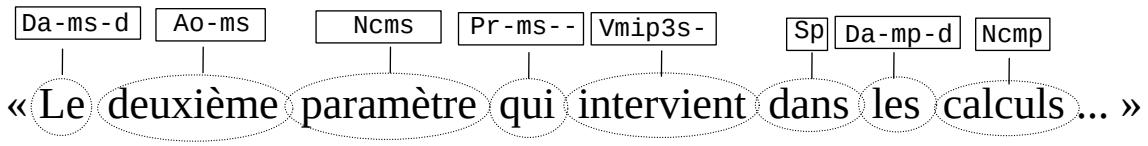
Lexicométrie sur corpus étiquetés

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



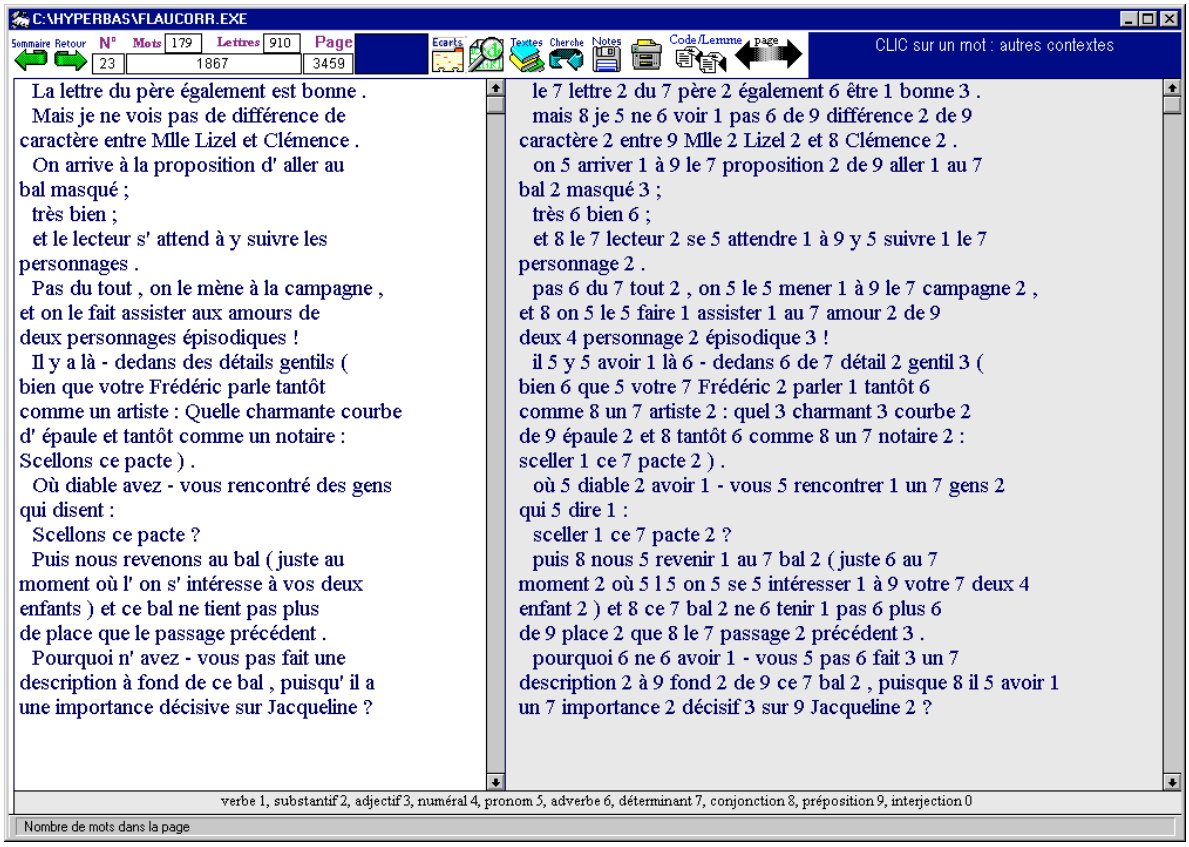
Le	le	Da-ms-d
deuxième	deuxième	Ao-ms
paramètre	paramètre	Ncms
qui	qui	Pr-ms--
intervient	intervenir	Vmip3s-
dans	dans	Sp
les	le	Da-mp-d
calculs	calcul	Ncmp

```
<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>
```

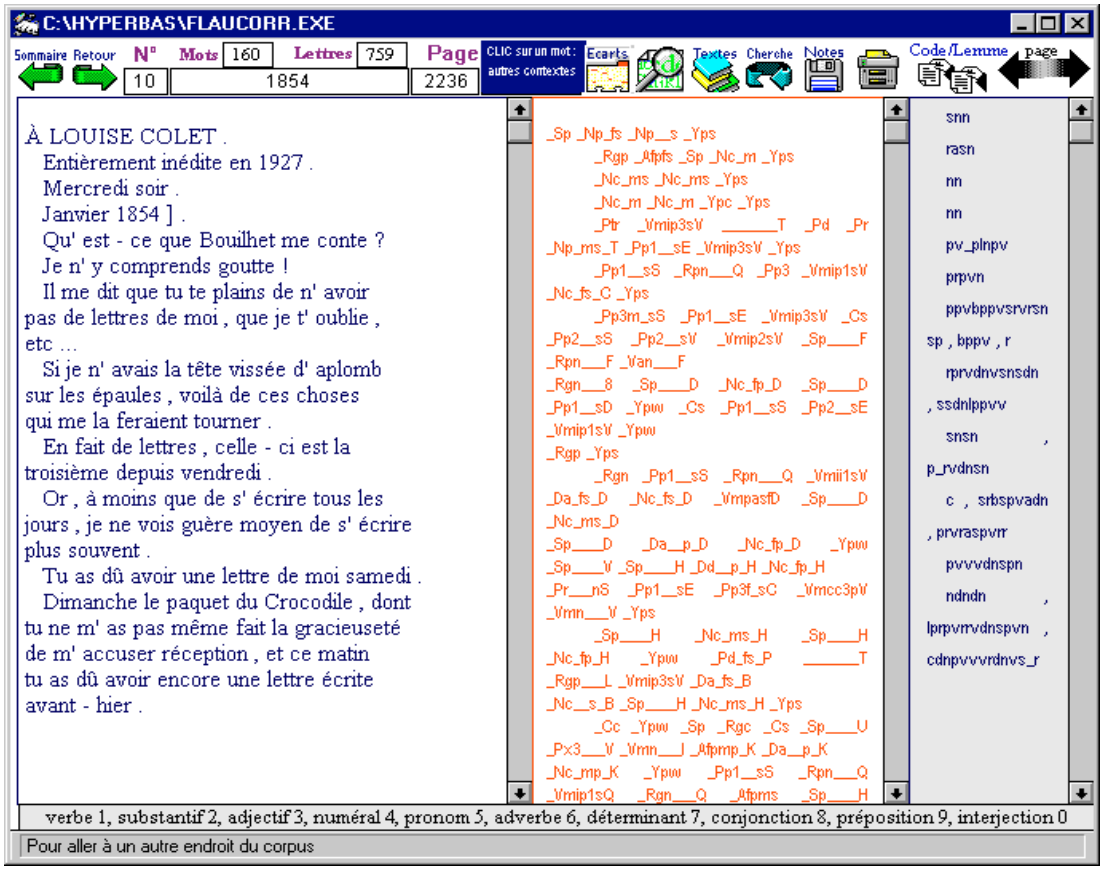
Quelle prise en compte proposent les logiciels lexicométriques ?

- 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

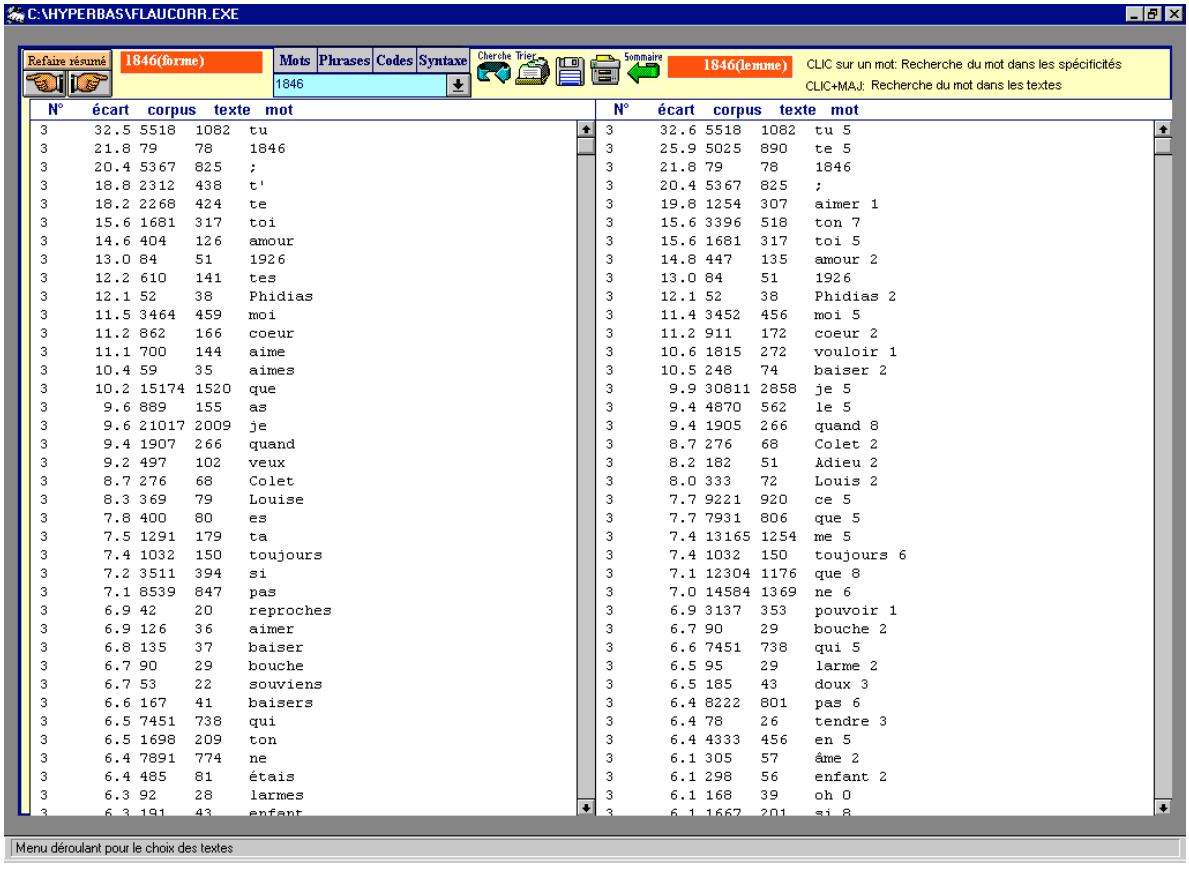
Exemple : lecture parallèle graphie/lemme dans Hyperbase



Exemple : lecture parallèle graphie/code dans Hyperbase

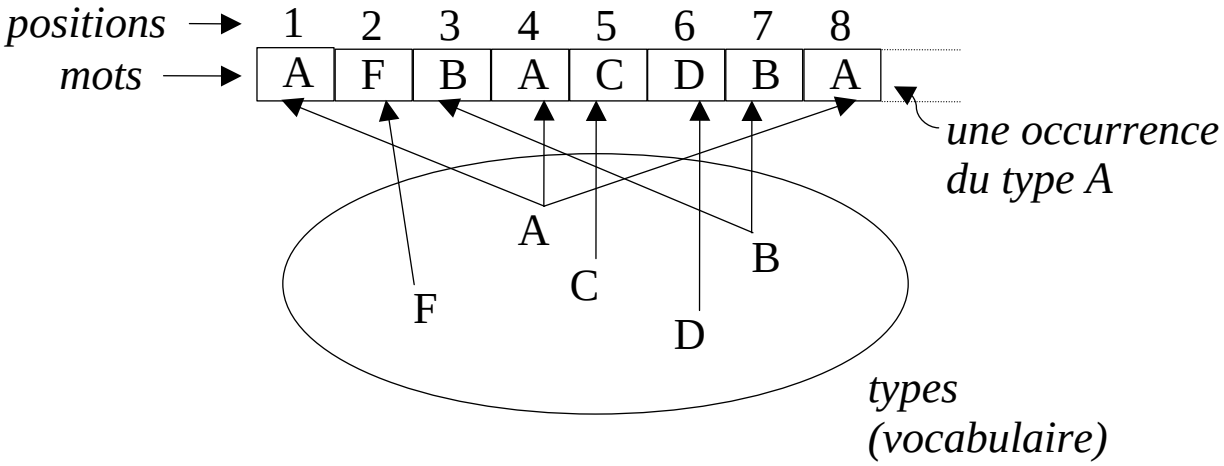


Exemple : spécificités parallèles graphie/lemme dans Hyperbase



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

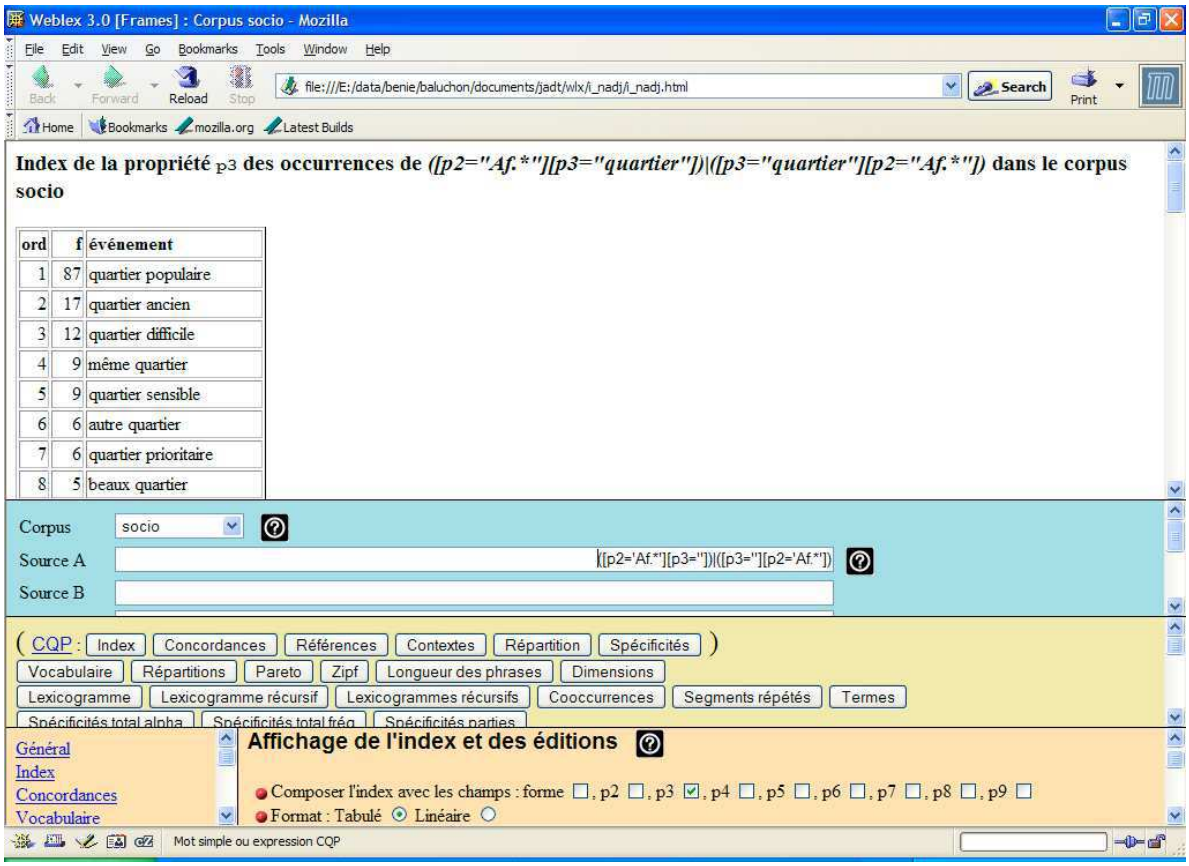
Texte multidim. dans Weblex

- 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.

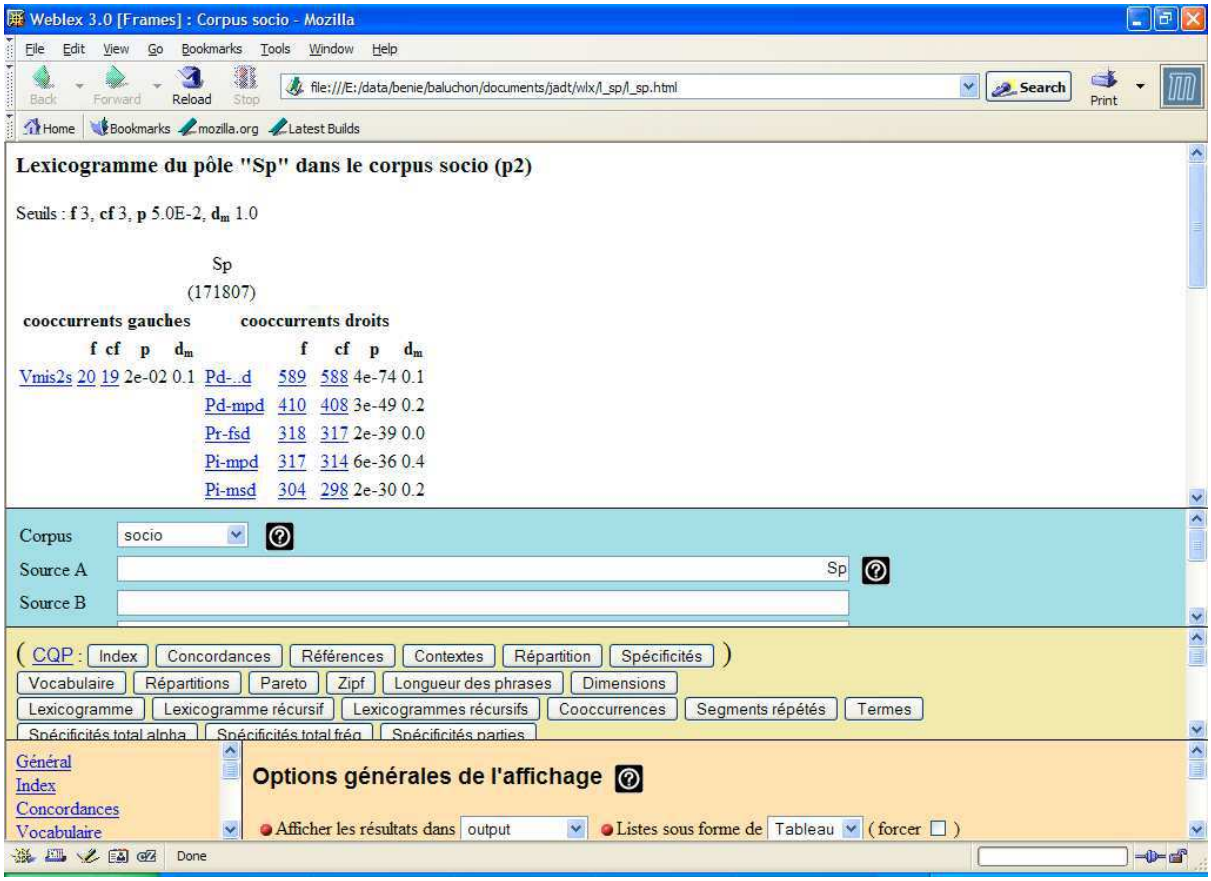
Exemples d'interrogations mono- et multidimensionnelles

dimension d'analyse	***sélection***	***calcul***
lemmes	code = N-Adj ou Adj-N	index
codes	code = préposition	lexicogramme

Exemple : index de lemmes avec une sélection sur les codes



Exemple : lexicogramme sur les codes grammaticaux



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	...
	graphie fléchie							
	graphie du lemme							
partie du discours	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)

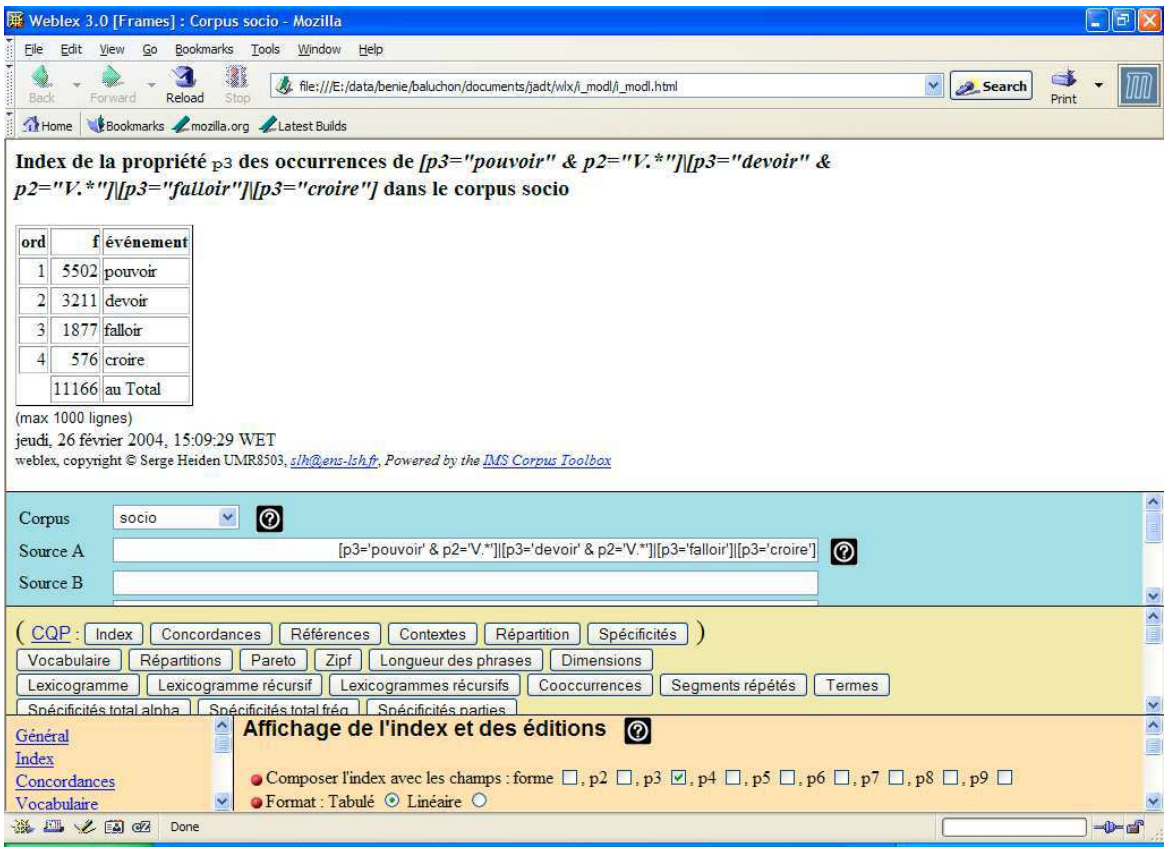
- 2e cas : Résultats = liste de types
  - ex. : vocabulaire, index, spécificités, cooccurrents
  - affichage de la dimension d'analyse
  - sur des colonnes complémentaires, valeurs correspondantes pour les dimensions d'affichage demandées

Dimensions d'affichage (3)

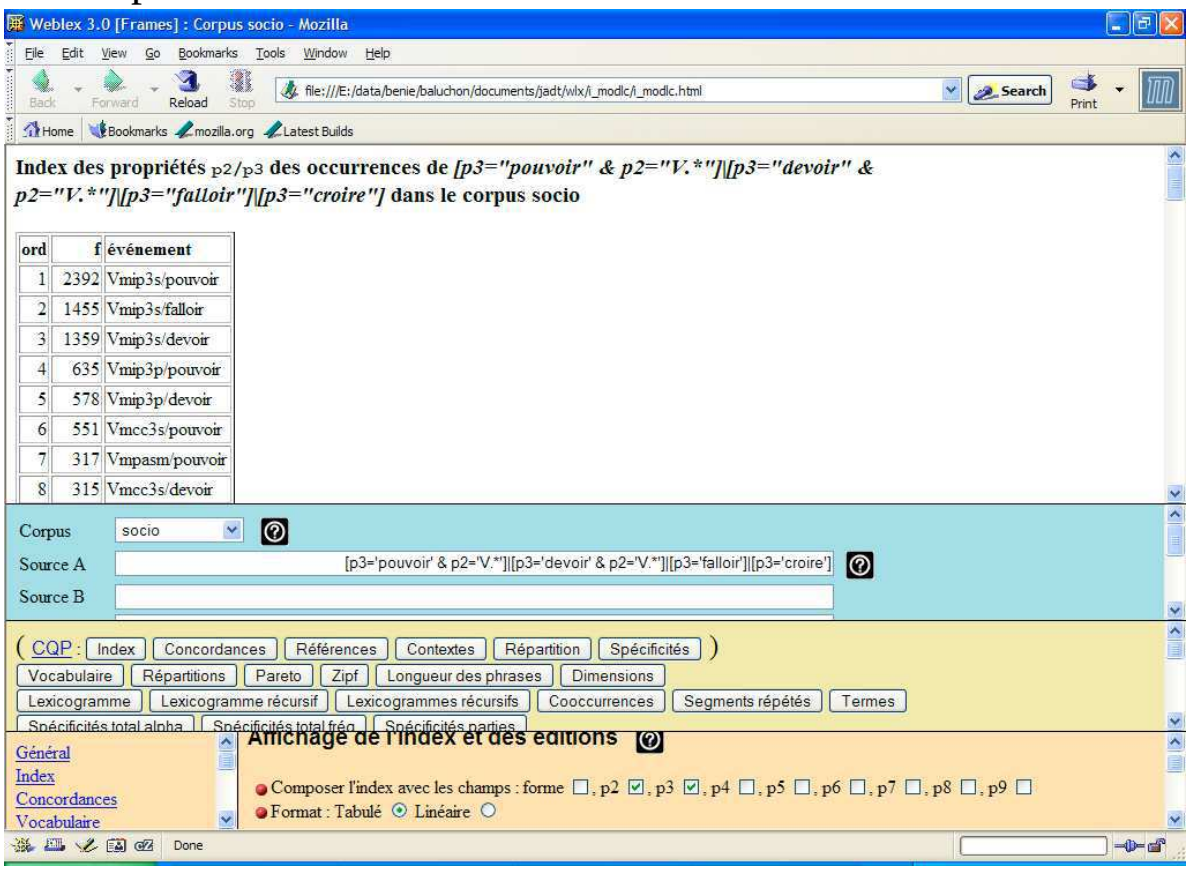
Exemple :
plutôt que liste de lemmes seuls
ou liste détaillée de toutes les flexions,
liste de lemmes avec quelques indications de flexion :

fréq.	*lemme*	*mode*	*temps*	*personne*
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

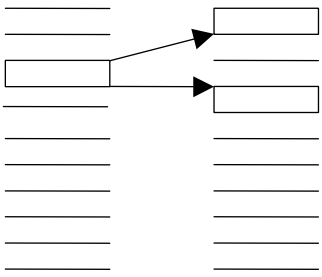


Exemple : liste des lemmes + flexions

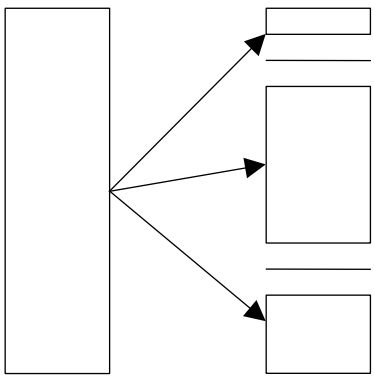


Affichage : projections

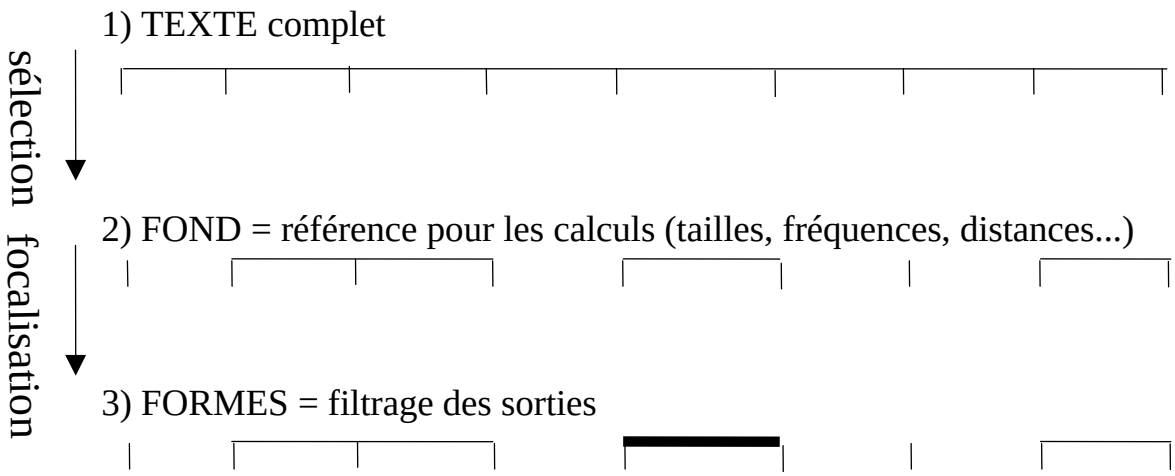
d'un élément :



d'une liste entière :

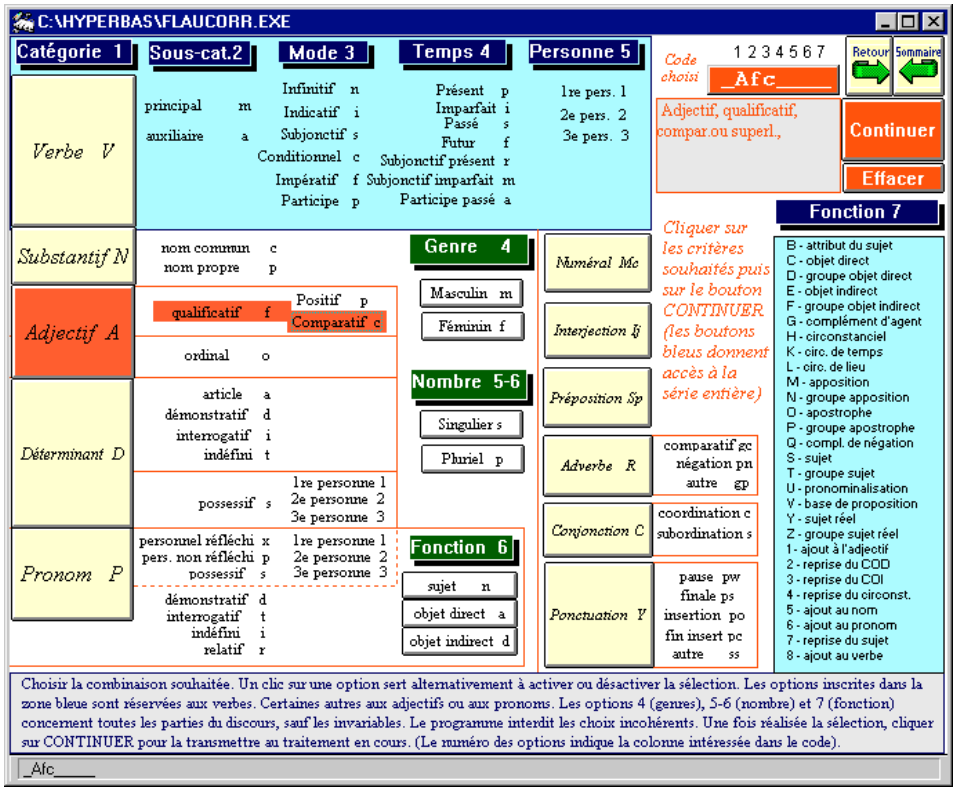


Et l'axe syntagmatique du tableau ?



Le deuxième paramètre *qui* **intervient** *dans les* calculs

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



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 :