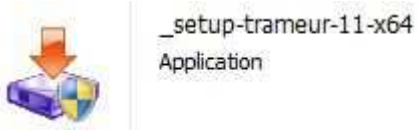
Le Trameur

Trameur: A Framework for Annotated Text Corpora Exploration

Serge Fleury (Sorbonne Nouvelle – Paris 3)
serge.fleury@univ-paris3.fr
Maria Zimina (Paris Diderot – Sorbonne Paris Cité)
maria.zimina@eila.univ-paris-diderot.fr

Le Trameur *in Brief*

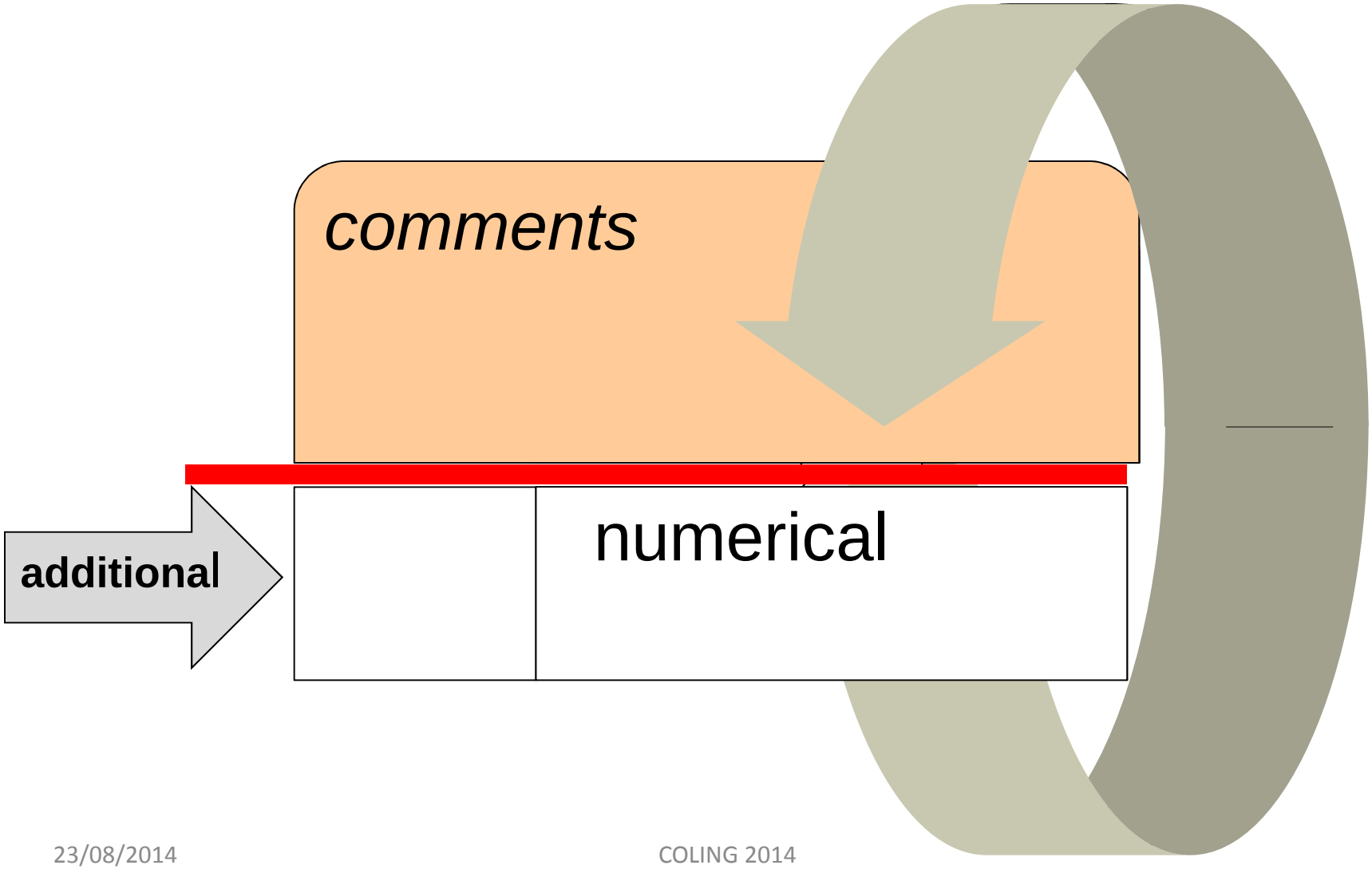
- Le Trameur **(Fleury, 2013a)** is a tool for exploration of *Treebanks*, or parsed text corpora that annotate syntactic or semantic sentence structure.
- The software is distributed with a *graphical user interface*.
- A reference package for *Windows* is available for free download from the official website: **http://www.tal.univ-paris3.fr/trameur**



The novelty of Le Trameur

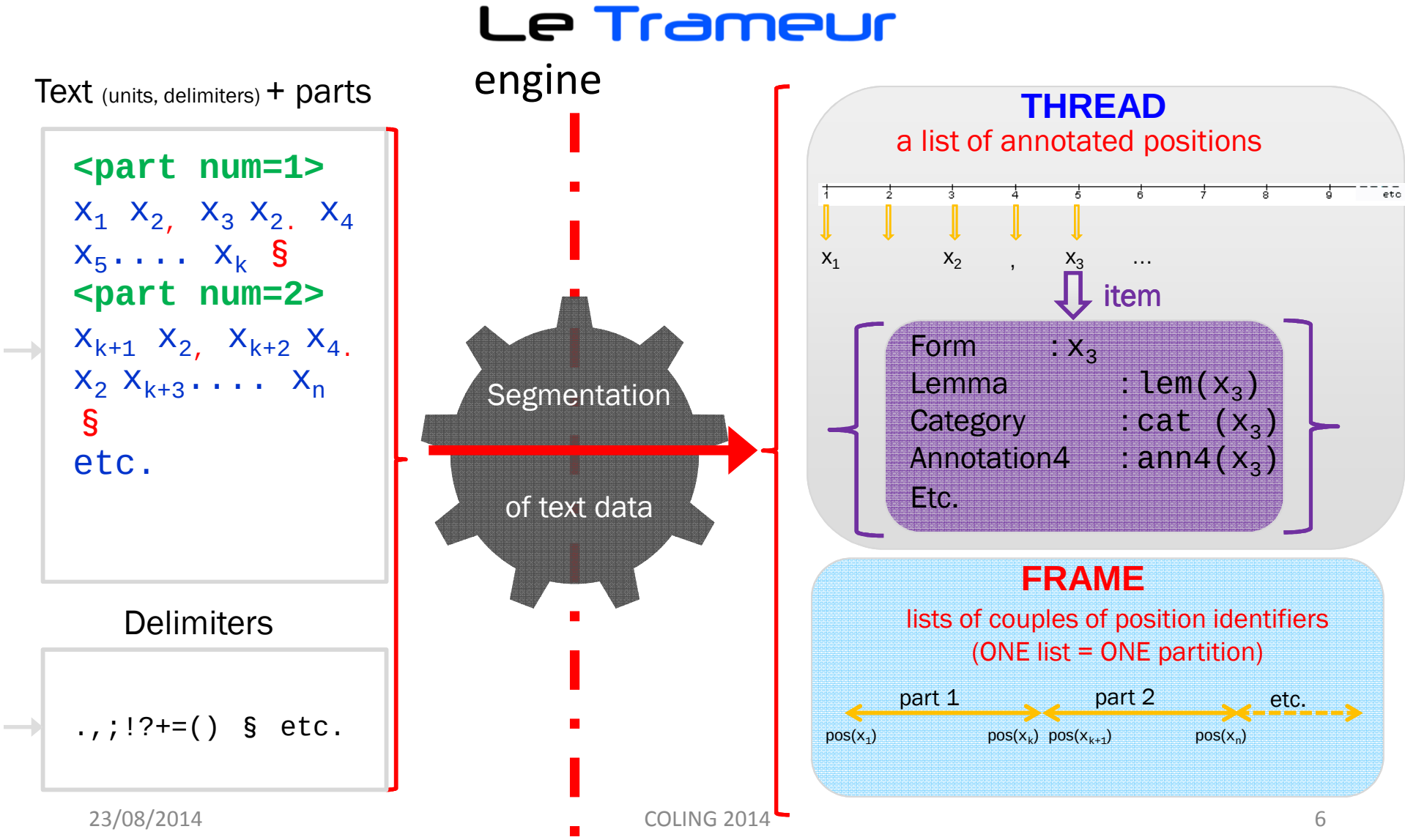
- Search engine, context return, text mapping, graphs and statistical analysis of dependency relations within a single graphical user interface.
- Simultaneous access to multiple corpus layers and their interactions statistics.
- Possibility to implement ***incremental textual resources for Treebanks.***

Incremental Textual Resources for *Treebanks*



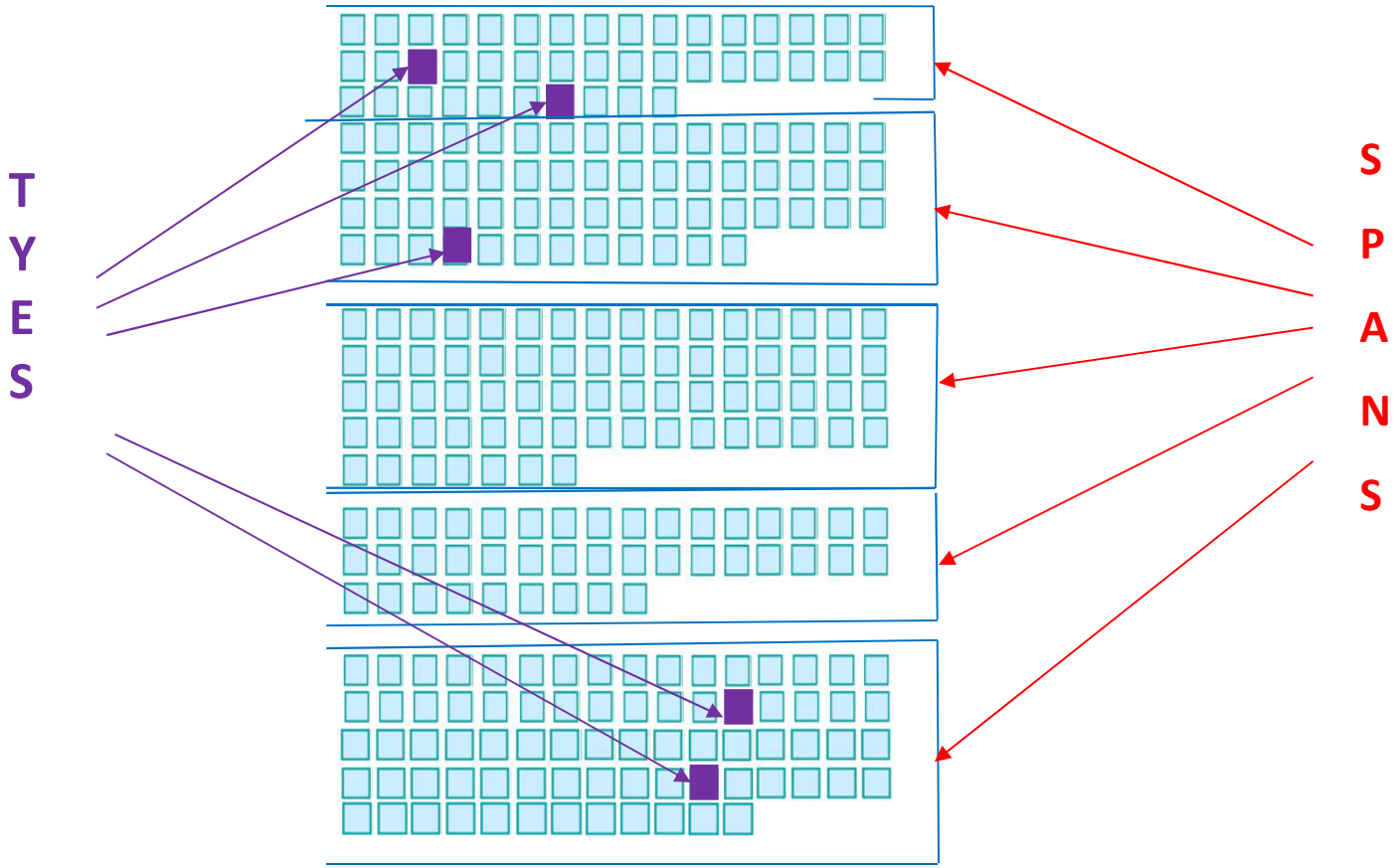
FRAMEWORK

System architecture



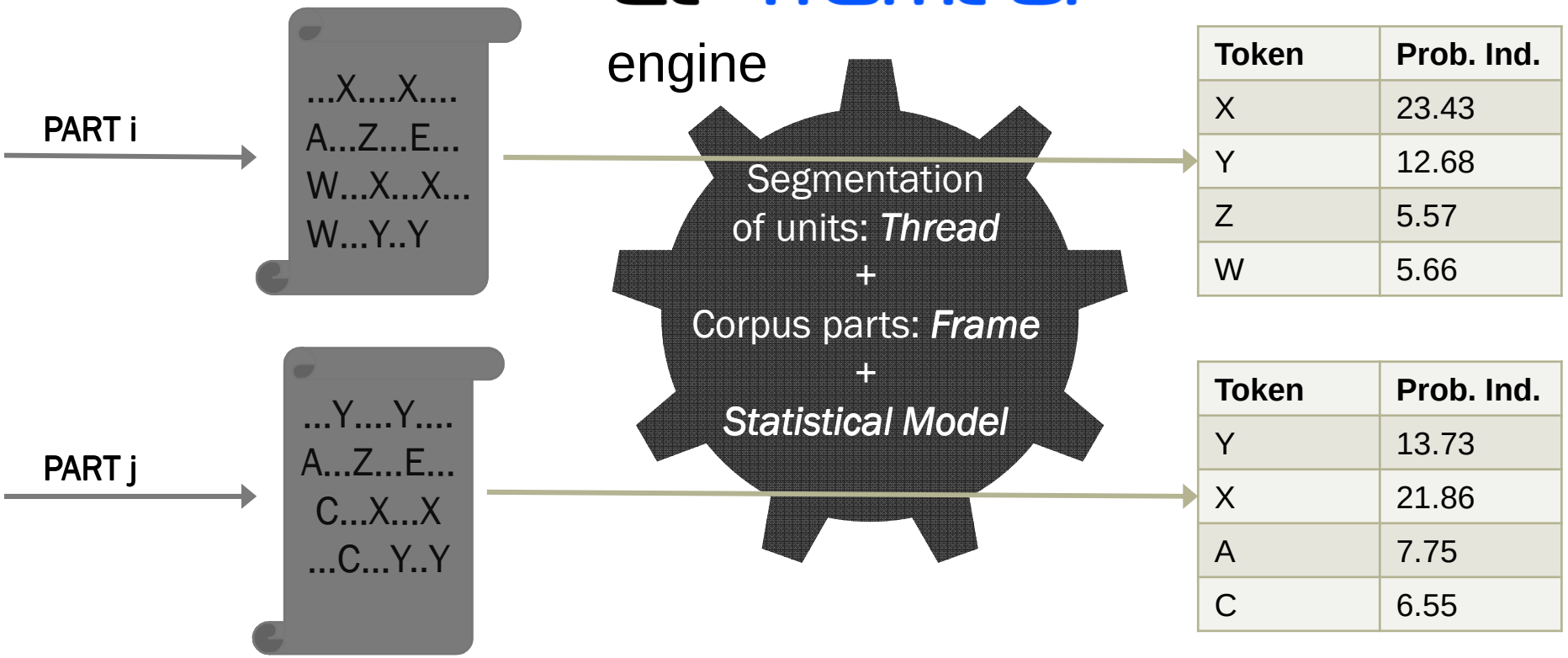
Statistical tables, *Types* and *Spans*

Systems of ***item types*** and ***text spans*** allow automatic counts in statistical tables.



How it works…

Statistical tables are processed using quantitative methods



Token	Prob. Ind.
X	23.43
Y	12.68
Z	5.57
W	5.66

Token	Prob. Ind.
Y	13.73
X	21.86
A	7.75
C	6.55

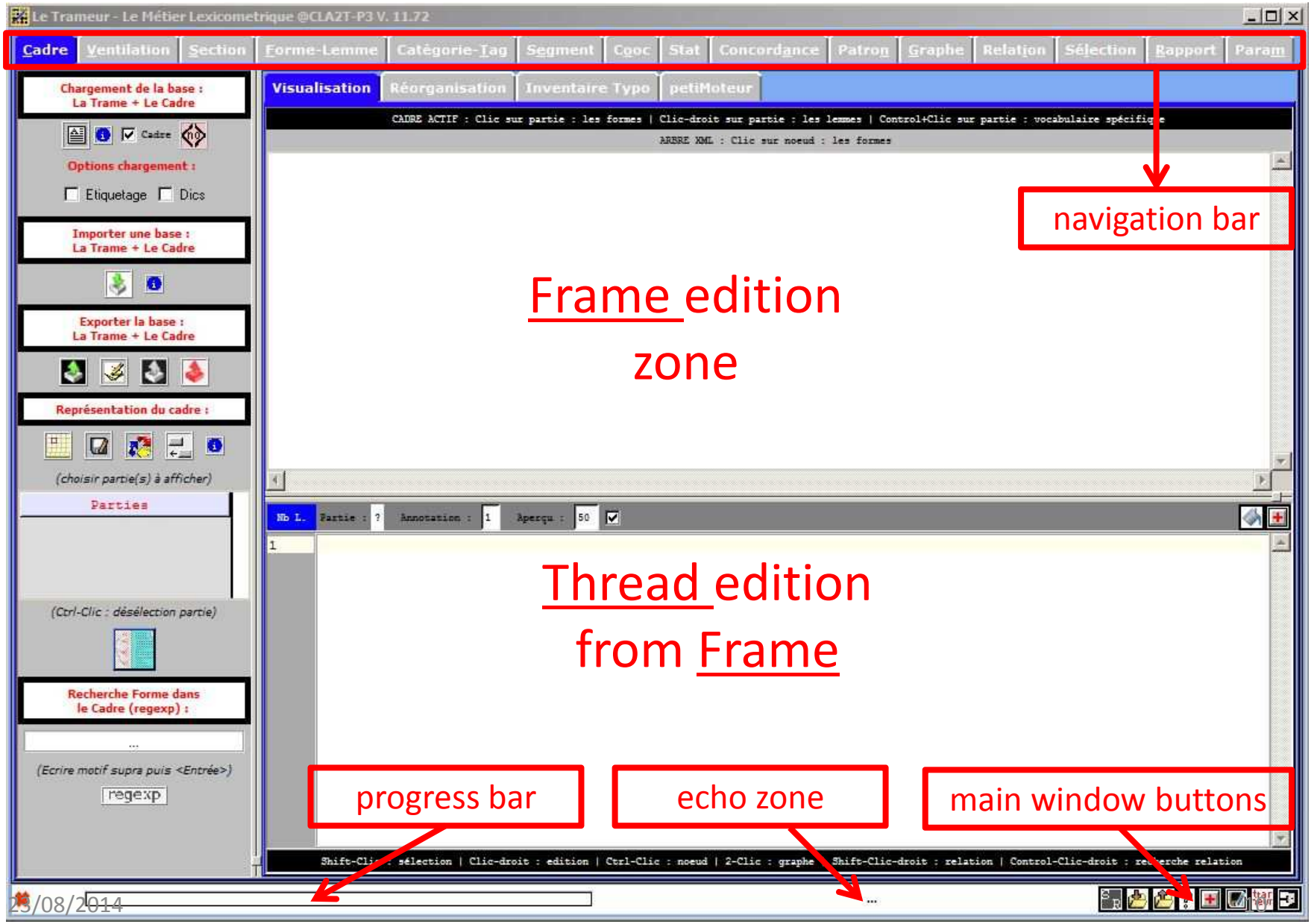
http://www.tal.univ-paris3.fr/trameur

Le Trameur

GRAPHICAL USER INTERFACE

User interface

Le Trameur



Base file (Thread + Frame)

- **Creating a new base file:**
  - Text only
  - Tagged texts
  - XML encoding (TEI etc.)

- **Importing an existing base file:**
  - Predefined XML format: encoding *Thread* and *Frame* annotations
    - Example of a multi-annotated base file : *Rhapsodie2Trameur*

    http://www.tal.univ-paris3.fr/trameur/bases/baseTrameurFromRhapsodie.zip

Example: *Rhapsodie*

a Prosodic-Syntactic Treebank for Spoken French

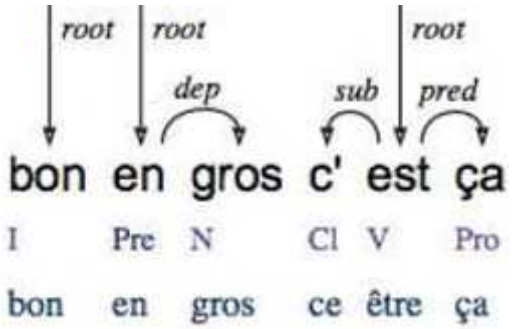
http://www.projet-rhapsodie.fr

57 short samples of spoken French (≈5 min each)

≈ 33 000 words, tabular layout

- Microsyntactic annotations
- Macrosyntactic analysis
- Prosody

arborator.ilpga.fr

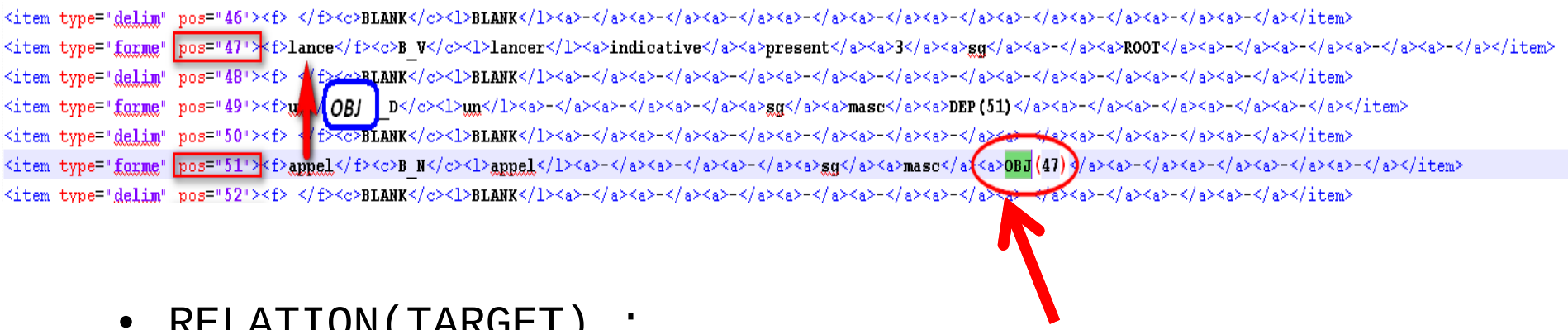


(Gerdes *et al.*, 2012)

" bon " en gros < c' est ça

Rhapsodie data within a base file

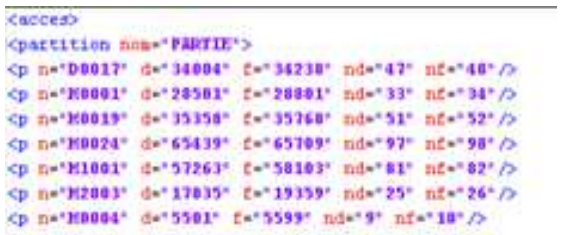
Le Trameur



- RELATION(TARGET) :
  - Position 51: "appel" OBJ(47) → "lance" (position 47)

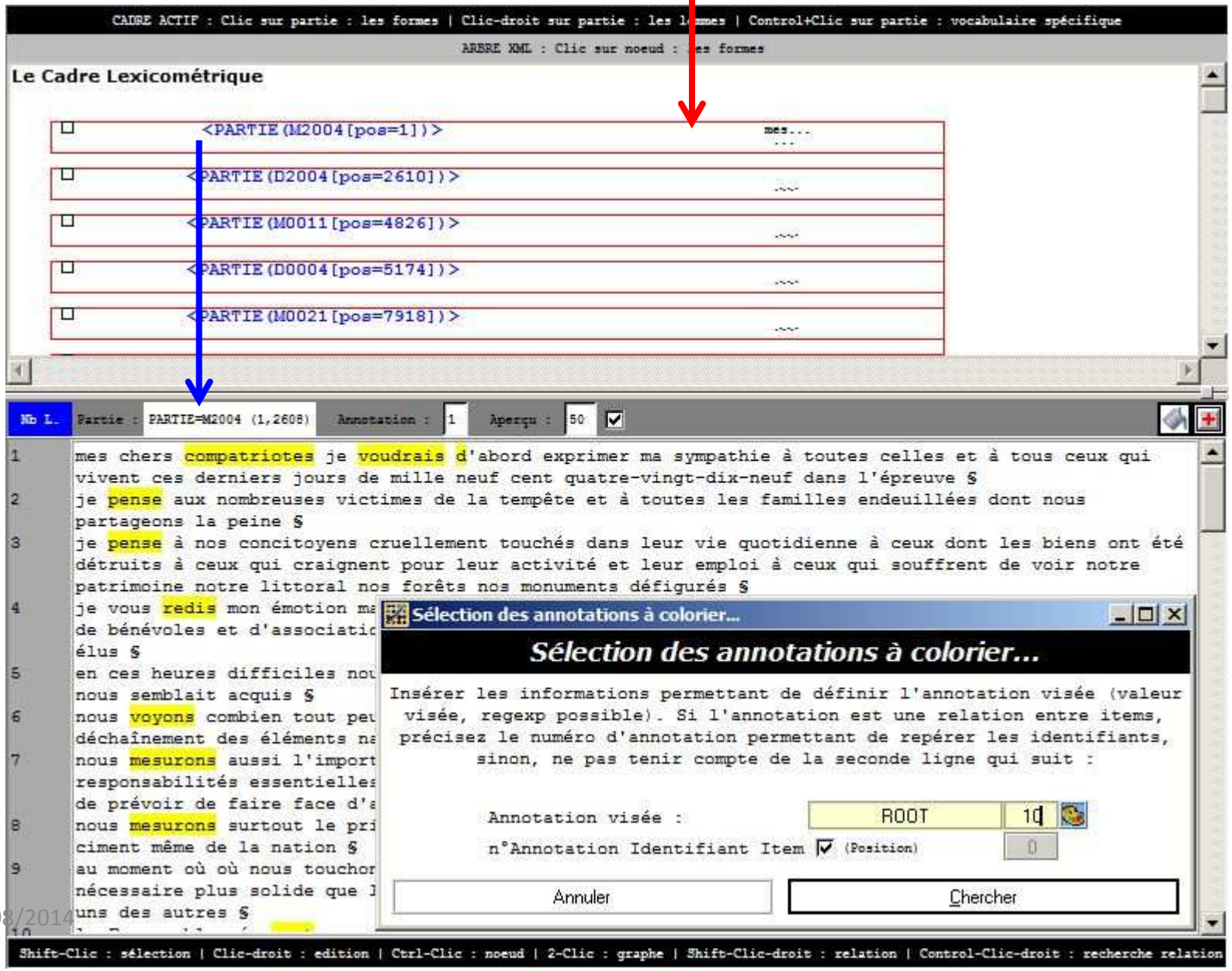
- RELATION is a character string showing the name of the target relation
- TARGET is a number indicating a specific position on the Thread

Corpus parts are identified on the **Thread** using **position identifiers**:



Thread and Frame display

Le Trameur



Le Trameur

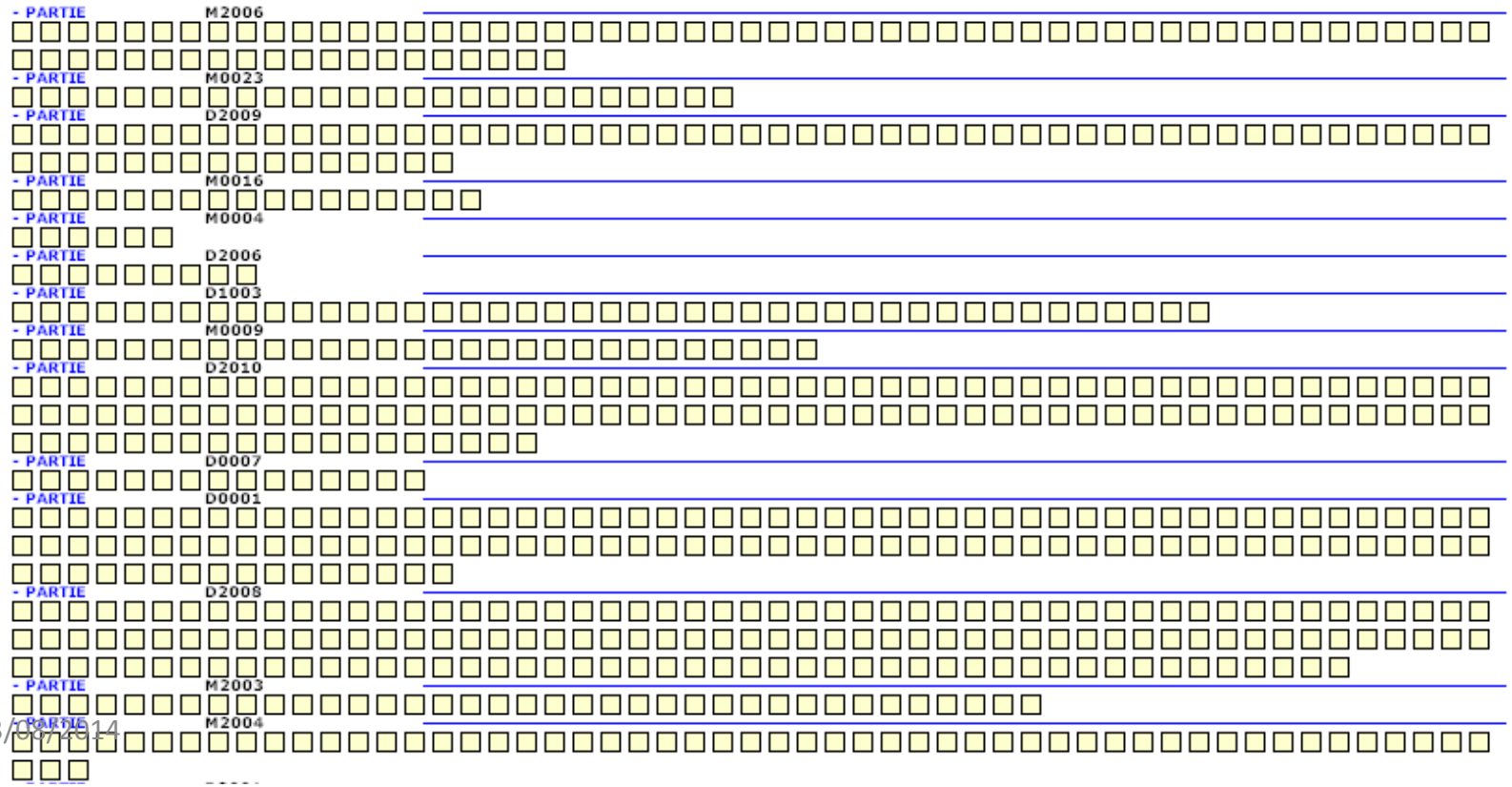
SECTIONS MAP

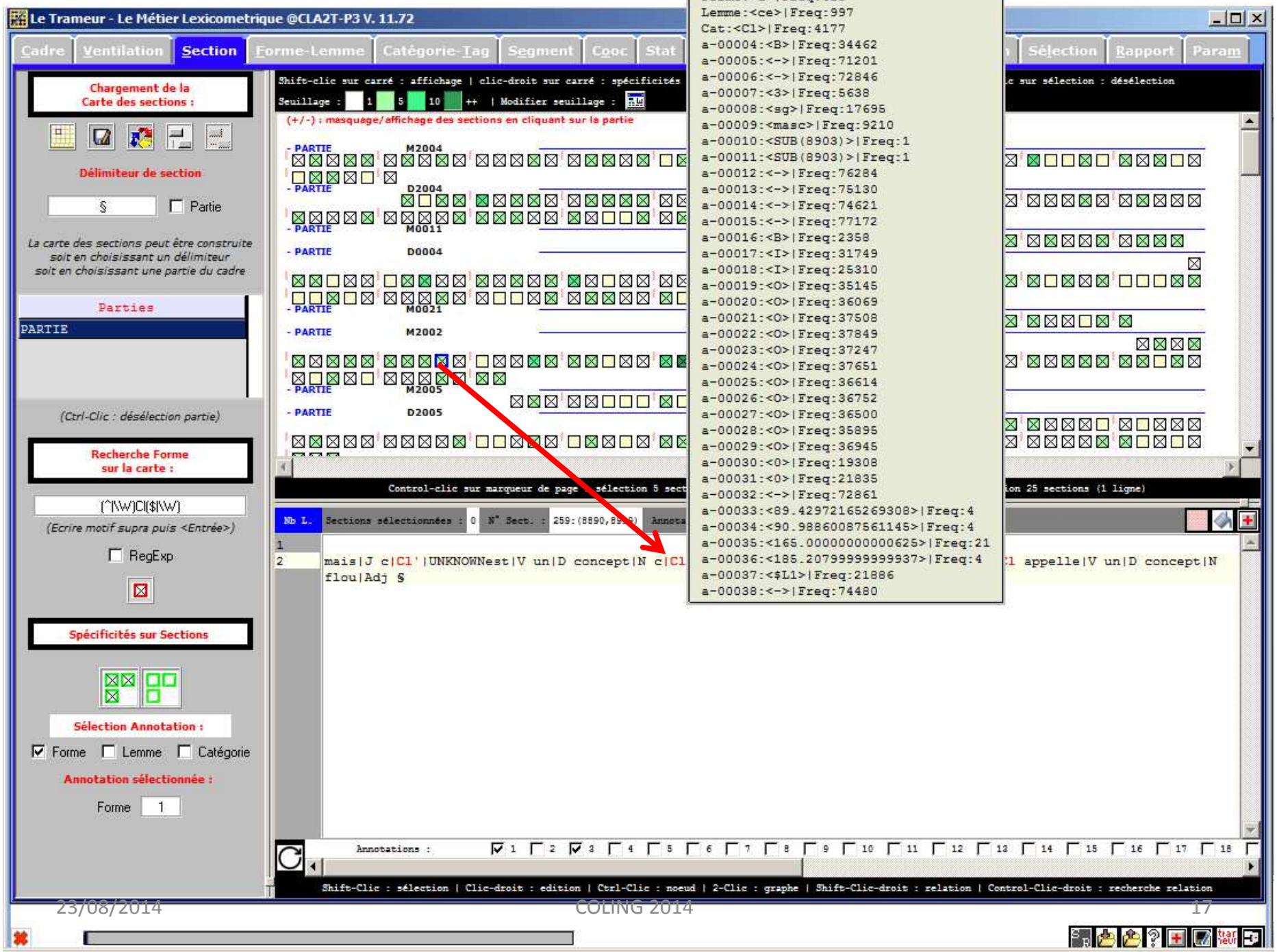
Sections map

Le Trameur

After each *Illocutionary Unit* (IU), a delimiting character (§) is used to build a ***map of sections***.

Sections map from **Rhapsodie** (extract):




Le Trameur - Le Métier Lexicometrique @CLA2T-P3 V. 11.72
Cadre
Ventilation
Section
Forme-Lemme
Catégorie-Tag
Segment
Cooc
Stat
Sélection
Rapport
Param
Chargement de la Carte des sections :
Délimiteur de section
§
Partie
La carte des sections peut être construite soit en choisissant un délimiteur soit en choisissant une partie du cadre
Parties
PARTIE
(Ctrl-Clic : désélection partie)
Recherche Forme sur la carte :
(^|\W)Cl($|\W)
(Ecrire motif supra puis <Entrée>)
RegExp
Spécificités sur Sections
Sélection Annotation :
Forme
Lemme
Catégorie
Annotation sélectionnée :
Forme
1
Shift-clic sur carré : affichage | clic-droit sur carré : spécificités
Seuillage :
Modifier seuillage :
(+/-) : masquage/affichage des sections en cliquant sur la partie
- PARTIE M2004
- PARTIE D2004
- PARTIE M0011
- PARTIE D0004
- PARTIE M0021
- PARTIE M2002
- PARTIE M2005
- PARTIE D2005
Control-clic sur marqueur de page
Nb L.
Sections sélectionnées : 0
N° Sect. : 259:(8890,8902)
mais|J c|Cl'|UNKNOWNest|V un|D concept|N c|Cl
flou|Adj $
Annotations :
Shift-Clic : sélection | Clic-droit : edition | Ctrl-Clic : noeud | 2-Clic : graphe | Shift-Clic-droit : relation | Control-Clic-droit : recherche relation
Position:<8901>
Forme:<c>|Freq:618
Lemme:<ce>|Freq:997
Cat:<Cl>|Freq:4177
a-00004:<B>|Freq:34462
a-00005:<->|Freq:71201
a-00006:<->|Freq:72846
a-00007:<3>|Freq:5638
a-00008:<sg>|Freq:17695
a-00009:<masc>|Freq:9210
a-00010:<SUB(8903)>|Freq:1
a-00011:<SUB(8903)>|Freq:1
a-00012:<->|Freq:76284
a-00013:<->|Freq:75130
a-00014:<->|Freq:74621
a-00015:<->|Freq:77172
a-00016:<B>|Freq:2358
a-00017:<I>|Freq:31749
a-00018:<I>|Freq:25310
a-00019:<O>|Freq:35145
a-00020:<O>|Freq:36069
a-00021:<O>|Freq:37508
a-00022:<O>|Freq:37849
a-00023:<O>|Freq:37247
a-00024:<O>|Freq:37651
a-00025:<O>|Freq:36614
a-00026:<O>|Freq:36752
a-00027:<O>|Freq:36500
a-00028:<O>|Freq:35895
a-00029:<O>|Freq:36945
a-00030:<0>|Freq:19308
a-00031:<0>|Freq:21835
a-00032:<->|Freq:72861
a-00033:<89.42972165269308>|Freq:4
a-00034:<90.98860087561145>|Freq:4
a-00035:<165.00000000000625>|Freq:21
a-00036:<185.20799999999937>|Freq:4
a-00037:<$L1>|Freq:21886
a-00038:<->|Freq:74480

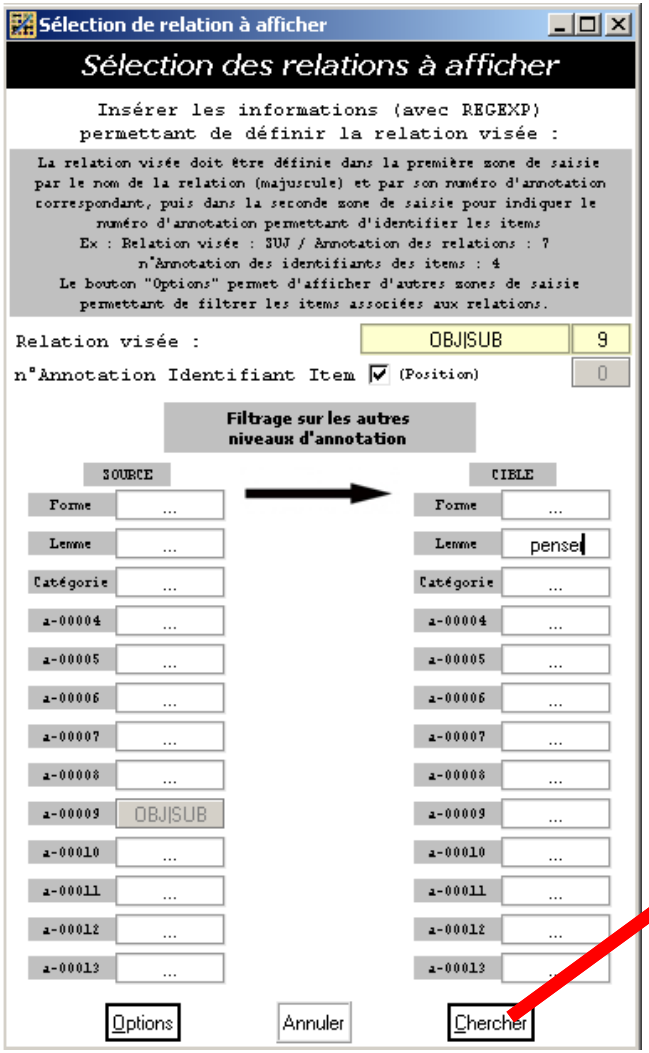
Demo 1: Sections map / statistics

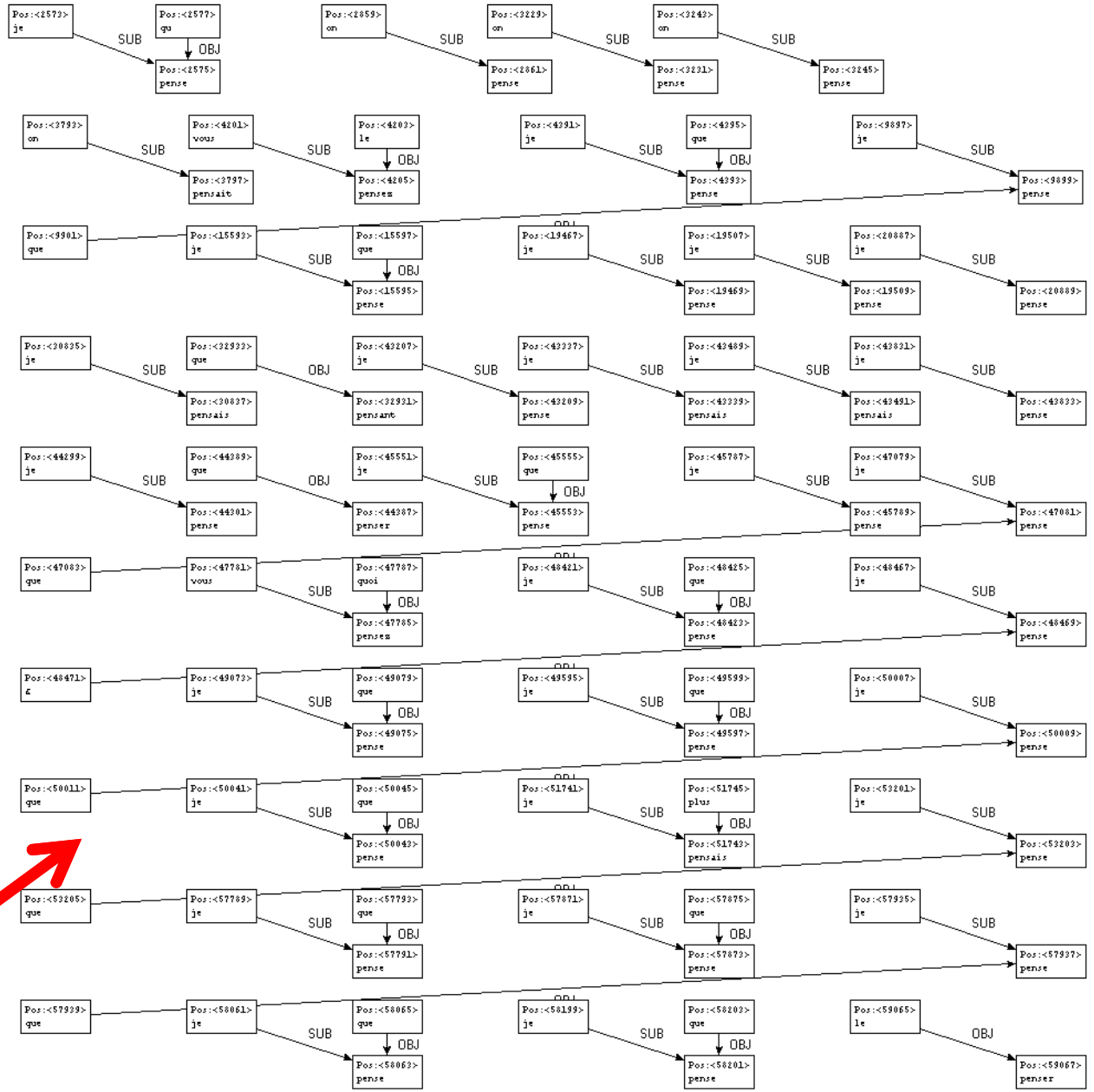
What POS patterns are underrepresented in Illocutionary Units with a clitic?

POS N-gram	Ind-Spécif	Fq-Totale	Fq-Partie
Adj V	-10.1	49	25
I I I I	-10.1	16	3
I I I	-19.2	57	21
Pro V	-22.0	133	71
D N V	-24.0	216	130
N	-28.6	6308	5232
N V	-39.4	384	234
I I	-48.0	273	140
I	-91.3	1978	1397

DEPENDENCY GRAPHS

SUB -> penser <-OBJ





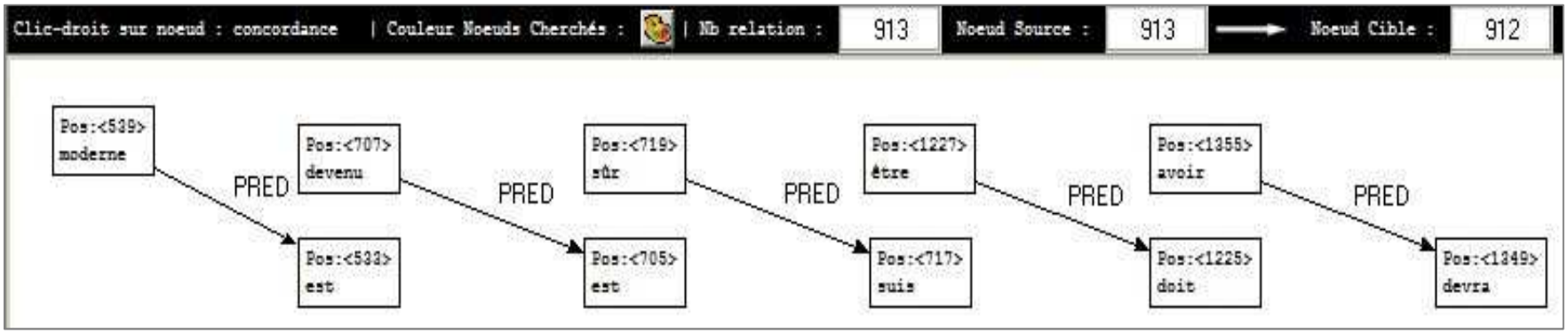
Context return

----------PARTIE{PARTIE=M0023 }----------
$ et il se dit que & $ ouais moi **je** *pense* qu'en fait il voit d'abord la
il se dit que & $ ouais moi je *pense* **qu**'en fait il voit d'abord la la jeune
----------PARTIE{PARTIE=D2009 }----------
il y a des moments où on écrit parce qu'**on** *pense* participer à un combat $ ça a été
$ mais d'abord nous savons bien qu'en fait **on** *pense* toujours avec du langage que on *pense* en
qu'en fait on *pense* toujours avec du langage que **on** *pense* en parlant qu'on parle en pensant qu'
littérature sont l'histoire des écrivains des écoles $ et **on** ne *pensait* jamais au lecteur $ la la lecture
sa perte $ c'était un mot $ ou bien **vous** le *pensez* vraiment $ ou peut-être y a-
perte $ c'était un mot $ ou bien vous **le** *pensez* vraiment $ ou peut-être y a-t-
preuve c'est que actuellement voyez-vous en soixante-quinze **je** *pense* que vous serez d'accord avec moi il
est que actuellement voyez-vous en soixante-quinze je *pense* **que** vous serez d'accord avec moi il n'y
----------PARTIE{PARTIE=D2010 }----------
était & qu'elles sont dans les musées hein $ **je** *pense* que si on s'était assis l'un
qu'elles sont dans les musées hein $ je *pense* **que** si on s'était assis l'un devant l'
----------PARTIE{PARTIE=D2008 }----------
anti-viraux $ il y a pas de vaccins $ **je** *pense* que aujourd'hui euh s~ euh même une
$ il y a pas de vaccins $ je *pense* **que** aujourd'hui euh s~ euh même une grippe aussi
----------PARTIE{PARTIE=M2004 }----------
neuf cent quatre-vingt-dix-neuf dans l'épreuve $ **je** *pense* aux nombreuses victimes de la tempête et à
toutes les familles endeuillées dont nous partageons la peine $ **je** *pense* à nos concitoyens cruellement touchés dans leur vie
l'Homme et ne se retournent jamais contre lui $ **je** *pense* par exemple aux manipulations génétiques au clonage $
----------PARTIE{PARTIE=D2007 }----------
étiez un peu choquée quand même hein non mais $ **je** *pensais* à son procès en fait $ tu peux
au téléphone vous avez déjà chauffé la mère en *pensant* **que** vous vous adressiez à la fille $ euh non
----------PARTIE{PARTIE=D0004 }----------
y a maintenant pff peut-être au moins douze ans **je** *pense* hein douze treize ans $ donc oui euh
un petit peu au-dessus $ mais & $ non **je** *pensais* à ça parce que j'imagine que des
n'a jamais vraiment travaillé euh mh mh non non **je** *pensais* aux albums oui avec $ du du point
des enfants sont à peu près autour des trente ans **je** *pense* hein $ mh mh donc euh ils sont
le soir $ ça c'est le problème de Paris **je** *pense* mh mh et de la région parisienne aussi
travailler $ donc euh & $ on aurait pu *penser* **que** les gens du quatorzième étaient privilégiés restaient dans la
----------PARTIE{PARTIE=D0006 }----------
c'était que l'argent $ mh $ et euh **je** *pense* que ça a été très bénéfique $ vous
que l'argent $ mh $ et euh je *pense* **que** ça a été très bénéfique $ vous vous faites
bon quand-même $ non $ c'est les parents **je** *pense* $ et euh XXX et j~ & $
la psychiatrie c'est c'est quelque chose $ euh **je** *pense* que la chirurgie c'est pareil hein $
c'est c'est quelque chose $ euh je *pense* **que** la chirurgie c'est pareil hein $ n'importe
$ les animaux dans les dans les quartiers de Paris **vous** en *pensez* quoi ici $ euh alors un changement
dans les dans les quartiers de Paris vous en *pensez* **quoi** ici $ euh alors un changement mh euh les
----------PARTIE{PARTIE=D0002 }----------
mh mh mh mh euh les lycées du & $ **je** *pense* que à l'intérieur de Paris euh globalement
mh mh euh les lycées du & $ je *pense* **que** à l'intérieur de Paris euh globalement euh les
écoles sont sont d'un bon niveau voilà $ mh **je** *pense* & $ moi je vois euh dans le
sont d'un bon niveau voilà $ mh je *pense* **&** $ moi je vois euh dans le vingtième j'
mesures du plan banlieue le busing $ mh mh mais **je** *pense* surtout que euh dans le sixième arrondissement les
banlieue le busing $ mh mh mais je *pense* surtout **que** euh dans le sixième arrondissement les maternelles euh il
$ non $ mh mh $ donc euh non je **je** *pense* que le problème & $ je dis pas
$ mh mh $ donc euh non je je *pense* **que** le problème & $ je dis pas que tous
sûr bien sûr que je parlais oui XXX oui mais **je** *pense* que vous avez raison de mélange XXX euh
sûr que je parlais oui XXX oui mais je *pense* **que** vous avez raison de mélange XXX euh $ je
mélange XXX euh $ je crois que là & $ **je** *pense* que vous avez raison $ mais mh si
euh $ je crois que là & $ je *pense* **que** vous avez raison $ mais mh si vous voulez
----------PARTIE{PARTIE=D0009 }----------
de quarante euros quelque chose comme ça ah oui ah **je** *pensais* plus que ça oui quarante quarante-cinq $
euros quelque chose comme ça ah oui ah je *pensais* **plus** que ça oui quarante quarante-cinq $ et est-
que c'est des heures ça hein $ oui mais **je** *pense* que non ça va ça va pas très
est des heures ça hein $ oui mais je *pense* **que** non ça va ça va pas très très loin

Searches in dependency graphs

Demo 2: Selected relations

Complex predicates which are <u>not</u> <u>governed</u> by another element (ROOTS)

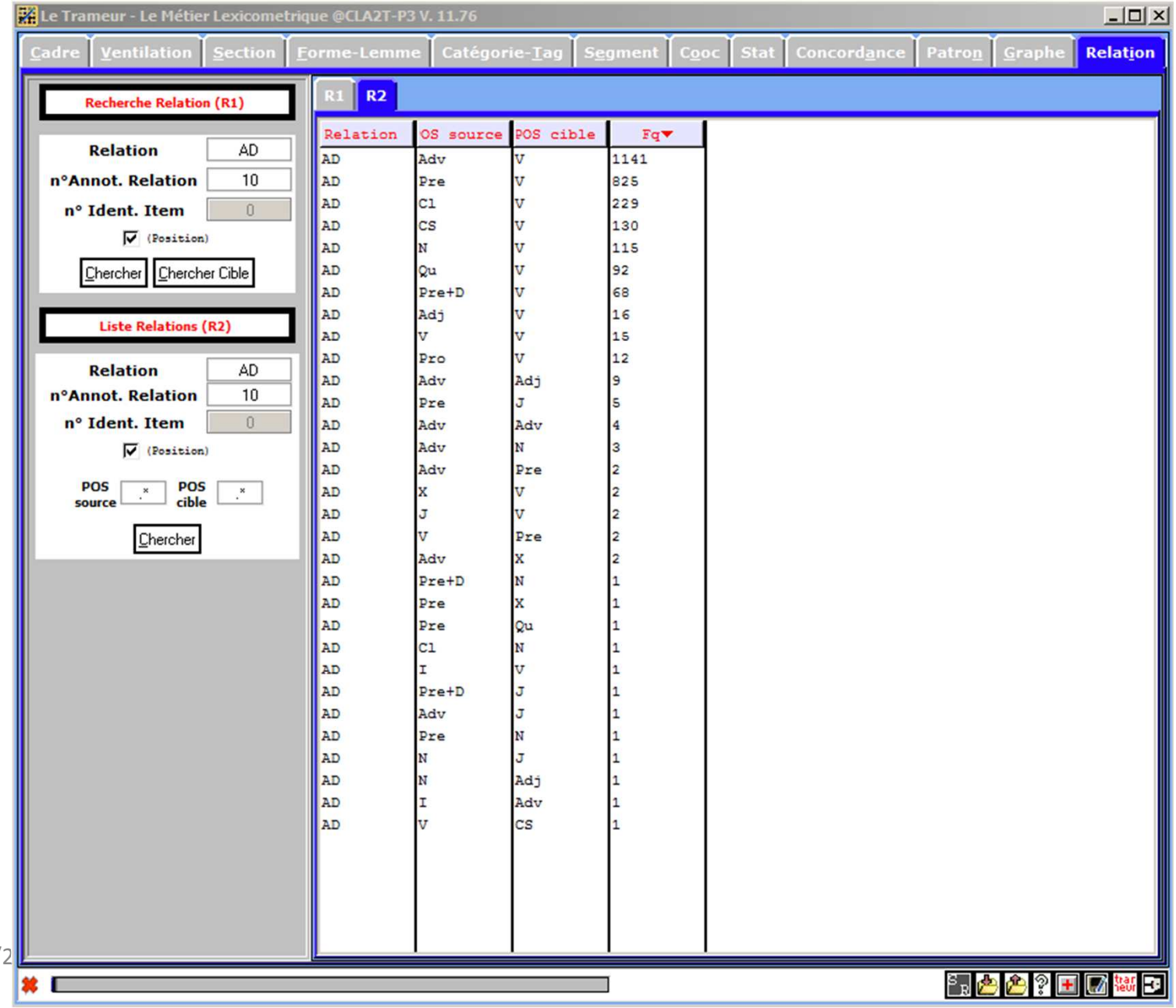


----------PARTIE{PARTIE=M2004}----------

rien n'est décidément plus moderne plus nécessaire plus solide
l'an deux mille est devenu contemporain immédiat § je
contemporain immédiat § je suis sûr que beaucoup d'entre
vingt-et-unième siècle doit être le siècle de l'
il ne devra plus y avoir de repos pour les
nation une exigence se fait entendre toujours plus forte pour
dégradent le patrimoine naturel doit être recherchée et sanctionnée §
modernité ne doit pas nous diviser § elle doit profiter
nous diviser § elle doit profiter à chacun § nous
notre avenir § nous avons choisi ensemble de faire grandir
le monde § nous avons choisi aussi de prendre part
§ ce sera tout le sens du combat de la

RELATIONS

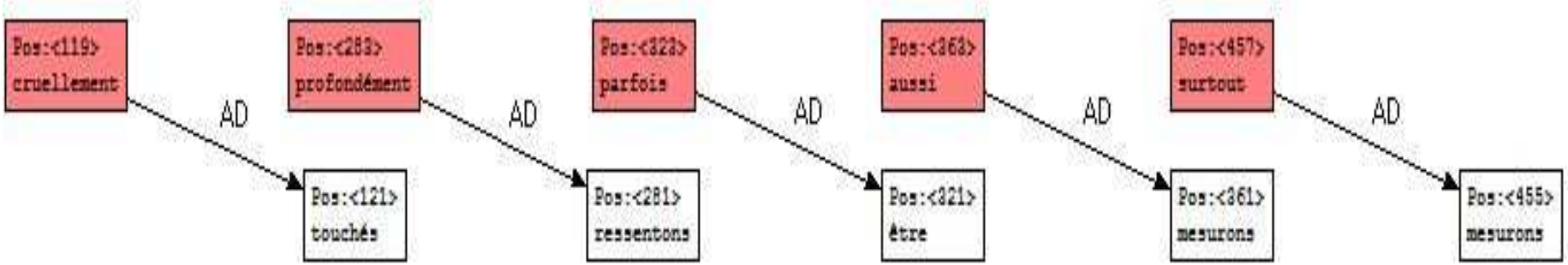
POS -> RELATION -> *POS* (statistics)



Relation	OS source	POS cible	Fq
AD	Adv	V	1141
AD	Pre	V	825
AD	Cl	V	229
AD	CS	V	130
AD	N	V	115
AD	Qu	V	92
AD	Pre+D	V	68
AD	Adj	V	16
AD	V	V	15
AD	Pro	V	12
AD	Adv	Adj	9
AD	Pre	J	5
AD	Adv	Adv	4
AD	Adv	N	3
AD	Adv	Pre	2
AD	X	V	2
AD	J	V	2
AD	V	Pre	2
AD	Adv	X	2
AD	Pre+D	N	1
AD	Pre	X	1
AD	Pre	Qu	1
AD	Cl	N	1
AD	I	V	1
AD	Pre+D	J	1
AD	Adv	J	1
AD	Pre	N	1
AD	N	J	1
AD	N	Adj	1
AD	I	Adv	1
AD	V	CS	1

Demo 3: V -> AD -> Adv

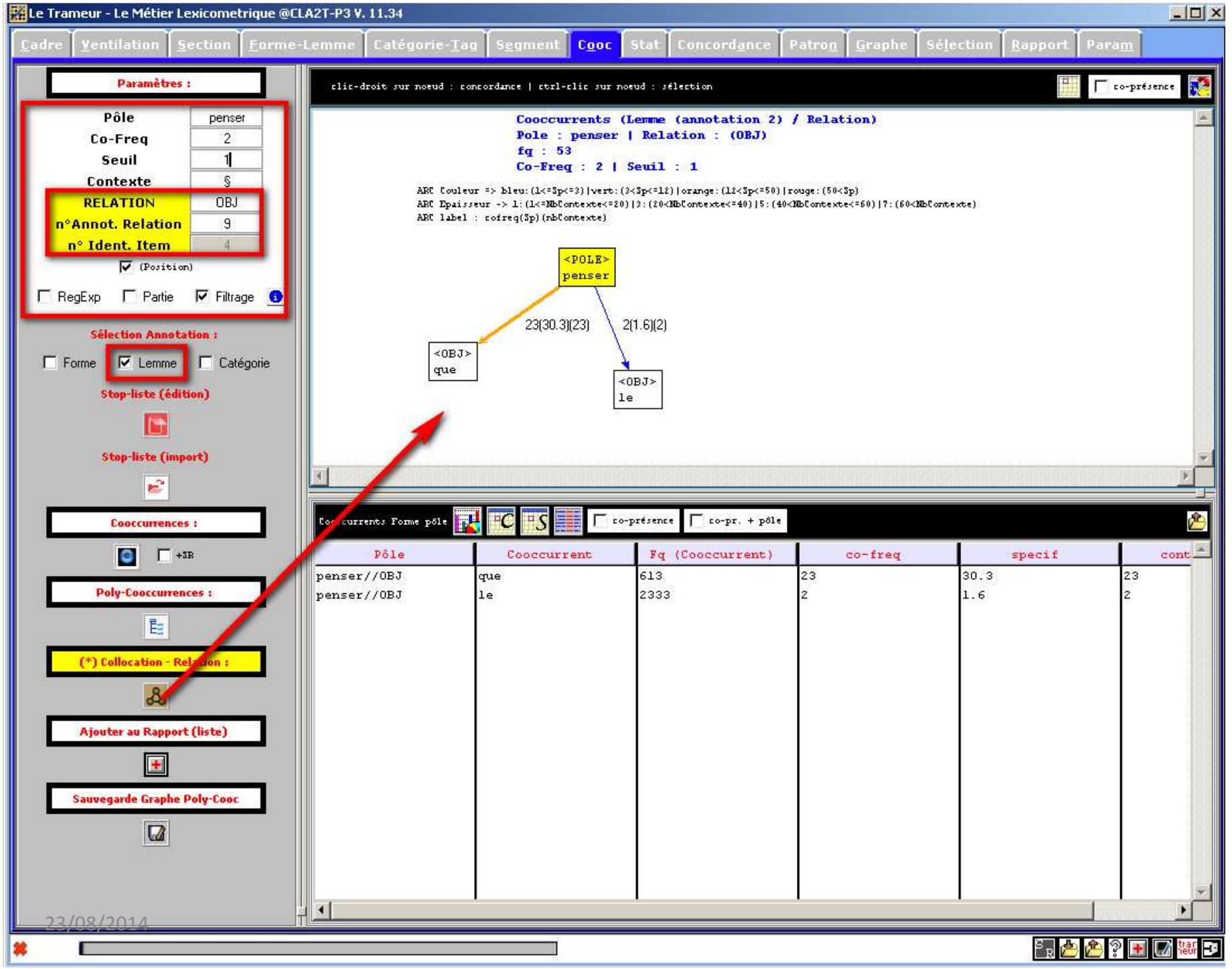
Verbs with <u>added adverbs</u> (Freq = 1141)



----------PARTIE{PARTIE=M2004}----------

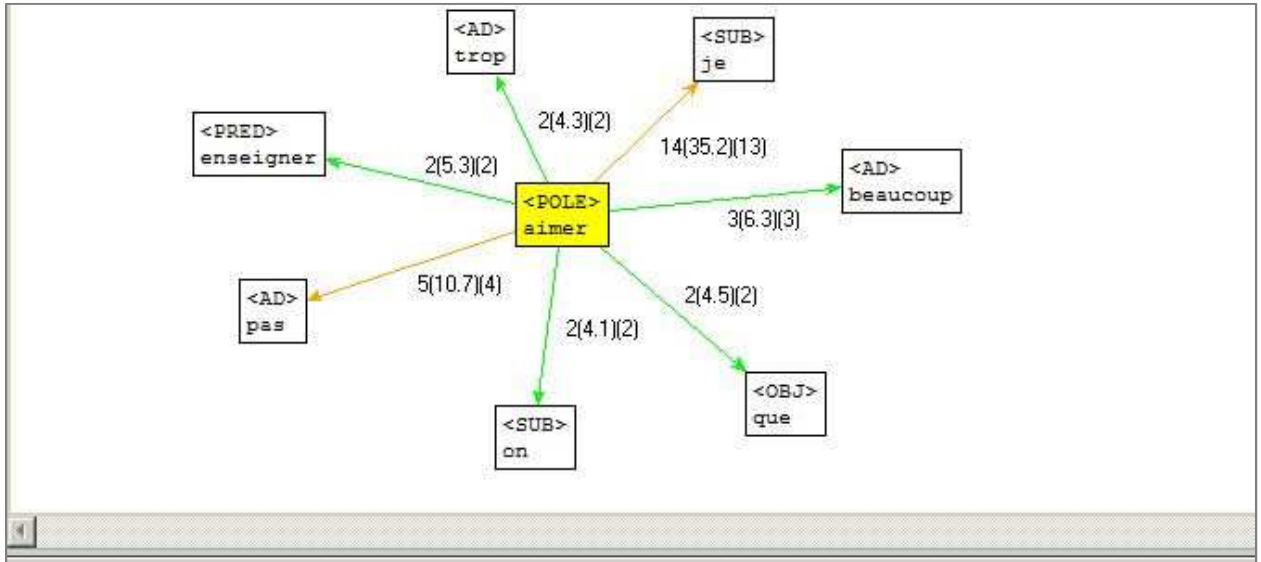
je pense à nos concitoyens cruellement touchés dans leur vie
ces heures difficiles nous ressentons profondément la fragilité des choses
voyons combien tout peut être parfois remis en cause du
éléments naturels § nous mesurons aussi l'importance du rôle
du pays § nous mesurons surtout le prix de l'
deux mille rien n'est décidément plus moderne plus nécessaire
parce que nos compatriotes ont toujours su dans l'épreuve
§ ce soir nous vivons ensemble un moment fort et
très lointain et qui a longtemps symbolisé le futur l'
d'étonnement une certaine appréhension parfois née du sentiment que

COLLOCATIONS CONSTRAINED BY DEPENDENCY RELATIONS


Le Trameur - Le Métier Lexicometrique @CLA2T-P3 V. 11.34
Cadre
Ventilation
Section
Forme-Lemme
Catégorie-Tag
Segment
Cooc
Stat
Concordance
Patron
Graphe
Sélection
Rapport
Param
Paramètres :
Pôle
penser
Co-Freq
2
Seuil
1
Contexte
§
RELATION
OBJ
n°Annot. Relation
9
n° Ident. Item
4
(Position)
RegExp
Partie
Filtrage
Sélection Annotation :
Forme
Lemme
Catégorie
Stop-liste (édition)
Stop-liste (import)
Cooccurrences :
+3R
Poly-Cooccurrences :
(*) Collocation - Relation :
Ajouter au Rapport (liste)
Sauvegarde Graphe Poly-Cooc
clic-droit sur noeud : concordance | ctrl-clic sur noeud : sélection
co-présence
Cooccurrents (Lemme (annotation 2) / Relation)
Pole : penser | Relation : (OBJ)
fq : 53
Co-Freq : 2 | Seuil : 1
ARC Couleur => bleu:(1<=Sp<=3)|vert:(3<Sp<=12)|orange:(12<Sp<=50)|rouge:(50<Sp)
ARC Epaisseur -> 1:(1<=NbContexte<=20)|3:(20<NbContexte<=40)|5:(40<NbContexte<=60)|7:(60<NbContexte)
ARC label : cofreq(Sp)(nbContexte)
<POLE>
penser
23(30.3)(23)
2(1.6)(2)
<OBJ>
que
<OBJ>
le
Cooccurrents Forme pôle
co-présence
co-pr. + pôle
Pôle
Cooccurrent
Fq (Cooccurrent)
co-freq
specif
cont
penser//OBJ
que
613
23
30.3
23
penser//OBJ
le
2333
2
1.6
2
23/08/2014

Demo 4: *X* -> RELATION -> *Y* (where *X* is collocate of *Y*)

Collocates of the verb "aimer" with dependency relations



Cooccurrents Forme pôle | co-présence | co-pr. + pôle

Pôle	Cooccurrent	Fq (Cooccurrent)	co-freq	specif	contextes
aimer//AD	pas	377	5	10.7	4
aimer//PRED	enseigner	7	2	5.3	2
aimer//AD	trop	19	2	4.3	2
aimer//SUB	on	305	2	4.1	2
aimer//SUB	je	786	14	35.2	13
aimer//OBJ	que	613	2	4.5	2
aimer//AD	beaucoup	56	3	6.3	3

Le Trameur

Conclusions

- In Progress…
- New perspectives for *Treebanks* exploration
- Successfully tested for monolingual text processing within several research projects in corpus linguistics and discourse analysis (Branca-Rosoff *et al*., 2012; Née *et al*., 2012)
- Potential for processing parallel and comparable text data in distant languages (Zimina and Fleury, 2014)

Le Trameur

References

- **Sonia Branca-Rosoff, Serge Fleury, Florence Lefeuvre and Mat Pires. 2012.** Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP 2000). Sorbonne nouvelle – Paris 3. Online publication: http://cfpp2000.univ-paris3.fr/CFPP2000.pdf
- **Serge Fleury. 2013a**. *Le Trameur. Propositions de description et d'implémentation des objets textométriques. Sorbonne nouvelle – Paris 3.* Online publication: http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf
- **Serge Fleury. 2013b.** *Annotations Rhapsodie pour le Trameur*. *Sorbonne nouvelle – Paris 3*. Online publication: http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur.pdf (v1 base) http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur-v3.pdf (v3 base)
- **Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong and Paola Pietrandrea**. 2012. *Intonosyntactic data structures: The Rhapsodie treebank of spoken French*. *Proceedings of the Linguistic Annotation Workshop, COLING 2012.* Jeju, Republic of Korea, July 2012.
- **Emilie Née, Erin MacMurray and Serge Fleury. 2012.** *Textometric Explorations of Writing Processes*: *A Discursive and Genetic Approach to the Study of Drafts*. *Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012)*. Liège (Belgium), June 2012.
- **Maria Zimina and Serge Fleury. 2014.** *Approche systémique de la résonance textuelle multilingue. Journées internationales d'Analyse statistique des Données Textuelles (JADT 2014).* Paris, June 2014.