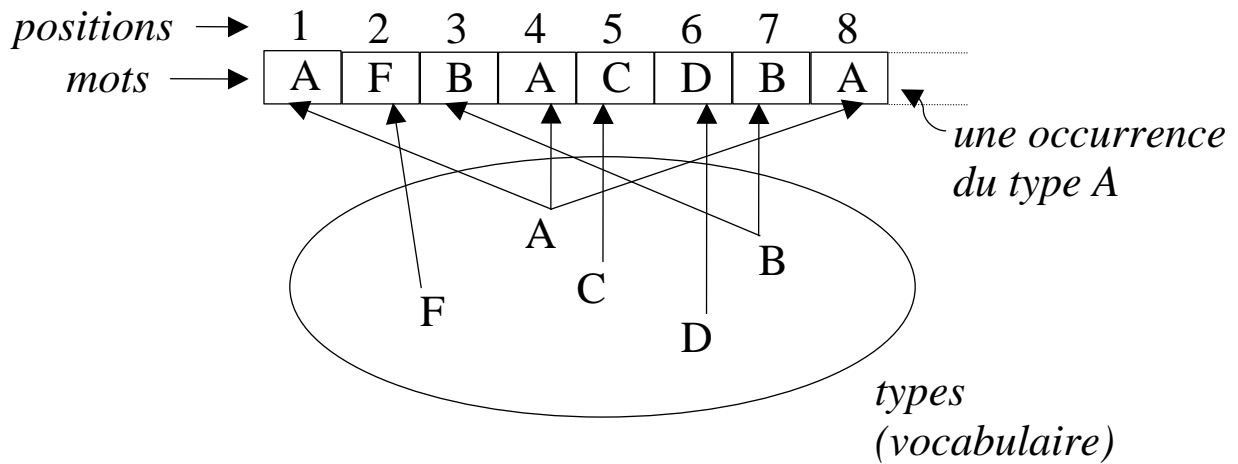
The main window displays a table with two columns of data. The left column is titled "Refaire résumé" and the right column is titled "1846(lemme)". Both columns have the same headers: "N°", "écart", "corpus", "texte", and "mot".

N°	écart	corpus	texte	mot	N°	écart	corpus	texte	mot
3	32.5	5518	1082	tu	3	32.6	5518	1082	tu 5
3	21.8	79	78	1846	3	25.9	5025	890	te 5
3	20.4	5367	825	;	3	21.8	79	78	1846
3	18.8	2312	438	t'	3	20.4	5367	825	;
3	18.2	2268	424	te	3	19.8	1254	307	aimer 1
3	15.6	1681	317	toi	3	15.6	3396	518	ton 7
3	14.6	404	126	amour	3	15.6	1681	317	toi 5
3	13.0	84	51	1926	3	14.8	447	135	amour 2
3	12.2	610	141	tes	3	13.0	84	51	1926
3	12.1	52	38	Phidias	3	12.1	52	38	Phidias 2
3	11.5	3464	459	moi	3	11.4	3452	456	moi 5
3	11.2	862	166	coeur	3	11.2	911	172	coeur 2
3	11.1	700	144	aime	3	10.6	1815	272	vouloir 1
3	10.4	59	35	aimes	3	10.5	248	74	baiser 2
3	10.2	15174	1520	que	3	9.9	30811	2858	je 5
3	9.6	889	155	as	3	9.4	4870	562	le 5
3	9.6	21017	2009	je	3	9.4	1905	266	quand 8
3	9.4	1907	266	quand	3	8.7	276	68	Colet 2
3	9.2	497	102	veux	3	8.2	182	51	Adieu 2
3	8.7	276	68	Colet	3	8.0	333	72	Louis 2
3	8.3	369	79	Louise	3	7.7	9221	920	ce 5
3	7.8	400	80	es	3	7.7	7931	806	que 5
3	7.5	1291	179	ta	3	7.4	13165	1254	me 5
3	7.4	1032	150	toujours	3	7.4	1032	150	toujours 6
3	7.2	3511	394	si	3	7.1	12304	1176	que 8
3	7.1	8539	847	pas	3	7.0	14584	1369	ne 6
3	6.9	42	20	reproches	3	6.9	3137	353	pouvoir 1
3	6.9	126	36	aimer	3	6.7	90	29	bouche 2
3	6.8	135	37	baiser	3	6.6	7451	738	qui 5
3	6.7	90	29	bouche	3	6.5	95	29	larme 2
3	6.7	53	22	souviens	3	6.5	185	43	doux 3
3	6.6	167	41	baisers	3	6.4	8222	801	pas 6
3	6.5	7451	738	qui	3	6.4	78	26	tendre 3
3	6.5	1698	209	ton	3	6.4	4333	456	en 5
3	6.4	7891	774	ne	3	6.1	305	57	âme 2
3	6.4	485	81	étais	3	6.1	298	56	enfant 2
3	6.3	92	28	larmes	3	6.1	168	39	oh 0
3	6.3	191	43	enfant	3	6.1	1667	201	si 8

At the bottom, it says "Menu déroulant pour le choix des textes".

3e solution : texte multidimensionnel (1)

La vision lexicométrique du texte :



3e solution : texte multidimensionnel (2)

Généralisation du modèle lexicométrique aux textes étiquetés

1	2	3	4	5	6	7	8
Le	deuxième	paramètre	qui	intervient	dans	les	calculs
le	deuxième	paramètre	qui	intervenir	dans	le	calcul
Da-ms-d	Ao-ms	Ncms	Pr-ms--	Vmip3s-	Sp	Da-mp-d	Ncmp

- en colonnes : les positions
- en lignes : les dimensions de codage

Texte multidim. dans Weblex

- Sélection du « pivot » d'un calcul par équation CQP (Christ 1994), avec croisement de plusieurs dimensions (**graphies**, **lemmes**, **codes gramm....**)
 - adv. en *-ment* (index) : [p2="R.*" & word=".*ment"]
 - être non aux. conditionnel 3e pers. suivi d'un adj. qual. (conc.) : [p3="être" & p2="Vmc.*3.*"][p2="Af.*"]
 - nom plu. non terminé par -s et non invariable (index) : [p2="N..p" & word!=".*s" & p3!=word]
- Et toujours, choix d'une dimension d'analyse.

Exemples d'interrogations mono- et multidimensionnelles

<i>dimension d'analyse</i>	<i>sélection</i>	<i>calcul</i>
lemmes	code = N-Adj ou Adj-N	index
codes	code = préposition	lexicogramme

Exemple : index de lemmes avec une sélection sur les codes

Weblex 3.0 [Frames] : Corpus socio - Mozilla

file:///E:/data/benie/baluchon/documents/jadt/wlx/i_nadj/i_nadj.html

Index de la propriété p3 des occurrences de $((p2="Af.*")[p3="quartier"])|((p3="quartier")[p2="Af.*"])$ dans le corpus socio

ord	f	événement
1	87	quartier populaire
2	17	quartier ancien
3	12	quartier difficile
4	9	même quartier
5	9	quartier sensible
6	6	autre quartier
7	6	quartier prioritaire
8	5	beaux quartier

Corpus: socio

Source A: $((p2="Af.*")[p3=""])|((p3="")|p2="Af.*"))$

Source B:

(CQP : Index Concordances Références Contextes Répartition Spécificités)

Vocabulaire Répartitions Pareto Zipf Longueur des phrases Dimensions

Lexicogramme Lexicogramme récursif Lexicogrammes récursifs Cooccurrences Segments répétés Termes

Spécificités total alpha Spécificités total fréq Spécificités parties

Général

Index

Concordances

Vocabulaire

Affichage de l'index et des éditions

Composer l'index avec les champs : forme p2 p3 p4 p5 p6 p7 p8 p9

Format : Tabulé Linéaire

Mot simple ou expression CQP

Exemple : lexicogramme sur les codes grammaticaux

Weblex 3.0 [Frames] : Corpus socio - Mozilla

file:///E:/data/benie/baluchon/documents/jadt/wlx/l_sp/l_sp.html

Lexicogramme du pôle "Sp" dans le corpus socio (p2)

Seuils : f3, cf3, p 5.0E-2, d_m 1.0

Sp
(171807)

cooccurents gauches					cooccurents droits				
f	cf	p	d _m		f	cf	p	d _m	
Vmis2s	20	19	2e-02	0.1	Pd-..d	589	588	4e-74	0.1
					Pd-mpd	410	408	3e-49	0.2
					Pr-fsd	318	317	2e-39	0.0
					Pi-mpd	317	314	6e-36	0.4
					Pi-msd	304	298	2e-30	0.2

Corpus: socio

Source A: Sp

Source B:

(CQP : Index Concordances Références Contextes Répartition Spécificités)

Vocabulaire Répartitions Pareto Zipf Longueur des phrases Dimensions

Lexicogramme Lexicogramme récursif Lexicogrammes récursifs Cooccurrences Segments répétés Termes

Spécificités total alpha Spécificités total fréq Spécificités parties

Général

Index

Concordances

Vocabulaire

Options générales de l'affichage

Afficher les résultats dans output

Listes sous forme de Tableau (forcer)

Done

Limites des dimensions de codage

- codes grammaticaux : émiettement (ex. verbes)
- lemmes : graphie du lemme seule peu intéressante (confusion des homographes alors même qu'on voulait la finesse linguistique)
- graphies fléchies : pas complètement représentative de l'approche traditionnelle puisque découpage non purement typographique (par caractères délimiteurs)
 - ex. : *est-ce que* (3 mots par caractères délimiteurs, 1 mot pour certains analyseurs)

Dimensions élémentaires à combiner en dimensions d'analyse

		1	2	3	4	5	6	...
partie du discours	graphie fléchie							
	graphie du lemme							
	catégorie							
	sous-catégorie							
	trait							
flexion	genre							
	nombre							
	personne							
conjugaison	mode							
	temps							
	aux. (tps composé)							

Dimensions d'affichage (1)

- 1er cas : Résultats = occurrences en contexte
 - ex. : extraction de contextes, concordance, surlignages
 - affichage des graphies (lisibles, et désambiguïsées par le contexte)
 - pour les occurrences concernées, indication de la dimension d'analyse et des dimensions d'affichage complémentaires
- Ex. (dim. d'analyse = catégorie et ss-cat.) :
 - « Le deuxième paramètre **qui [Pr]** intervient dans les calculs... »

Dimensions d'affichage (2)

- 2e cas : Résultats = liste de types
 - ex. : vocabulaire, index, spécificités, cooccurrents
 - affichage de la dimension d'analyse
 - sur des colonnes complémentaires, valeurs correspondantes pour les dimensions d'affichage demandées

Dimensions d'affichage (3)

Exemple :

plutôt que liste de lemmes seuls

ou liste détaillée de toutes les flexions,

liste de lemmes avec quelques indications de flexion :

<i>fréq.</i>	<i>lemme</i>	<i>mode</i>	<i>temps</i>	<i>personne</i>
5502	pouvoir	i, c, s, f	p, f, i, s	3, 1, 2
3211	devoir	i, c, s	p, f, i	3, 1, 2
1877	falloir	i, c	p, f, i	3
576	croire	i	p, i	1, 3

Exemple : liste de lemmes

The screenshot shows the Weblex 3.0 interface. The main content area displays the following table:

ord	f	événement
1	5502	pouvoir
2	3211	devoir
3	1877	falloir
4	576	croire
	11166	au Total

Below the table, there are search filters and a navigation menu. The search criteria are: `[p3='pouvoir' & p2='V.*'] [p3='devoir' & p2='V.*'] [p3='falloir'] [p3='croire']`. The navigation menu includes options like 'Index', 'Concordances', 'Références', 'Contextes', 'Répartition', 'Spécificités', 'Vocabulaire', 'Répartitions', 'Pareto', 'Zipf', 'Longueur des phrases', 'Dimensions', 'Lexicogramme', 'Lexicogramme récursif', 'Lexicogrammes récursifs', 'Cooccurrences', 'Segments répétés', 'Termes', 'Spécificités total alpha', 'Spécificités total fréq', and 'Spécificités nantes'. The 'Affichage de l'index et des éditions' section is active, showing options for 'Composer l'index avec les champs' and 'Format' (Tabulé or Linéaire).

Exemple : liste des lemmes + flexions

Index des propriétés p2/p3 des occurrences de [p3="pouvoir" & p2="V.*"]/[p3="devoir" & p2="V.*"]/[p3="falloir"]/[p3="croire"] dans le corpus socio

ord	f	événement
1	2392	Vmip3s/pouvoir
2	1455	Vmip3s/falloir
3	1359	Vmip3s/devoir
4	635	Vmip3p/pouvoir
5	578	Vmip3p/devoir
6	551	Vmcc3s/pouvoir
7	317	Vmpasm/pouvoir
8	315	Vmcc3s/devoir

Corpus: socio

Source A: [p3='pouvoir' & p2='V.*']/[p3='devoir' & p2='V.*']/[p3='falloir']/[p3='croire']

Source B:

(CQP : Index Concordances Références Contextes Répartition Spécificités)

Vocabulaire Répartitions Pareto Zipf Longueur des phrases Dimensions

Lexicogramme Lexicogramme récursif Lexicogrammes récursifs Cooccurrences Segments répétés Termes

Spécificités total.alpha Spécificités total.fréq Spécificités parties

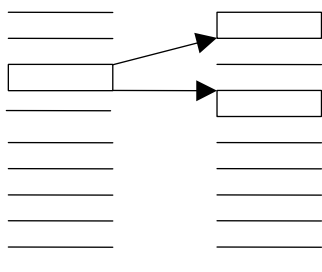
Affichage de l'index et des éditions

● Composer l'index avec les champs : forme p2 p3 p4 p5 p6 p7 p8 p9

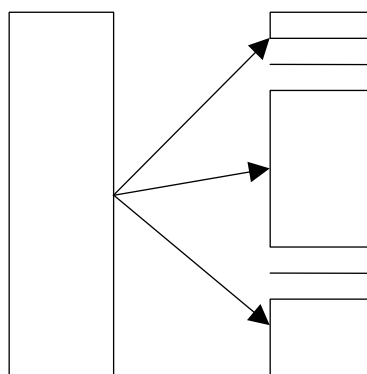
● Format : Tabulé Linéaire

Affichage : projections

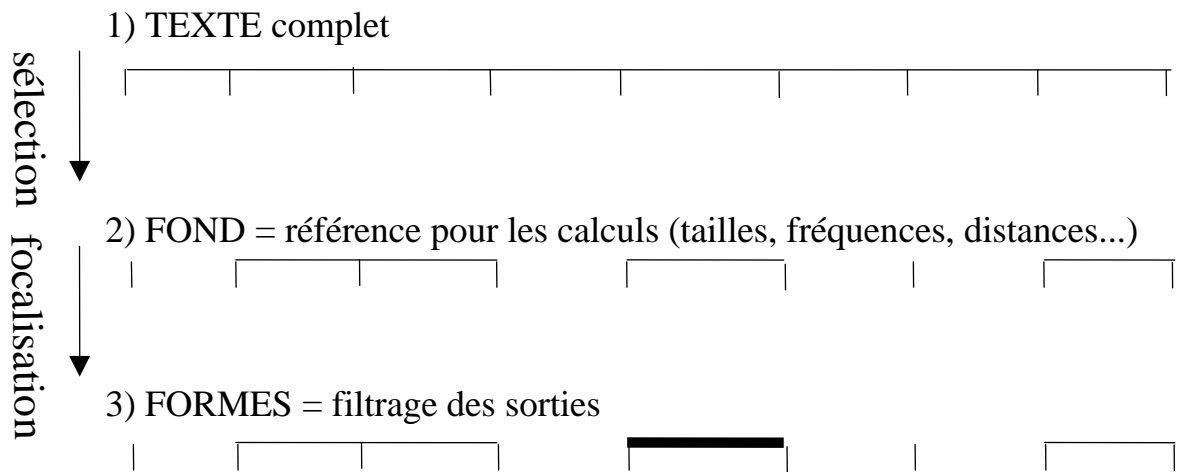
d'un élément :



d'une liste entière :



Et l'axe syntagmatique du tableau ?



Le deuxième paramètre qui intervient dans les calculs

Expression des filtres : l'exemple d'Hyperbase pour les informations morphosyntaxiques

C:\HYPERBAS\FLAUCORR.EXE

Catégorie 1	Sous-cat.2	Mode 3	Temps 4	Personne 5
Verbe V	principal m	Infinitif n	Présent p	1re pers. 1
	auxiliaire a	Indicatif i	Imparfait i	2e pers. 2
		Subjonctif s	Passé s	3e pers. 3
		Conditionnel c	Subjonctif présent r	
		Impératif f	Subjonctif imparfait m	
		Participe p	Participe passé a	
Substantif N	nom commun c			
	nom propre p			
Adjectif A	qualificatif f	Positif p		
		Comparatif c		
	ordinal o			
Déterminant D	article a			
	démonstratif d			
	interrogatif i			
	indéfini t			
Pronom P	pers. réfléchi x	1re personne 1		
	pers. non réfléchi p	2e personne 2		
	possessif s	3e personne 3		
	démonstratif d			
	interrogatif t			
	indéfini i			
	relatif r			

Code choisi 1 2 3 4 5 6 7

Afc

Adjectif, qualificatif, compar. ou superl.

Continuer

Effacer

Fonction 7

Cliquer sur les critères souhaités puis sur le bouton CONTINUER (les boutons bleus donnent accès à la série entière)

B - attribut du sujet
C - objet direct
D - groupe objet direct
E - objet indirect
F - groupe objet indirect
G - complément d'agent
H - circonstanciel
K - circ. de temps
L - circ. de lieu
M - apposition
N - groupe apposition
O - apostrophe
P - groupe apostrophe
Q - compl. de négation
S - sujet
T - groupe sujet
U - pronominalisation
V - base de proposition
Y - sujet réel
Z - groupe sujet réel
1 - ajout à l'adjectif
2 - reprise du COD
3 - reprise du COI
4 - reprise du circonst.
5 - ajout au nom
6 - ajout au pronom
7 - reprise du sujet
8 - ajout au verbe

Genre 4

Masculin m

Féminin f

Numéral Mo

Interjection Ij

Préposition Sp

Adverbe R

comparatif gc

négation pn

autre gp

Conjonction C

coordination c

subordination s

Ponctuation Y

pause pw

finale ps

insertion po

fin insert pc

autre ss

Nombre 5-6

Singulier s

Pluriel p

Fonction 6

sujet n

objet direct a

objet indirect d

Choisir la combinaison souhaitée. Un clic sur une option sert alternativement à activer ou désactiver la sélection. Les options inscrites dans la zone bleue sont réservées aux verbes. Certaines autres aux adjectifs ou aux pronoms. Les options 4 (genres), 5-6 (nombre) et 7 (fonction) concernent toutes les parties du discours, sauf les invariables. Le programme interdit les choix incohérents. Une fois réalisée la sélection, cliquer sur CONTINUER pour la transmettre au traitement en cours. (Le numéro des options indique la colonne intéressée dans le code).

Afc

En résumé :

étapes d'un calcul lexicométrique

- 1) Sélection fond / texte
- 2) Détermination d'une dimension d'analyse
- 3) Focalisation :

pour un élément
(une position) :

repérage	
(dimension d'affichage)	ordre de tri
(dimension d'affichage)	ordre de tri

formation de motifs :

