

Approches textométriques des brouillons



Serge Fleury
Université Paris 3 Sorbonne nouvelle / ILPGA
CLESTHIA/SYLED/CLA²T
Site : <http://sfweb.no-ip.org>
Email : sergefleury@gmail.com



Le Trameur
<http://www.tal.univ-paris3.fr/trameur>



Sommaire

- Textométrie
 - Objet, objectif, outils
- Application sur les brouillons
 - Objet, objectif, adaptation aux outils
- Le Trameur
 - Logiciel de textométrie
 - Traitements textométriques des brouillons

La textométrie : préambule

L'objectif de la textométrie est de :

compter des éléments

(des **contenus** textuels)

dans des ensembles

(des **contenants**)

regroupant des unités élémentaires d'un texte ou des zones de texte couvrant un certain nombre ou un certain type d'unités élémentaires

➔ Mettre au jour les particularités de certains contenus dans certains contenants

Contraster les contenants par leurs contenus

Etc.



- Les *contenus* se réalisent sous la forme de ressources textuelles
 - une séquence de caractères organisés en unités élémentaires (les mots par exemple)
- Les *contenants* existent sous la forme de système de masques ou de calques définis sur les contenus
 - Il s'agit de systèmes d'annotations sur tout ou partie des zones textuelles, ces annotations constituant en retour des accès sur les parties textuelles qu'elles définissent
 - Le marquage des phrases ou des paragraphes étant un exemple d'annotation particulier pour décrire un certain niveau de la structure du texte
- Le processus de comptage nécessite au préalable *d'identifier* les *contenus* et les *contenants*.
 - Ce préalable consiste à expliciter une *segmentation* du texte conduisant à la mise au jour d'une *Trame* sur laquelle des annotations pourront se greffer ultérieurement (le *Cadre*)

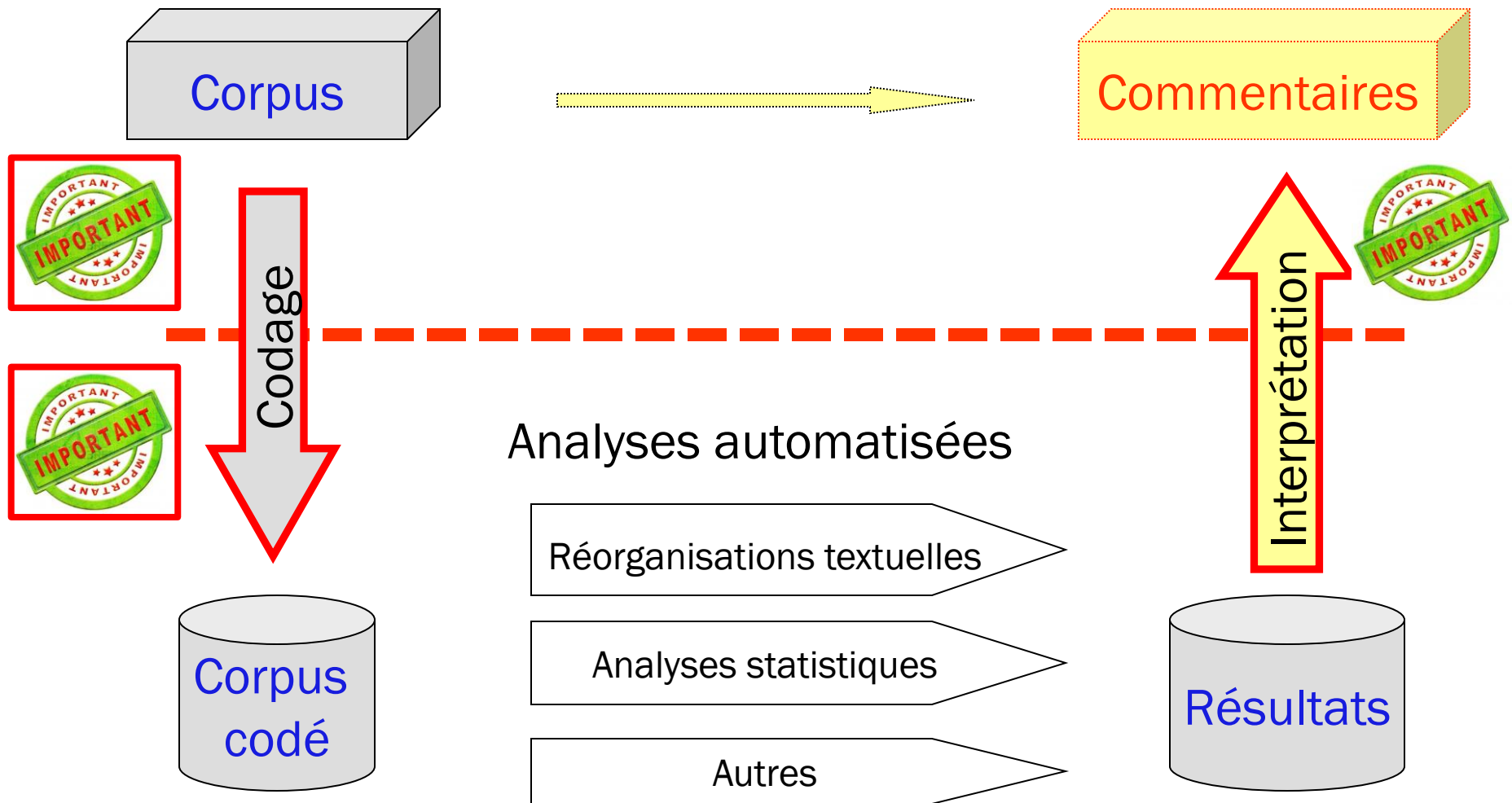
Tableau des formes graphiques du corpus Brouillons (1 dossier : état 1 -> état 24)

Forme (Fq > 0) ▲	Fq. 01	Sp. 01	Fq. 02	Sp. 02	Fq. 03	Sp. 03	Fq. 04	Sp. 04	Fq. 05	Sp. 05	Fq. 06	Sp. 06	Fq. 07	Sp. 07	Fq. 08	Sp. 08	Fq. 09	Sp. 09	Fq. 10	Sp. 10	Fq. 11	Sp. 11
1	0	-1.0	0	-1.1	0	-1.1	1	1.6	1	1.5	1	1.5	1	1.4	1	1.3	1	1.3	1	1.3	1	1.2
16	0	-1.0	0	-1.0	1	2.9	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
1997	0	-1.0	0	-1.0	1	2.9	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
2	0	-1.0	0	-1.1	0	-1.1	1	1.6	1	1.5	1	1.5	1	1.4	1	1.3	1	1.3	1	1.3	1	1.2
2009	0	-1.0	0	-1.1	0	-1.1	0	-1.1	1	1.5	1	1.5	1	1.4	1	1.3	1	1.3	1	1.3	1	1.3
3	0	-1.0	0	-1.1	0	-1.1	1	1.6	1	1.5	1	1.5	1	1.4	1	1.3	1	1.3	1	1.3	1	1.2
4	0	-1.0	0	-1.1	0	-1.1	1	1.6	1	1.5	1	1.5	1	1.4	1	1.3	1	1.3	1	1.3	1	1.2
5	0	-1.0	0	-1.1	0	-1.1	1	1.6	1	1.5	1	1.5	1	1.4	1	1.3	1	1.3	1	1.3	1	1.2
6	0	-1.0	0	-1.0	1	2.9	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
a	0	-1.5	0	-1.8	3	-1.4	3	-1.3	5	-1.2	6	-1.2	7	-1.3	7	-1.5	10	-1.2	10	-1.3	11	-1.3
abordés	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1
absence	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accentué	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.2	0	-1.2	1	1.4
accessible	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	1	1.4	1	1.4	1	1.3
accompagnement	0	-1.1	1	1.4	3	1.8	3	1.9	3	1.7	3	1.5	3	1.4	3	1.3	3	-1.2	3	-1.2	4	1.3
accompagné	1	2.4	1	2.1	1	1.8	1	1.8	1	1.7	1	1.6	1	1.5	1	1.5	1	1.4	1	1.4	1	1.4
accordaient	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accueil	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.2	0	-1.2	1	1.4
activité	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.3	0	-1.3	1	1.2	1	-1.1	1	-1.2	1	-1.2	1	-1.2
activités	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.2
actuel	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	1	1.6	1	1.5	1	1.5	1	1.4	0	-1.2	0	-1.2
actuellement	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.2
adapté	0	-1.0	1	2.2	1	1.8	1	1.8	1	1.7	1	1.7	1	1.6	1	1.5	1	1.5	1	1.5	1	1.4
adaptées	0	-1.0	0	-1.1	0	-1.1	0	-1.1	1	1.5	1	1.5	1	1.4	1	1.3	1	1.3	1	1.3	1	1.3
ado	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1
adolescence	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1
adulte	0	-1.1	0	-1.1	0	-1.3	0	-1.3	0	-1.3	0	-1.4	1	-1.2	1	-1.2	1	-1.3	2	1.3	3	1.5
adultes	1	1.7	1	1.5	1	1.2	1	1.2	1	-1.2	1	-1.2	2	1.3	3	1.5	4	1.8	4	1.7	4	1.6
affectionne	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.2	0	-1.2
agit	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
agressifs	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.3	0	-1.3	0	-1.4	2	1.6	2	1.5	2	1.5	2	1.4
agressivité	0	-1.0	1	1.9	1	1.6	1	1.6	1	1.5	1	1.4	2	2.0	2	1.9	2	1.8	1	1.3	1	1.2
aide	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.1
aider	0	-1.0	0	-1.0	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.2	0	-1.2

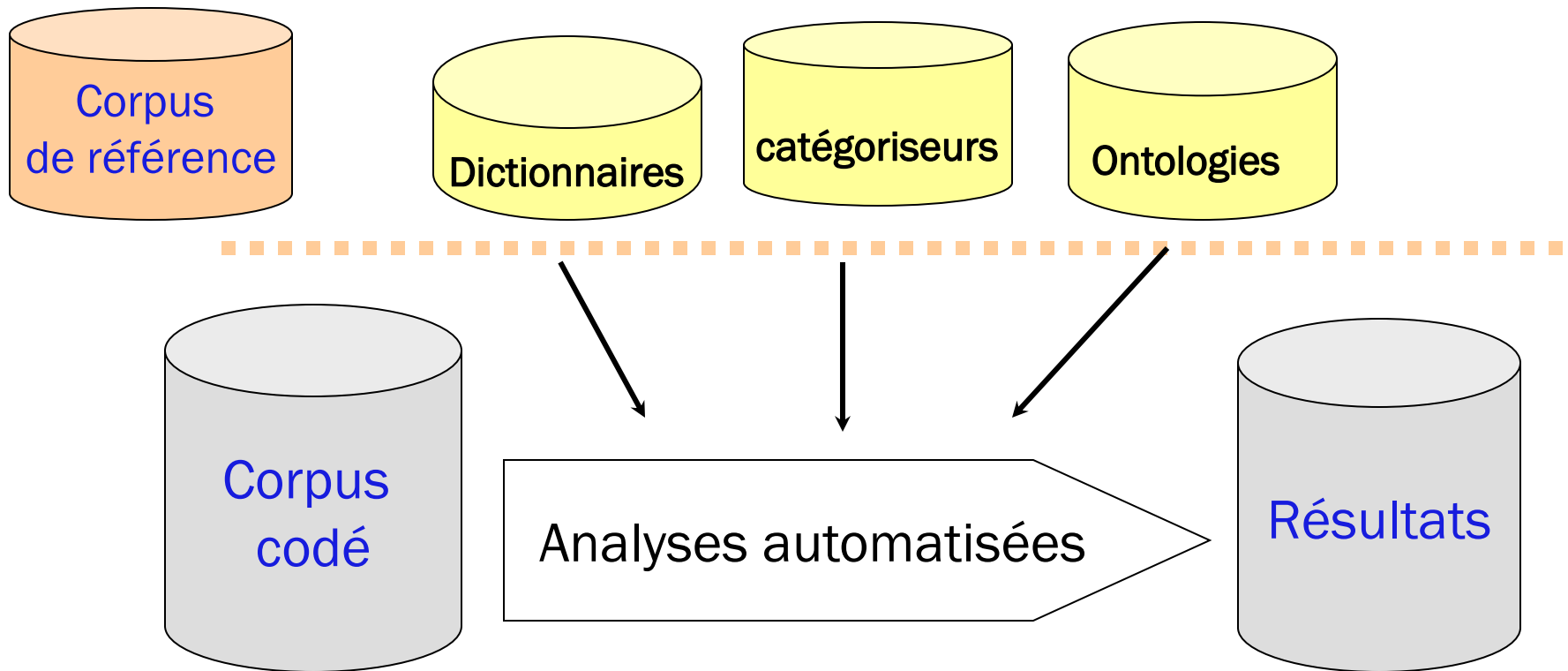
Tableau des formes graphiques du corpus Brouillons (états finaux)

Forme (Fq > 0) ▲	Fq. afara	Sp. afara	Fq. baillo	Sp. baillo	Fq. cabarry	Sp. cabarry	Fq. chapel	Sp. chapel	Fq. correia	Sp. correia	Fq. desmoulins	Sp. desmoulins	Fq. dumont	Sp. dumont
8	0	-1.1	0	-1.2	2	2.4	0	-1.1	2	2.1	0	-1.2	0	-1.2
9	1	1.6	0	-1.2	0	-1.2	0	-1.1	0	-1.2	0	-1.2	0	-1.2
96	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
97	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
a	15	-1.3	28	1.9	12	-2.4	11	-2.2	25	-1.4	21	-1.8	46	3.2
abandon	0	-1.0	0	-1.1	0	-1.1	0	-1.0	0	-1.1	0	-1.1	0	-1.1
abattue	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
abcd	3	1.5	5	1.9	6	2.4	0	-2.0	6	1.9	8	2.8	4	1.3
abcdavait	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
abord	0	-1.0	0	-1.0	1	2.4	0	-1.0	0	-1.0	0	-1.0	0	-1.0
aborde	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.1
aborder	1	1.8	1	1.6	0	-1.1	0	-1.1	0	-1.1	0	-1.1	1	1.5
abordé	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	1	2.2
abordés	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
absence	1	1.3	0	-1.4	1	1.2	0	-1.3	0	-1.5	0	-1.5	8	5.6
absences	0	-1.1	0	-1.1	1	1.8	0	-1.1	0	-1.1	0	-1.1	0	-1.1
absente	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.2	0	-1.2	0	-1.2
absolument	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
abus	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1	0	-1.1
abîme	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accentuer	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	1	2.3	0	-1.0
accentué	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
acceptait	1	2.5	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accepte	0	-1.1	1	1.4	0	-1.2	0	-1.2	0	-1.3	0	-1.3	0	-1.3
accepter	0	-1.1	1	1.7	0	-1.1	0	-1.1	1	1.6	0	-1.1	0	-1.1
accepté	0	-1.1	0	-1.2	0	-1.2	0	-1.1	0	-1.2	0	-1.2	0	-1.2
acceptée	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
acceptées	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accessible	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accident	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.1
accidentel	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accompagne	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0
accompagne	1	-1.2	1	-1.3	2	1.4	1	-1.2	1	-1.4	5	2.4	1	-1.5
accompagne	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.0	0	-1.1

Analyses textuelles sur corpus



Ressources textuelles



Outils et méthodes pour la Textométrie

Lectures :

- *Pratique des Segments répétés* (1987) A. Salem
- *Statistique textuelle* (1994) A. Salem, L. Lebart
http://issuu.com/sfleury/docs/st-1994-lebart_salem
- *Exploring Textual Data* (1998) A. Salem, L. Lebart & L. Berry
- *Les linguistiques de corpus* (2000) A. Salem, B. Habert & A. Nazarenko
http://lexicometrica.univ-paris3.fr/livre/les_linguistiques_de_corpus_1997/
- *Explorations Textométriques* (2009) A. Salem, S. Fleury,
<http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm>
- wiki (TAL-Lexicométrie)
<http://tal.univ-paris3.fr/wakka/wakka.php?wiki=PagePrincipale>



Textométrie

- Objet traité :
 - le texte (les textes)
- Objectif :
 - définir et *compter* des unités dans ces textes
- Moyens :
 - outils / méthodes statistiques textuelles

Le texte

- Flux textuel : séquence organisée des caractères
 - la *Trame** : une liste ordonnée de positions annotées

Annotation importée n°1	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	
Catégorie		Cat(Le)		Cat(dormeur)		Cat(du)		Cat(val)	...
Lemme		Lemme(Le)		Lemme(dormeur)		Lemme(du)		Lemme(val)	...
Forme		Le		dormeur		du		val	...
Positions	1	2	3	4	5	6	7	8	...

- Mise au jour d'une organisation du texte
 - le *Cadre** : ensemble de sélections de positions

<STRUCTURE="TITRE"> Le dormeur du val									
<STRUCTURE="TEXTEPOEME">									
LIGNE="VERS1"> C' est un trou de verdure où chante une rivière,									
LIGNE="VERS2"> Accrochant follement aux herbes des haillons									
LIGNE="VERS3"> D' argent où le soleil ; de la montagne fière,									
Annotation n°1	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	
Catégorie	Cat(Le)			Cat(dormeur)		Cat(du)		Cat(val)	...
Lemme	Lemme(Le)			Lemme(dormeur)		Lemme(du)		Lemme(val)	...
Forme	Le			dormeur		du		val	...
positions	1	2	3	4	5	6	7	8	...
LIGNE="VERS10">Sourirait un enfant malade, il fait un somme :									
LIGNE="VERS11">Nature, berce-le chaudement : il a froid.									
LIGNE="VERS12">Les parfums ne font pas frissonner sa narine ;									
LIGNE="VERS13">Il dort dans le soleil, la main sur sa poitrine,									
LIGNE="VERS14">Tranquille. Il a deux trous rouges au côté droit.									
<STRUCTURE="AUTEUR">Arthur Rimbaud									

Trame+Cadre → exploration textométrique

- A partir d'un texte segmenté, la numérotation des items découpés dans le texte de départ permet de constituer un système de coordonnées sur le texte dans lequel chaque item est repéré par son numéro d'ordre : ***la Trame textométrique***
- Ce même système de coordonnées permet de définir et de localiser, au sein du corpus, des zones textuelles (zones formées par une suite d'items consécutifs, entre la position x_1 la position x_2 , réunion d'un certain nombre de zones de ce type, etc.) qui correspondent aux contenants de l'analyse textométrique (parties, paragraphes, phrases, sections, chapitres etc.) : ***le Cadre textométrique***
- La mise au jour d'une ressource textuelle constituée sous la forme **Trame/Cadre** constitue une solution suffisante pour servir de base à toute exploration textométrique

Moteur textométrique

Texte (unités, délimiteurs) + parties¹

<partie num=1>

x_1 x_2 , x_3 x_2 . x_4

x_5 ... x_k §

<partie num=2>

x_{k+1} x_2 , x_{k+2} x_4 .

x_2 x_{k+3} ... x_n

§

etc.

Délimiteurs

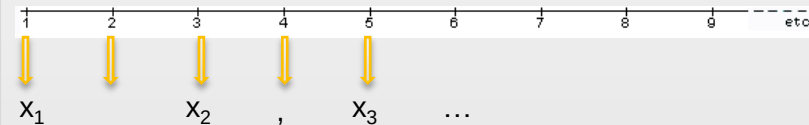
., ; ! ? + = () § etc.

Segmentation

des données

TRAME

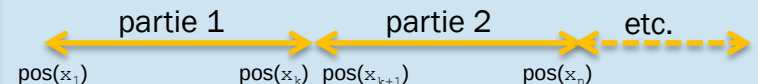
une liste de positions annotées



Forme : x_3
Lemme : $\text{lem}(x_3)$
Catégorie : $\text{cat}(x_3)$
Annotation4 : $\text{ann4}(x_3)$
Etc.

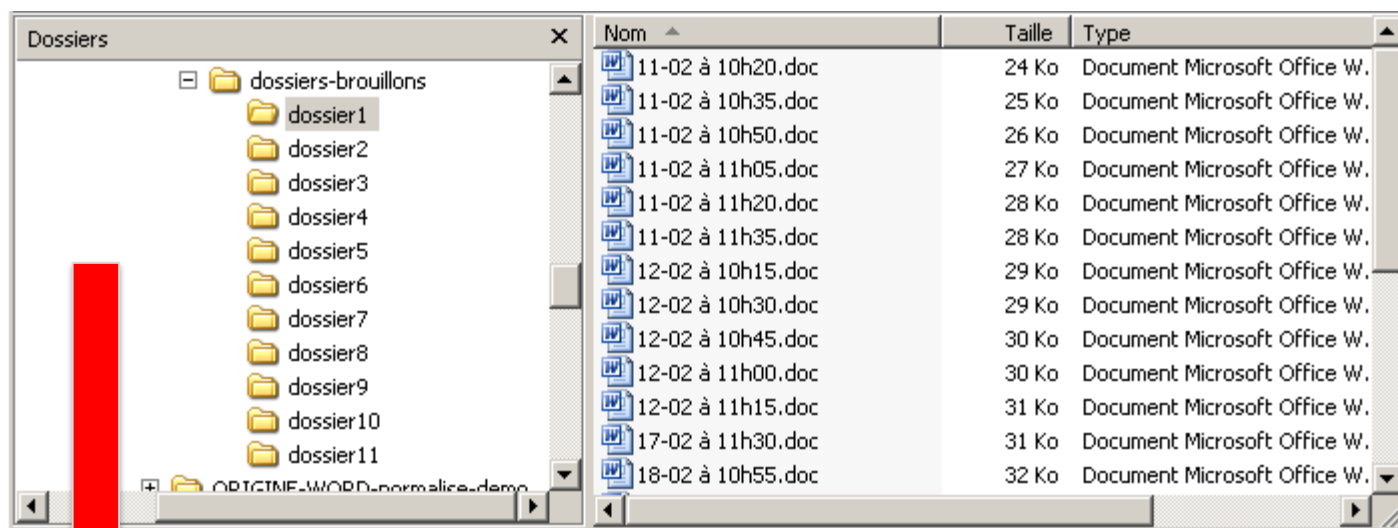
CADRE

ensemble de listes de couples de positions
(1 liste = 1 partition)



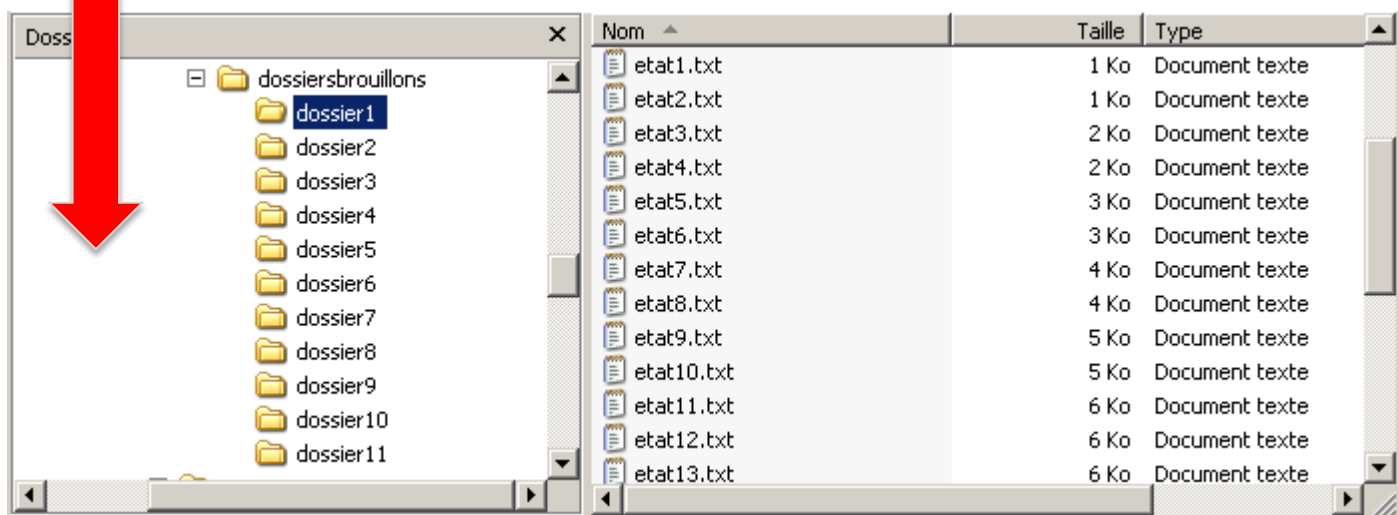
(1) : le codage des parties peut être réalisé suivant différents formats (XML ou assimilés)

Le corpus Brouillons



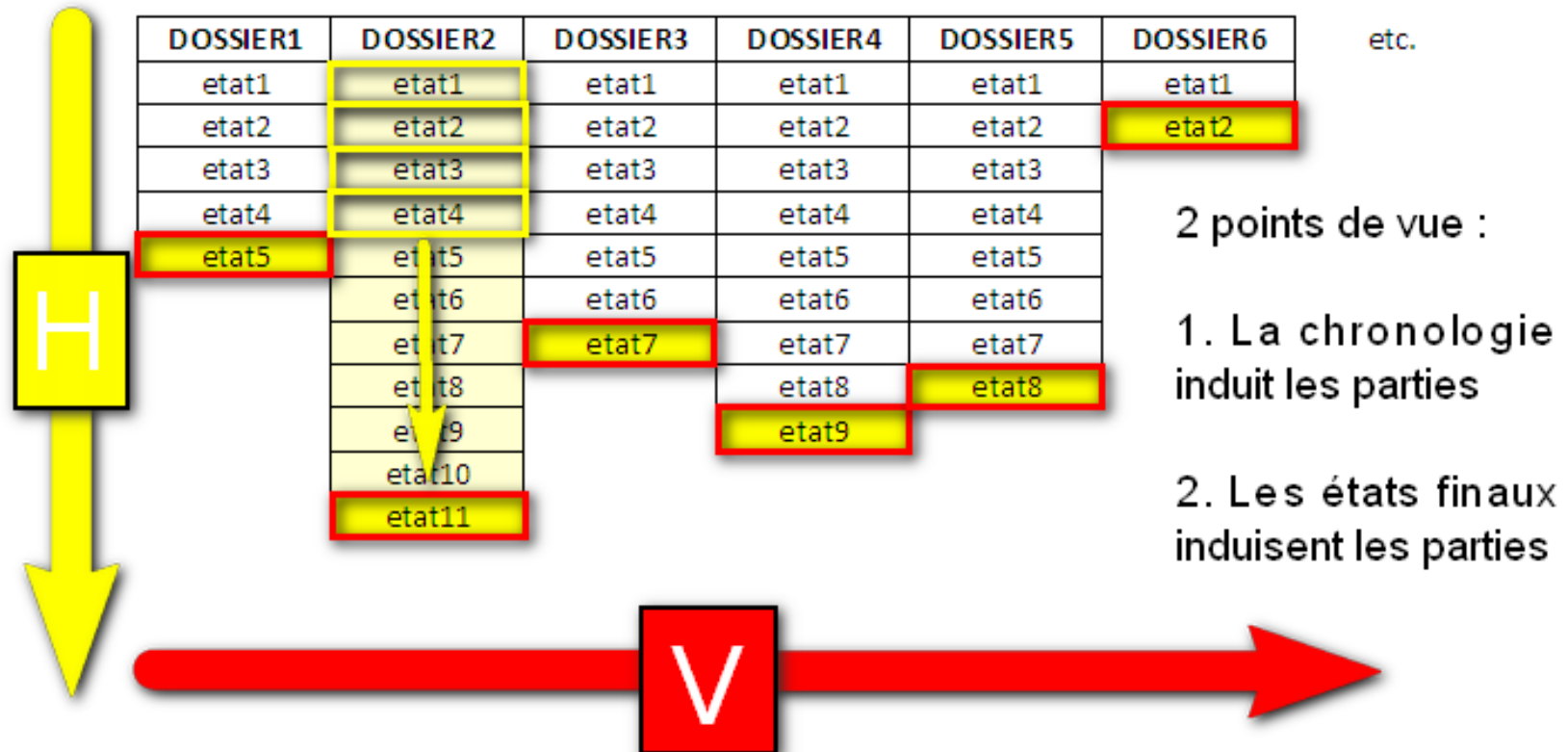
Le corpus : une pile de dossiers

1 dossier : différents états de rédaction horodatés



Première étape : anonymisation et conversion en texte brut

Les parties dans les brouillons



On pourrait aussi mettre au jour des parties au sein des états : « rubriquage »

Différents points de vue sur le texte

prépa synthèse anthony viti. \$
plan \$
information concernant le placement. \$
connaissance de l'histoire familiale \$
raison du placement (demande parentale, du jeune, du service demandeurs). \$
observation depuis son arrivée sur le groupe : \$
attitudes relationnelles avec ses pairs, les adultes \$
comportement en général sur le groupe, à l'extérieur, \$
trouble encopréatique observé et accompagné- rapport au corps : soin, \$
hygiène, apparence. \$
l'illusion (faux self ou non) \$
la colère contenu \$
investissement du placement \$

prépa synthèse anthony viti. \$
plan \$
information concernant le placement. \$
connaissance de l'histoire familiale \$
raison du placement (demande parentale, du jeune, du service demandeurs). \$
observation depuis son arrivée sur le groupe : \$
attitudes relationnelles avec ses pairs, les adultes \$
comportement en général sur le groupe, à l'extérieur, \$
trouble encopréatique observé et accompagné \$
- rapport au corps : soin, \$
hygiène, apparence. \$
l'illusion (faux self ou non) \$
la colère contenu \$
investissement du placement \$

insertion de caractères "marqueurs de sections" (phrases, paragraphes, etc.

etat 01

num

prépa synthèse anthony viti. \$
plan \$
information concernant le placement. \$
connaissance de l'histoire familiale \$
raison du placement (demande parentale, du jeune, du service demandeurs). \$
observation depuis son arrivée sur le groupe : \$
attitudes relationnelles avec ses pairs, les adultes \$
comportement en général sur le groupe, à l'extérieur, \$
trouble encopréatique observé et accompagné \$
- rapport au corps : soin, \$
hygiène, apparence. \$
l'illusion (faux self ou non) \$
la colère contenu \$
investissement du placement \$

extrait de la Trame : chaque unité (mot) occupe une position sur la Trame

section

découpage en parties :
une partie peut être
définie par une position
de début et une position
de fin sur la Trame

investissement du placement

Position: <158>
Forme: <investissement> | Freq: 14

Shift-clic sur carré : affichage | clic-droit sur carré : spécificités | Control-clic sur carré : sélection | Shift-Control-clic sur sélection : désélection

Seuillage : 1 5 10 ++ | Modifier seuillage

Mo L Sections sélectionnées : 0 N° Sect. : 18: (177,184) Annotation : 1 Agerçu : 56

1 investissement du placement \$

etat num_01

etat num_02

etat num_03

etat num_04

etat num_05

etat num_06

etat num_07

etat num_08

etat num_09

Control-clic sur marqueur de page : sélection 5 sections | Shift-control-clic sur marqueur de page : sélection 25 sections (1 ligne)

Mo L Sections sélectionnées : 0 N° Sect. : 18: (177,184) Annotation : 1 Agerçu : 56

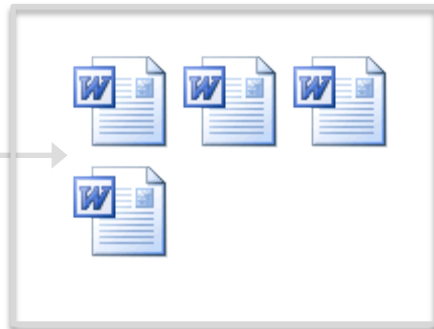
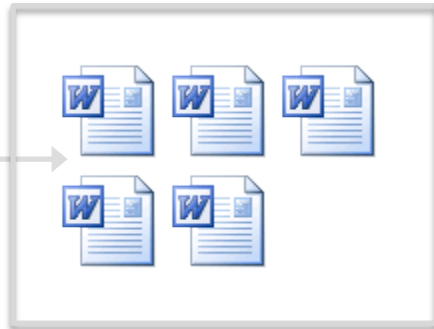
1 investissement du placement \$

Position: <178>
Forme: <investissement> | Freq: 14

Des brouillons aux objets « texte »

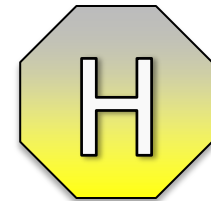
Dossier n° 1:

5 états de rédaction



Dossier n° 2 :

4 états de rédaction



Corpus de travail

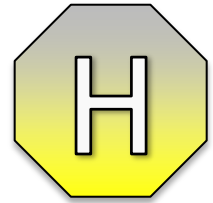


Phase de travail fondamentale
Préparation et organisation des
données

Le « texte » et les brouillons

- Approches retenues

- Approche horizontale



- Analyse chronologique d'un dossier



- On privilégie *l'aspect dynamique*

- Approche verticale



- Analyse de tous les états finaux de tous les dossiers disponibles



- On privilégie *l'aspect contrastif*

Trame/Cadre sur les brouillons (1)

- Approche horizontale
 - Un dossier : un corpus chronologique (de l'état 1 à l'état final)
 - Tous les états d'un même dossier sont empilés « les uns derrière les autres »
 - Mise en place d'un Cadre : partition en état de rédaction

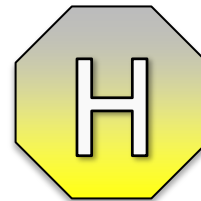


```
<etat=01> texte-01
```

```
<etat=02> texte-02
```

```
...
```

```
<etat=XX> texte-XX
```



➤ Exemple : dossier1-chronologique.txt

➤ **Autre solution** : dossier1-alignement.tmx (cf allongos)

Trame/Cadre sur les brouillons (2)

- Approche verticale
 - Regroupement de tous les états finaux
 - Tous les états finaux de chaque dossier sont regroupés « les uns derrière les autres »
 - Mise en place d'un Cadre : partition en dossier



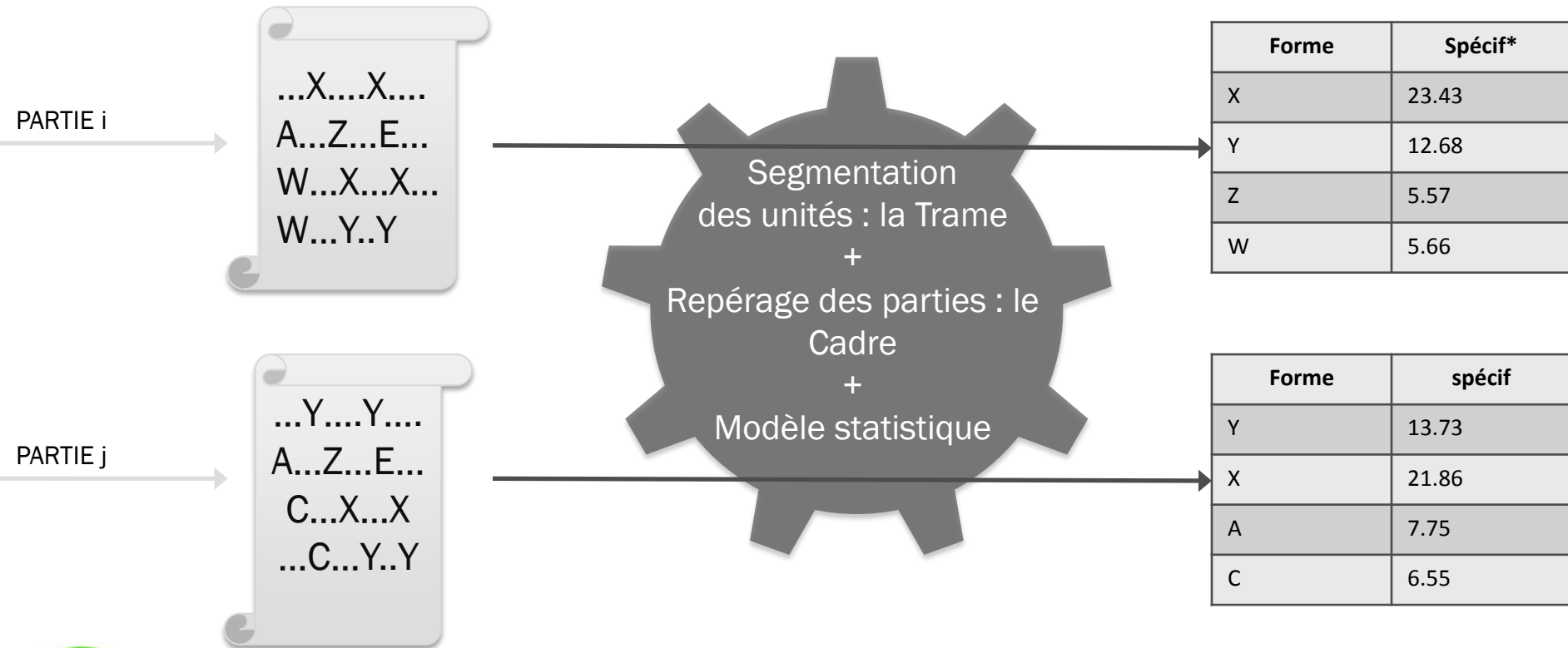
```
<dossier=01-final> texte-01-final  
<dossier=02-final> texte-02-final  
...  
<dossier=XX-final>texte-XX-final
```



➤ exemple : corpus-etat-final.txt

Moteur textométrique

« Classiquement », la textométrie construit des tableaux lexicaux à partir des données à analyser sur la base de processus de segmentation + modèle statistique



Dans le cas du Trameur, aucun tableau lexical n'est disponible directement (calcul dynamique en cas de besoin) : les unités ne sont pas « figées »

Spécif* : spécificité positive (vs négative)

pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

Le modèle hypergéométrique a vu de nombreuses applications : les *spécificités* par partie du corpus (Lebart & Salem, 1994 ; Lafon, 1980) le calcul des cooccurrences (Lafon, 1981 ; Martinez, 2003).

La formule classique de la distribution hypergéométrique est la suivante¹⁴ :

T : le nombre d'occurrences dans le corpus

t : le nombre d'occurrences dans la partie étudiée (ici le mois)

F : la fréquence du mot dans le corpus

f : la fréquence du mot dans la partie étudiée (ici le mois)

$$P[X=f] = \frac{\binom{F}{f} \binom{T-F}{t-f}}{\binom{T}{t}}$$

On commence par imaginer une *population* d'objets d'effectif total T . Parmi tous ces objets, on s'intéresse à une catégorie d'objets, dits marqués (ex : toutes les occurrences d'une forme distribuée dans l'ensemble T), les objets restant tous confondus en un même sous-ensemble d'effectif $T-F$. Prélevons maintenant, en pratiquant des tirages aléatoires sans remise, un *échantillon*, contenant exactement t objets dans la population totale. À l'issue de ce tirage, le nombre d'objets qui nous intéressent que contient l'échantillon est noté f . Pour porter un jugement sur le résultat f , il nous faut le situer parmi des comptages de même nature qui correspondent à l'ensemble de tous les échantillons, composé de t objets, qu'il est possible de prélever à partir de la population de départ. Pour chaque échantillon de taille t , le nombre f des objets marqués peut prendre une valeur obligatoirement comprise entre 0 et F . Inversement, toujours pour des nombres T , t , et F fixés et pour chaque nombre n compris entre 0 et F , il est possible de recenser le nombre $N(n)$ des échantillons de longueur t pour lesquels f est strictement égale à n . Si l'on divise chacun des nombres $N(n)$ par le nombre total des échantillons de longueur t on obtient une distribution de probabilité (de paramètres T , t et F) sur l'ensemble des nombres compris entre 0 et F .

On peut maintenant utiliser la distribution de probabilité construite à partir des paramètres T , t et F pour porter un jugement sur la fréquence absolue n_0 observée lors du tirage de l'échantillon. Pour cela nous commencerons par situer n_0 par rapport au *mode* (valeur la plus probable) de la distribution. Si cette valeur est très proche du mode nous ne pourrions pas dire grande chose à propos du résultat. Si en revanche elle lui est nettement supérieure, nous calculerons la quantité $P_{sup}(n_0)$ qui est la probabilité de voir apparaître toujours sous les hypothèses retenues plus, un nombre d'objets qui nous intéressent égal ou supérieur à n_0 parmi les t objets prélevés au hasard. Si n_0 est inférieur au mode, nous calculerons de la même manière la quantité $P_{inf}(n_0)$ qui est la probabilité de voir apparaître, toujours sous les mêmes hypothèses, un nombre d'objets qui nous intéressent égal n ou inférieur à n_0 .

Le résultat de ce calcul, le degré de probabilité d'apparition d'une unité textuelle dans une partie du corpus, est alors comparé à un seuil préétabli. Il s'agit d'un seuil de probabilité fixé arbitrairement au début de l'analyse et qui reste stable pour toutes les unités analysées¹⁵. En fonction de la probabilité $P_{sup}(f)$ ou $P_{inf}(f)$ (position de f par rapport au mode de la distribution), trois situations peuvent se présenter :

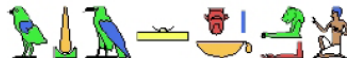
- b) spécificité dite **banale** - si aucun des deux probabilités $P_{sup}(f)$ et $P_{inf}(f)$ ne se révèle inférieur au seuil, on dira dans ce cas que la forme est *banale* pour la partie du corpus.
- S+) spécificité **positive** - si la probabilité $P_{sup}(f)$ est en dessous du seuil, la forme est sur-employée dans la partie étudiée¹⁶ et considérée comme forme spécifique positive de cette partie.
- S-) spécificité **négative** - si la probabilité $P_{inf}(f)$ est inférieure au seuil, la forme est sous-employée dans la partie étudiée et considérée comme forme spécifique négative.

La méthode des spécificités fournit donc un indice qui signale un emploi **atypique d'une** forme pour une partie donnée du corpus. Une spécificité positive indique qu'une unité textuelle est abondamment employée au sein de la partie alors qu'une spécificité négative montre une tendance de cette même partie à éviter l'emploi de l'unité dans la partie considérée.

14. Lebart & Salem, (1994 : 173-177)
 15. Dans le logiciel Lexico 3, le seuil par défaut est fixé à 5 et le calcul est fait sur toutes les formes qui ont une fréquence égale ou supérieure à 10 dans le corpus. Ces paramètres peuvent être modifiés, pour l'analyse ici les paramètres par défaut ont été maintenus. Rappelons que la codification du seuil peut avoir des conséquences sur les résultats obtenus, il convient de rester vigilant lorsqu'on choisit un seuil et éviter d'utiliser plusieurs seuils de probabilités au cours de comparaisons multiples.
 16. Dans les logiciel Lexico 3 et Le Trameur la spécificité positive est indiquée par le signe + suivi d'une valeur a . À l'inverse La spécificité négative est indiquée par le signe - suivi d'une valeur a . L'indice a indique une probabilité de l'ordre de 10^{-a} que l'unité ait une fréquence supérieure à la valeur constatée.

Les unités textuelles

- Quelles unités ?
- Identification des unités
 - comment on segmente le texte ?

Le conte du naufragé	
1	 I10:D46-M17-N35-T18-G43-A1-M17-N29:D21:Y1
2	 G43-U28-G1-Y1-F34*Z1:V31-F4:D36-A1
3	 G17-D36:V31-F22:D54-N35:N35:Z2-F26:N35-W24-G43-O1
4	 O42:Q3-D40-J1:D21-Q3*Z7:M3-V28-A25-A24-Y5:N35-M17-X1-P11-M3
5	 F4:X1*X1-Z7-D21:D36:X1-D2:Z1-N16:Z1*N23
6	 D21:D36-V28-V31:N35-W24:Z7-A2-R8-N14-A30-A2-O34:A1*Z1-V30-D2:Z1-V28-Q3:X1-D32:D36-T22-N35:W24-G43-A1-Z4:I9

Les unités dans les brouillons

- Segmentation en mots
- Liste de délimiteurs de mots

Délimiteurs utilisés pour segmenter les formes du texte de travail..... `.,:;!?/_-'""00$!*"><=+!«»`

➤ Démo Trameur :

- Segmentation en « séquences de mots »
 - Les unités minimales sont des séquences de mots
 - Repérage des unités « phrase » spécifiques ?
- Segmentation en mots
- Attention au choix des délimiteurs...

Annoter/enrichir les unités

- Etiquetage morphosyntaxique
 - Processus intégré dans *Le Trameur* via *TreeTagger*
- Lemmatisation
 - idem
- Annotation sémantique etc.
 - Possibilité d'importer des systèmes d'annotation sur une Trame donnée

➤ Démo Trameur

➤ La Trame = une liste de positions annotées

Son histoire a connu une fin tellement merveilleuse qu' elle mérite d' être contée dans le temps des Fêtes .

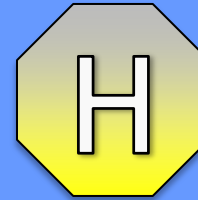
```
Position: <40>
Forme: <histoire> | Freq: 6
Lemme: <histoire> | Freq: 10
Cat: <NOM> | Freq: 3801
a-4: <21> | Freq: 1
a-5: <-> | Freq: 36858
a-6: <Topical_entity(23)> | Freq: 1
a-7: <SUJ(23)> | Freq: 1
```

Outils et méthodes

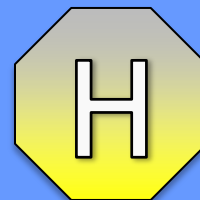
- Fréquence des unités
- Concordance
- Moteur de recherche
- Sélection des unités
- « Poids » des unités dans les parties (spécificités)
- Contraster les parties (par leur vocabulaire)
- Proximités de certaines unités (cooccurents)
- Au delà des unités primaires (les segments)

➤ **Démo Trameur**

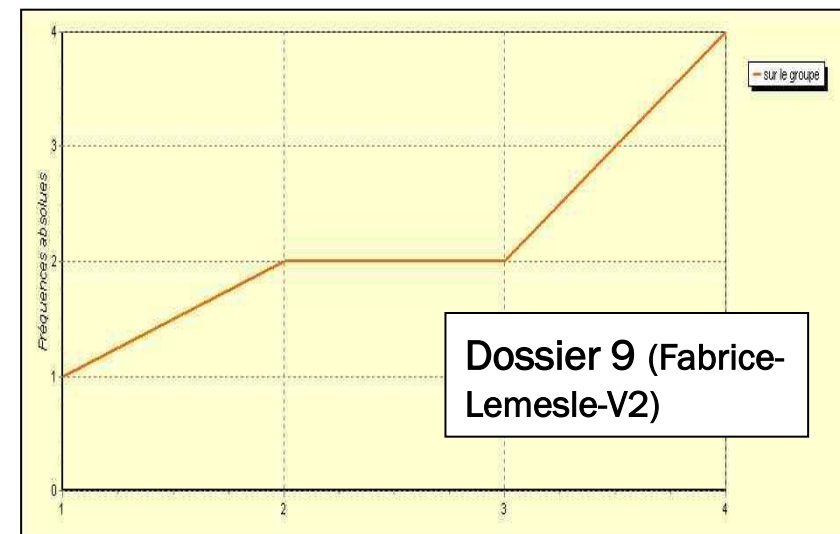
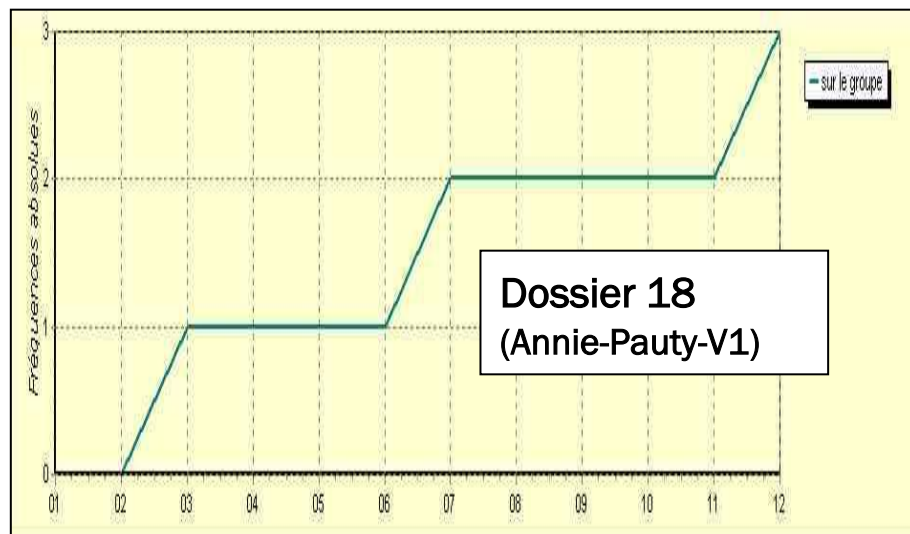
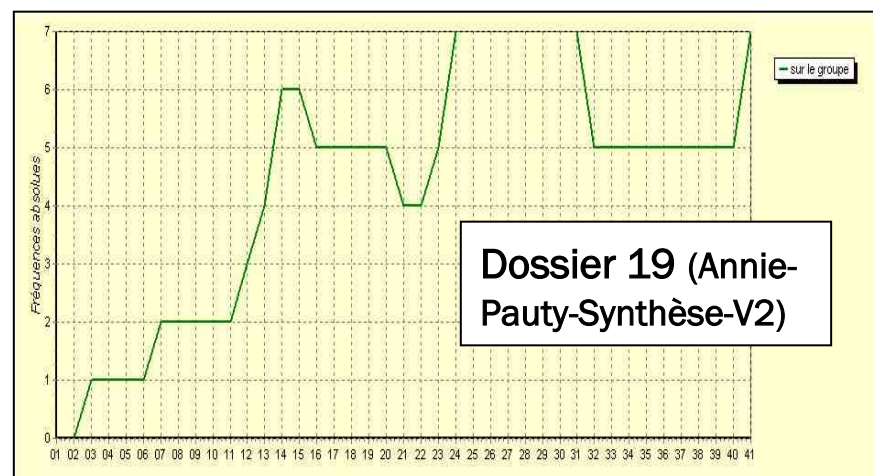
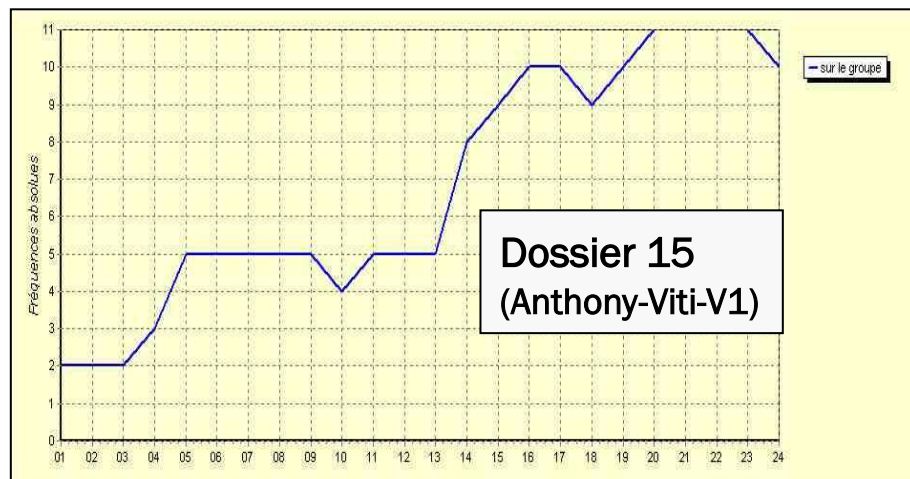
Ventilation



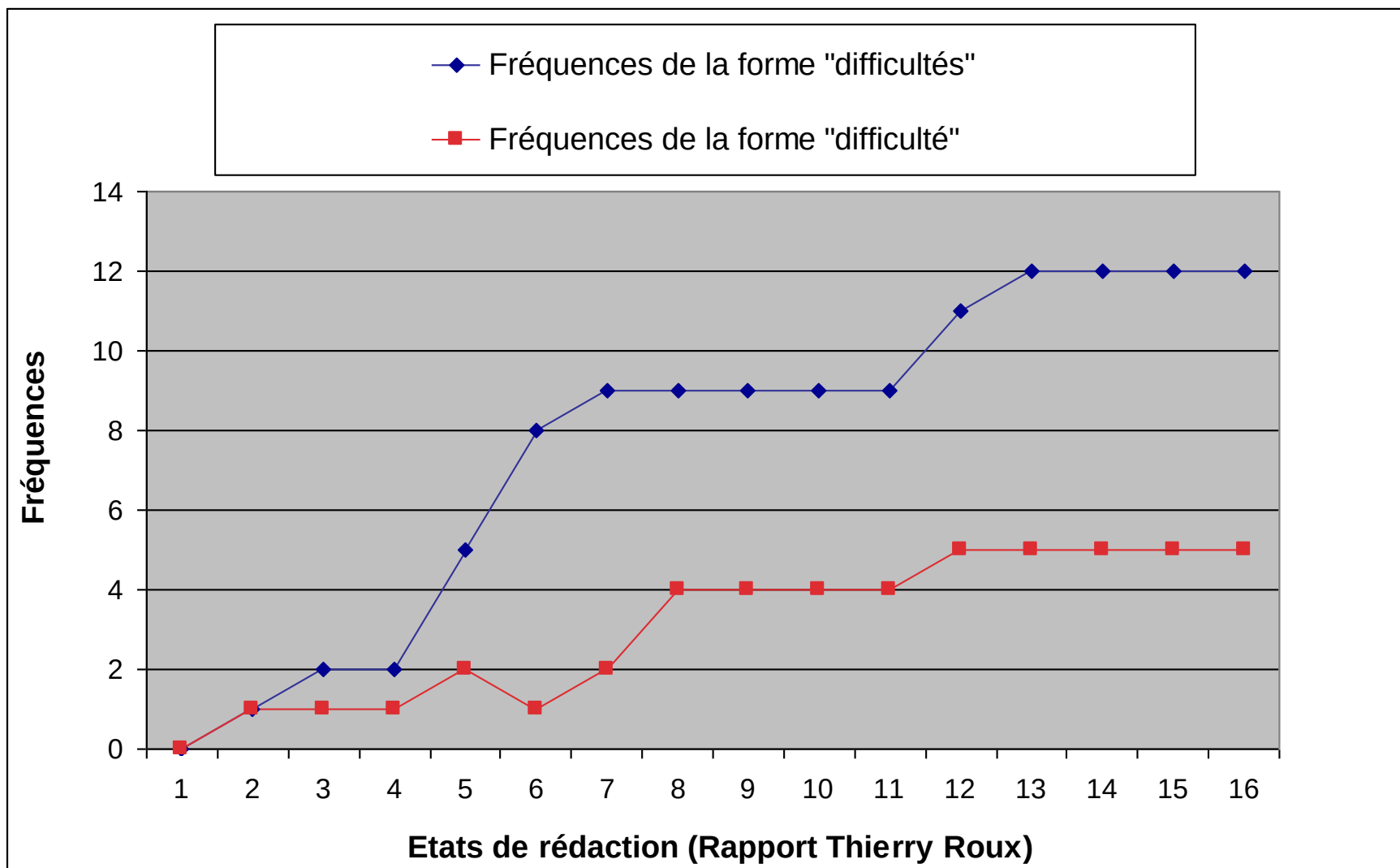
Ventilation



Densification de *sur le groupe*



Ce qui ne va pas



Distribution de fréquence

(axe horizontal)

Distributions de fréquences états par états de la sélection d'items :

→ densification dans le processus d'écriture de ces mots /expressions

→ mots /expressions présents dès les premiers états

Le groupe vs les autres

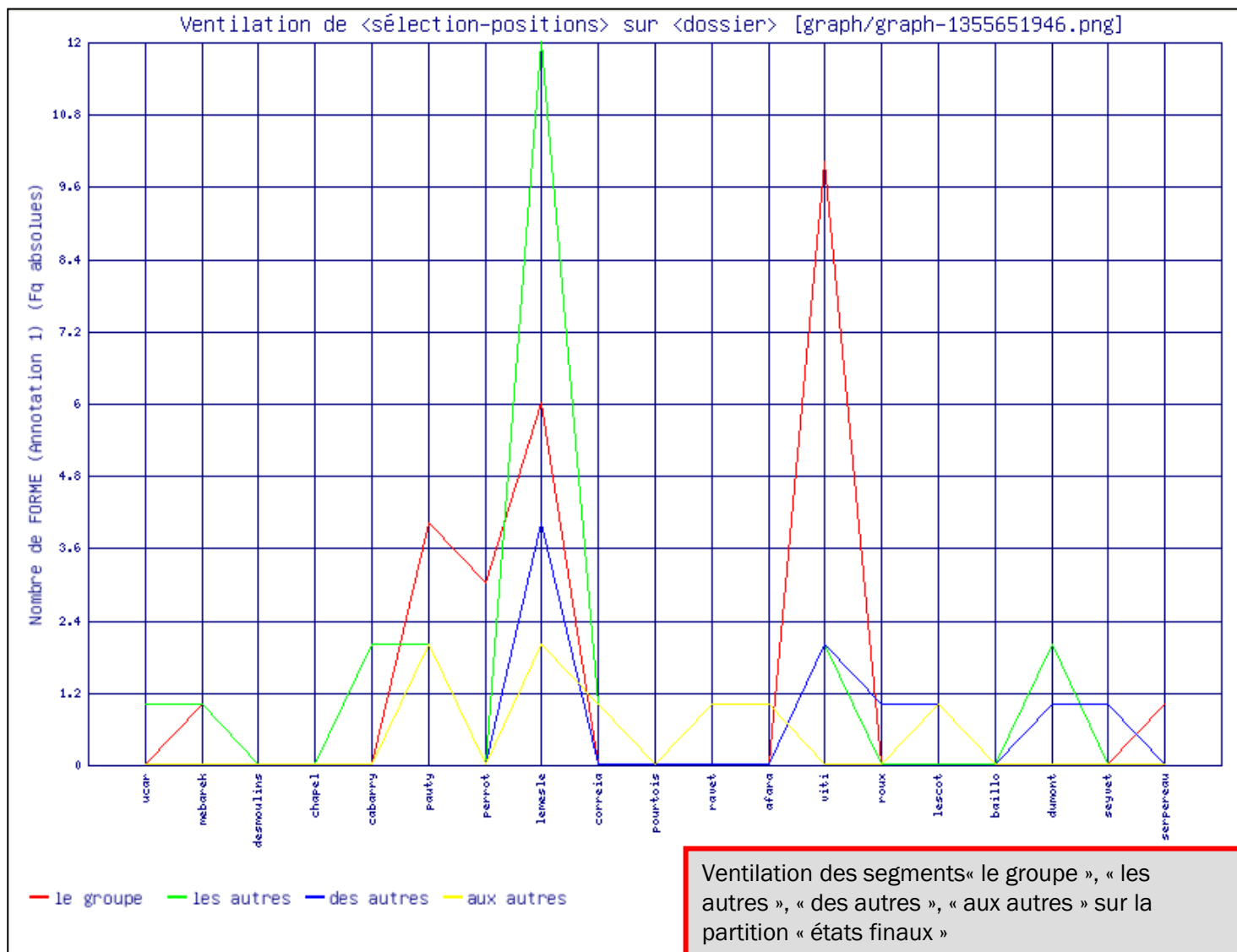
EVALUATION / CONTRASTE de la situation

→ Le rapport au groupe

→ Le rapport aux autres

→ « le groupe » vs « [les,des,aux] autres »

Ventilation



-----PARTIE{dossier=ravet}-----

\$ adèle est plus ouverte **aux** autres qu'auparavant, moins

-----PARTIE{dossier=pauty}-----

a. evolution sur **le** groupe : \$ jeune
mettre en danger) face **aux** autres. se laisse influencer
avec les réponses très agressives **des** autres garçons. \$
filles. sa place sur **le** groupe a évolué. elle
accepté et a préféré quitter **le** groupe (les collègues l
en difficulté lors des confrontations **aux** autres. a pu commencer
chambre est souvent investie par **les** autres filles et nous
la parole des uns et **des** autres. les versions se
dans de nombreuses histoires sur **le** groupe : vol de
du groupe. de plus, **les** autres jeunes évoquent des

-----PARTIE{dossier=viti}-----

l. evolution sur **le** groupe \$ anthony est
désir de venir sur **le** groupe de l'
discretion, son respect pour **les** autres, sa facilité d
trouvé une juste place sur **le** groupe de l'
rentre manger sur **le** groupe deux midis par
-vêtements sans la présence **des** autres jeunes. \$ l
qu'il aime partager avec **les** autres. la lecture
besoin de rester sur **le** groupe, s'isolant
groupe, s'isolant **des** autres. nous l
passer un week-end sur **le** groupe avant les
son bien-être sur **le** groupe ou une demande
et les retours sur **le** groupe sont généralement chaleureux.
son envie de rester sur **le** groupe. \$ 4
depuis son arrivée sur **le** groupe. la transformation physique

-----PARTIE{dossier=perrot}-----

puis, accueil sur **le** groupe d'internat
\$ emmanuel a donc quitté **le** groupe d'internat du
se rendant sans autorisation sur **le** groupe d'internat.

-----PARTIE{dossier=lescot}-----

de respect à l'égard **des** autres élèves voire des
gens les uns **aux** autres pour tenter

-----PARTIE{dossier=correia}-----

\$ dans les relations **aux** autres enfants, germain
les adultes ou **les** autres enfants à qui

-----PARTIE{dossier=ucar}-----

\$ en groupe, avec **les** autres enfants, il semble

-----PARTIE{dossier=nebarek}-----

, dans la mesure où **les** autres enfants du groupe
'autorisations. \$ observations sur **le** groupe : catherine scolarisation :

-----PARTIE{dossier=dumont}-----

être une adolescente comme **les** autres qui vit sa
agité et parfois agressif envers **les** autres enfants. \$ dans
assez gai, très apprécié **des** autres enfants que ce

-----PARTIE{dossier=afara}-----

'à ce jour, renvoyant **aux** autres la responsabilité de

-----PARTIE{dossier=roux}-----

uns **des** autres notamment concernant leurs

-----PARTIE{dossier=seyvet}-----

d'elise à l'égard **des** autres enfants ont

-----PARTIE{dossier=cabarry}-----

un bon contact avec **les** autres enfants.
avec son \$institutrice et **les** autres adultes de l'

-----PARTIE{dossier=lemesle}-----

avec **les** autres enfants, voire de
ii- evolution sur **le** groupe
lors de la visite sur **le** groupe, fabrice se
impulsif et agressif fait fuir **les** autres
entre dans les chambres **des** autres jeunes sans
insulte **les** autres, leur lance de
constamment. il empêche **les** autres de
murs, hurle **aux** autres « on fait
fabrice convoque l'agressivité chez **les** autres jeunes
son agitation se diffuse sur **le** groupe qui est mis
est mis à mal. **les** autres
fabrice **des** autres et de lui-
menace **les** autres. a titre d
lorsqu'il est isolé **des** autres, il peut
relation duelle ou sur **le** groupe.
fabrice est actuellement présent sur **le** groupe pour le repas
a son arrivée sur **le** groupe de le hameau,
(bruitages, insultes, embête **les** autres, détériore la
en classe, tape **les** autres. ils évoquent la
conflit trop vite avec **les** autres jeunes ?
limoges, la rencontre avec **les** autres s'est
par rapport **aux** autres mais il agresse
mais il agresse également allègrement **les** autres,
à la vie au milieu **des** autres interroge sa
-----PARTIE{dossier=serpereau}-----
elle est bien intégrée dans **le** groupe classe. \$activites

Concordance des segments :

« le groupe »

« les autres »

« des autres »

« aux autres »

sur la partition « états finaux »

Hautes fréquences de formes et de segments répétés

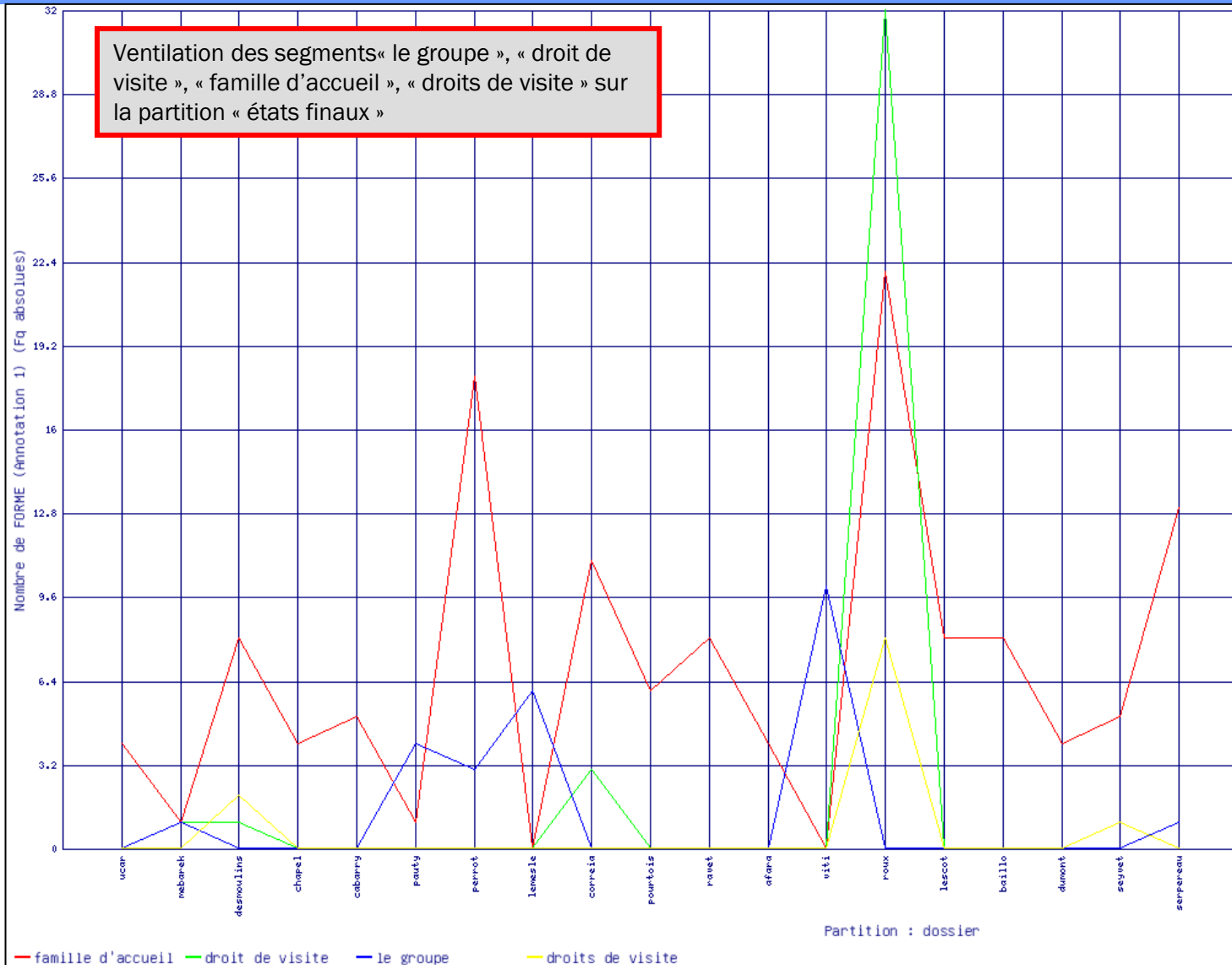
→ Repérage de termes / expressions du secteur éducatif désignant réalités spécifiques au champ (à la sphère) : *famille d'accueil, (droit) de visite, au service, assistante familiale, (pris/prennent/pris) en charge, au abcd, le service, placement familial, le groupe, de placement, etc.*

→ Repérage de mots et de séquences figées dont on a fait l'hypothèse qu'ils étaient des passages obligés du scripteur dans l'évaluation d'une situation: *cadre, difficulté(s), évolution (+ évolue, évoluer, etc.), est en/est dans, problématique, sur le groupe, symptôme...*

Ventilation

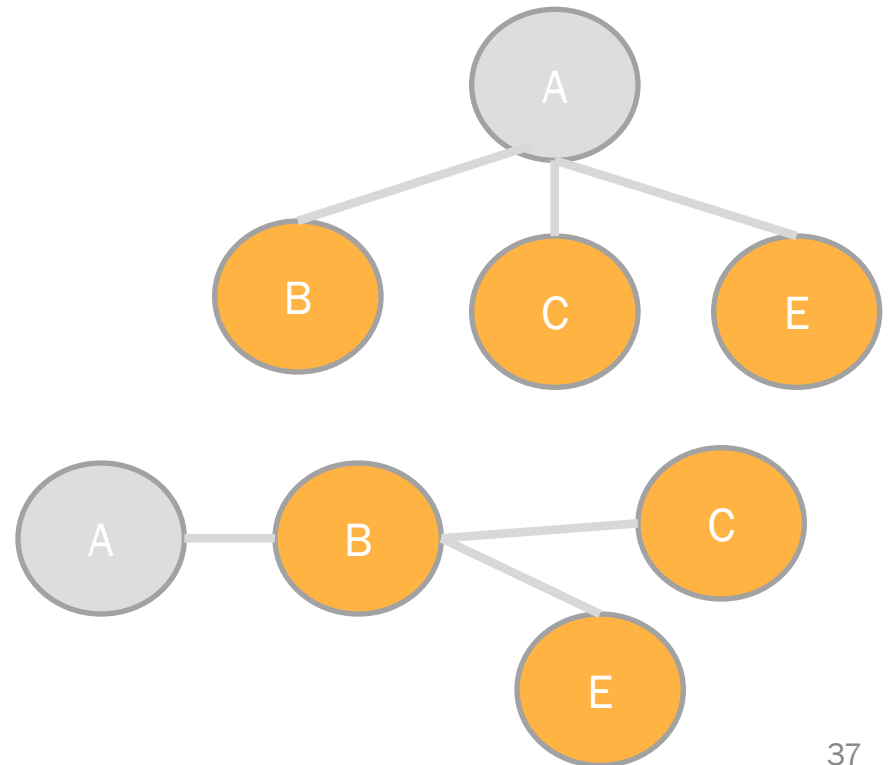
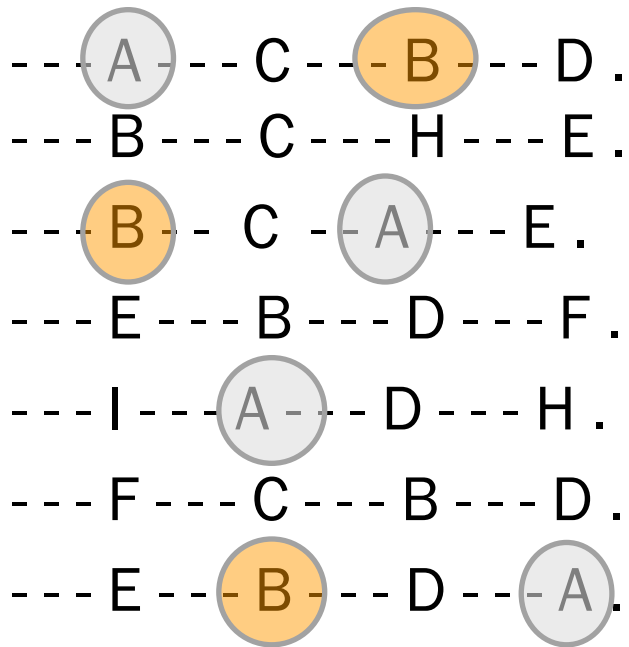


Ventilation des segments « le groupe », « droit de visite », « famille d'accueil », « droits de visite » sur la partition « états finaux »

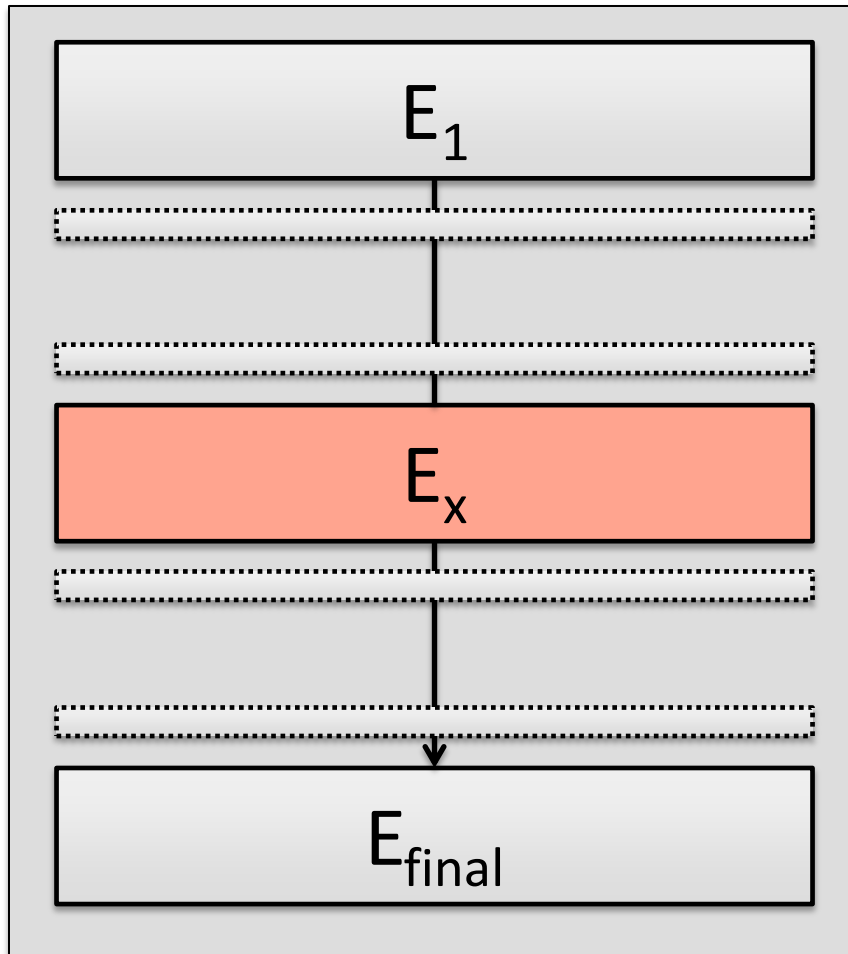
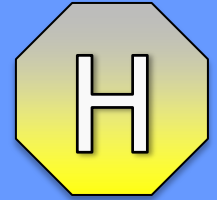


Cooccurents

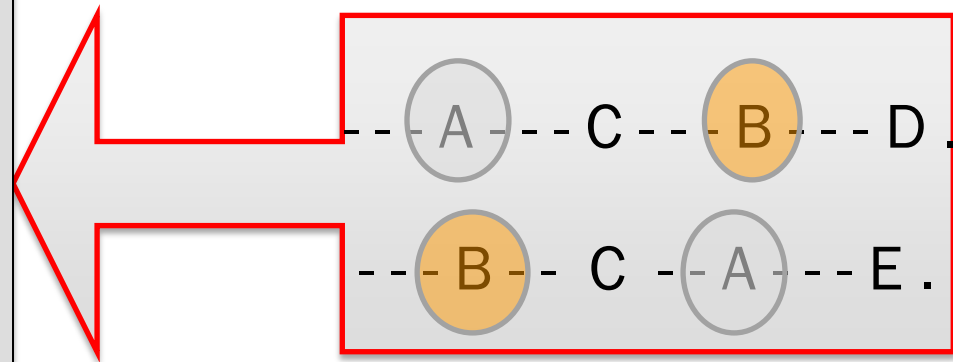
- **Cooccurrence**: la présence simultanée de deux mots dans le même contexte (Lafon, 1981; Martinez, 2003)
- **Résultat** : un réseau de relations



Cooccurents sur corpus chronologique



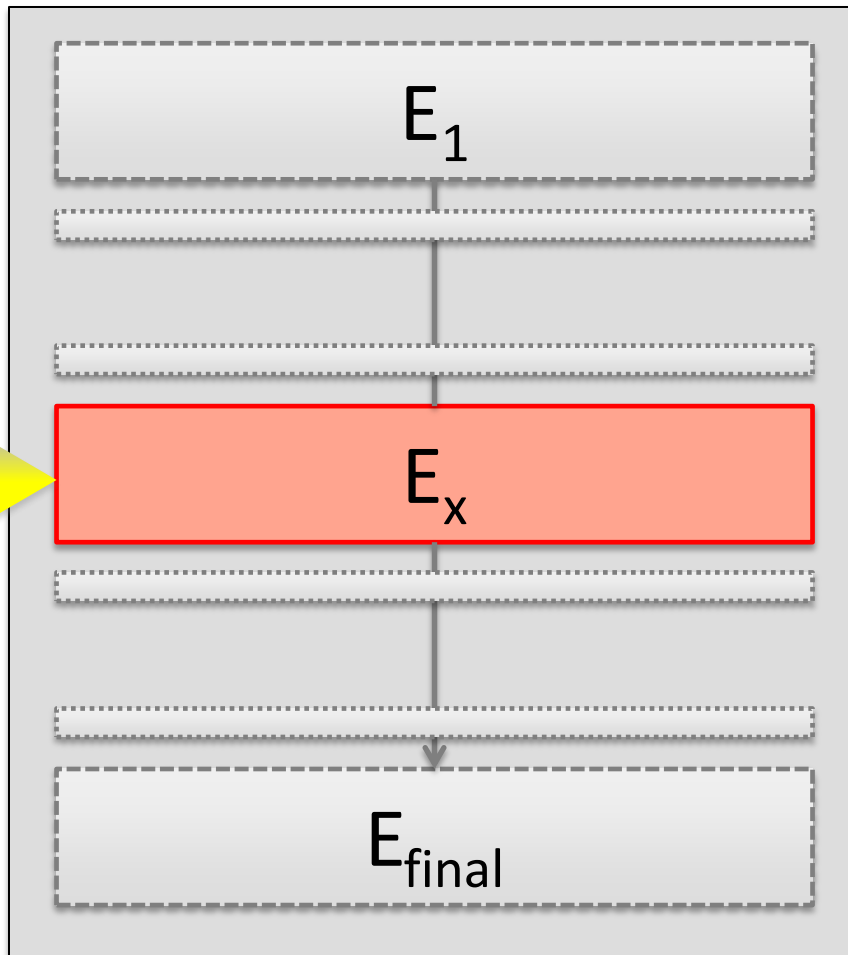
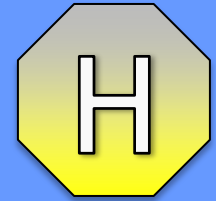
- Le calcul des cooccurents sur l'état E_x s'intéresse aux contextes contenant le pôle visé pour mettre au jour les unités qui y sont sur-représentées



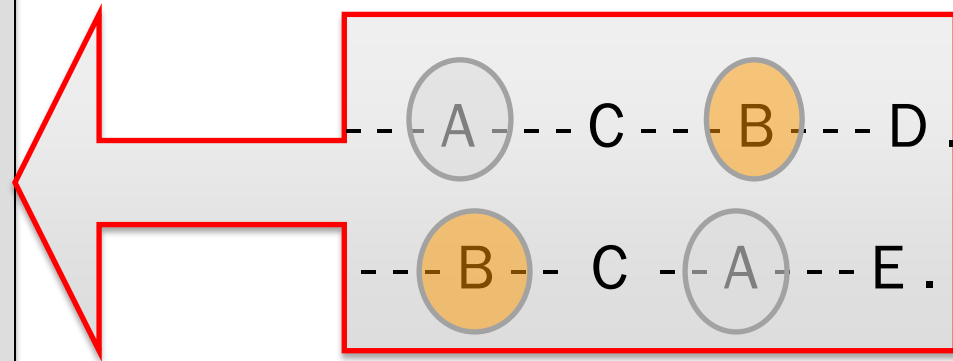
- 3 approches possibles :
 - Locale
 - Globale
 - Chronologique

(màj 10.66 dans Le Trameur)

Cooccurents : approche locale

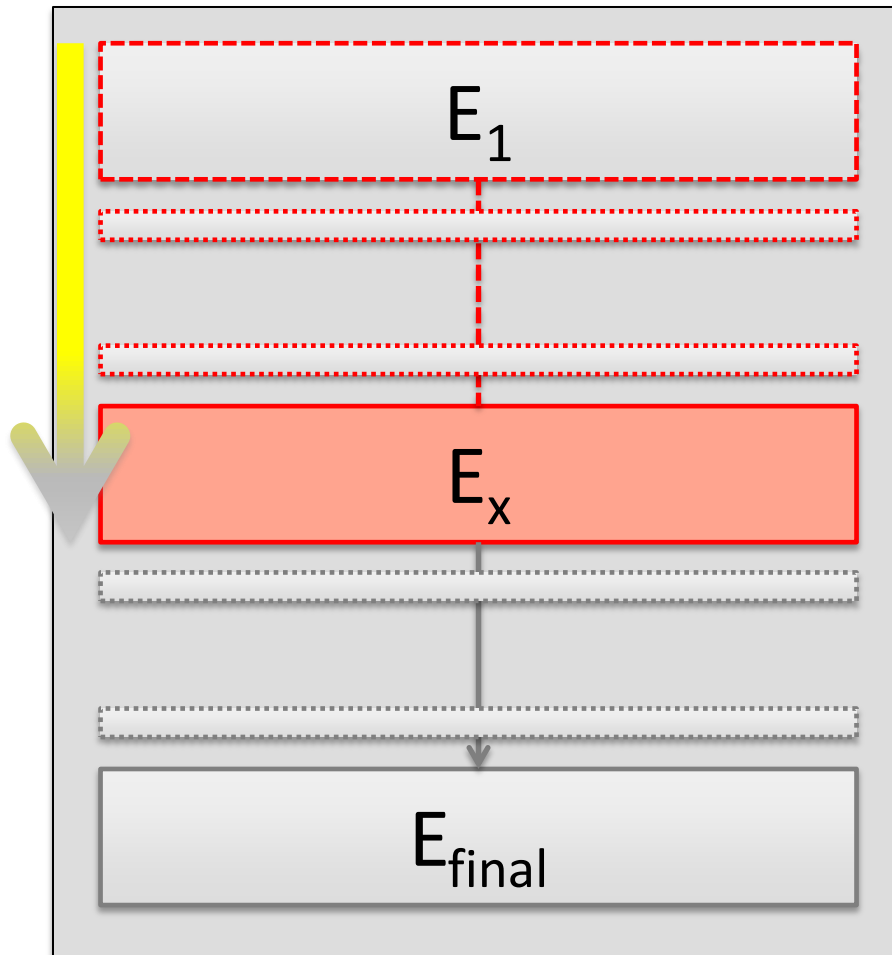
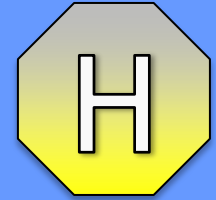


- On ne tient compte que des paramètres lexicométriques de la partie visée E_x

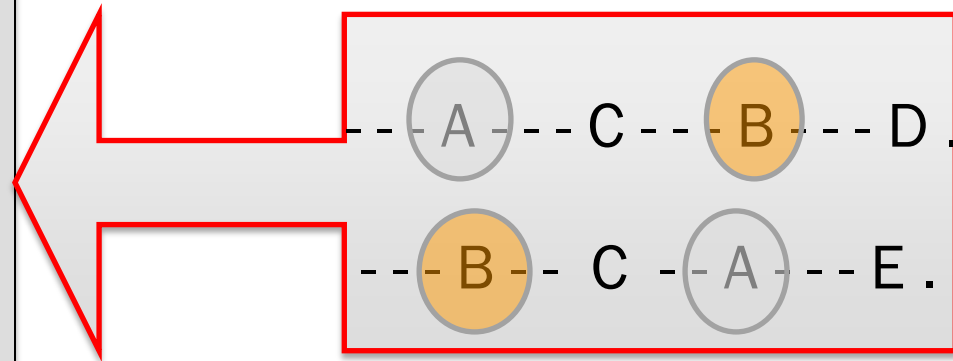


T : Nb occurrence de la partie E_x
 t : nb occurrence dans les contextes de E_x contenant le pôle
 F : fréquence d'une forme sur la partie E_x
 f : fréquence d'une forme dans les contextes de E_x contenant le pôle

Cooccurents : approche chronologique

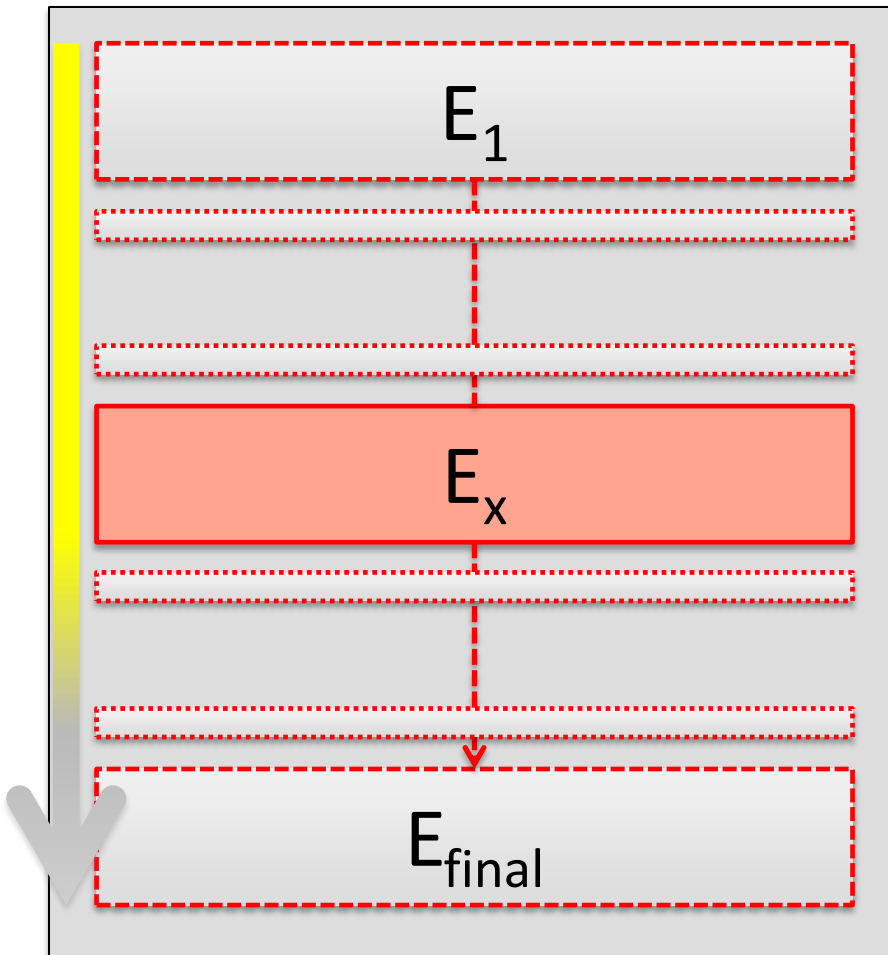
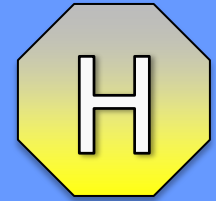


- Le calcul des cooccurents sur l'état E_x garde aussi « en mémoire » le « fond lexical » des parties précédentes

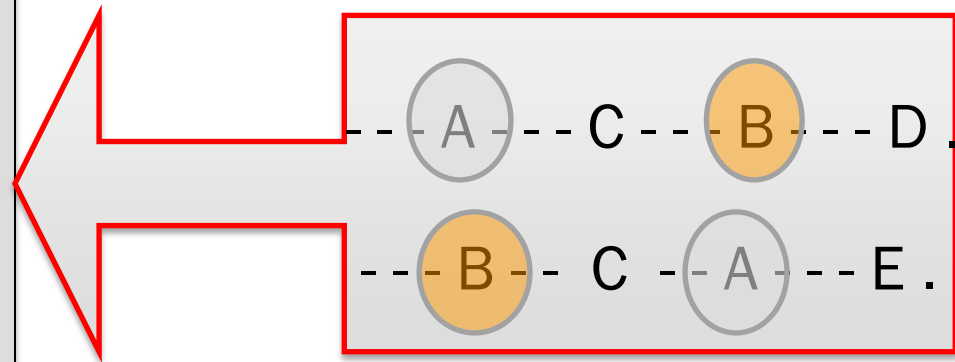


T : Nb occurrence de les parties E_1 à E_x
 t : nb occurrence dans les contextes de E_x contenant le pôle
 F : fréquence d'une forme sur la partie E_x
 f : fréquence d'une forme dans les contextes de E_x contenant le pôle

Cooccurents : approche globale



- Le calcul des cooccurents sur l'état E_x garde aussi « en mémoire » le « fond lexical » de l'intégralité du corpus

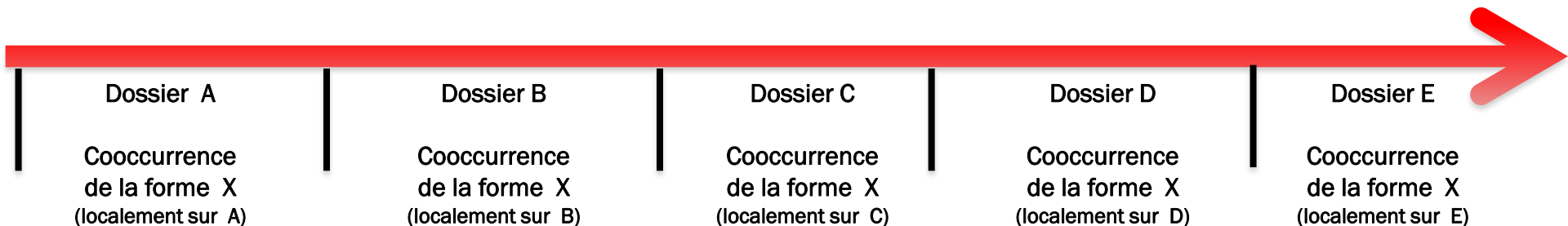


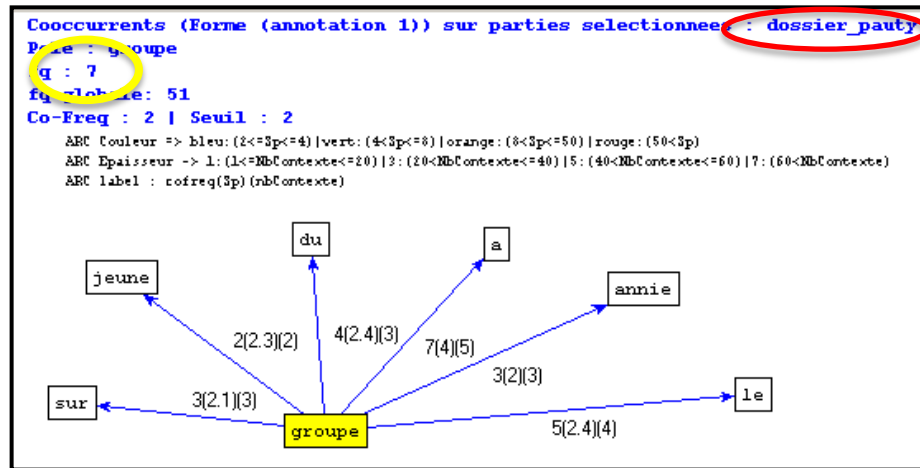
T : Nb occurrence sur le corpus complet
 t : nb occurrence dans les contextes de E_x contenant le pôle
 F : fréquence d'une forme sur le corpus complet
 f : fréquence d'une forme dans les contextes de E_x contenant le pôle

Cooccurrent

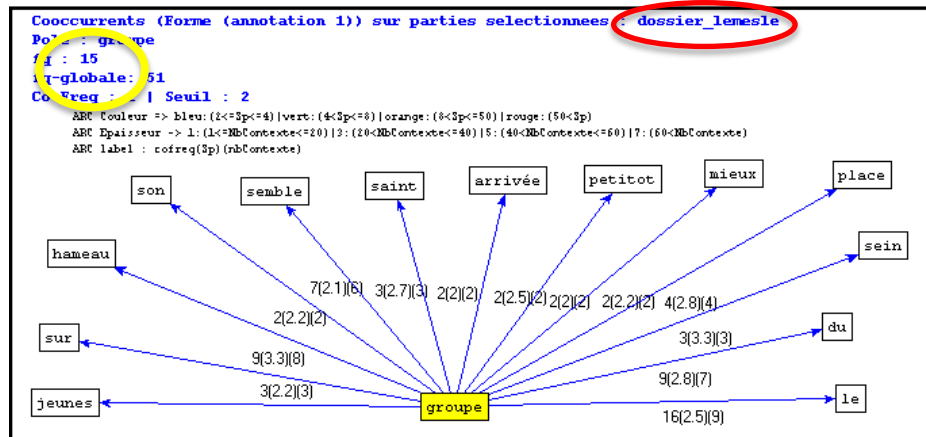


- Contraste des réseaux de cooccurrents d'une forme dans chacun des dossiers
- La partie sélectionnée (le dossier) devient le « corpus de travail » (calcul dynamique des propriétés lexicométriques de cette partie)

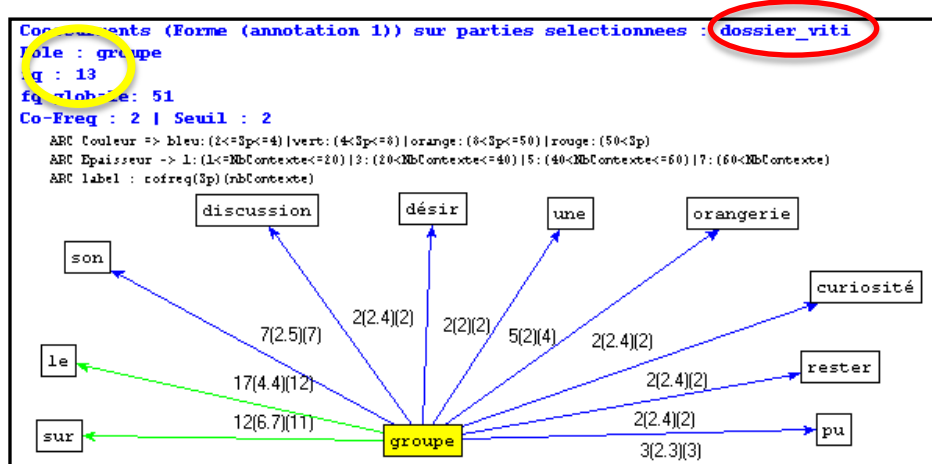




Cooccurents du pôle « groupe » sur une sélection de dossier (dans leur état final)



Pôle	groupe
Co-Freq	2
Seuil	1
Contexte	.





Cooccurents (Forme (annotation 1)) sur parties selectionnees : dossier_viti (option : global)

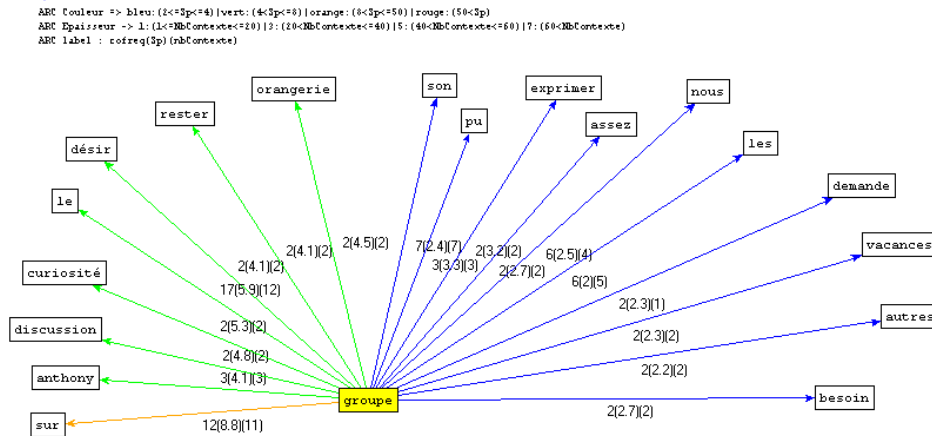
Pole : groupe

fq : 13

fq-globale: 51

Co-Freq : 2 | Seuil : 2

(en tenant compte du corpus complet)



Cooccurents (Forme (annotation 1)) sur parties selectionnees : dossier_pauty (option : global)

Pole : groupe

fq : 7

fq-globale: 51

Co-Freq : 2 | Seuil : 2

(en tenant compte du corpus complet)

ARC Couleur => bleu: (2<=5p<=4) | vert: (4<5p<=8) | orange: (8<5p<=50) | rouge: (50<5p)
ARC Epaisseur -> 1: (1<=NbContexte<=20) | 3: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)
ARC Label : cofreq(5p) (nbContexte)

Cooccurents (Forme (annotation 1)) sur parties selectionnees : dossier_lemesle (option : global)

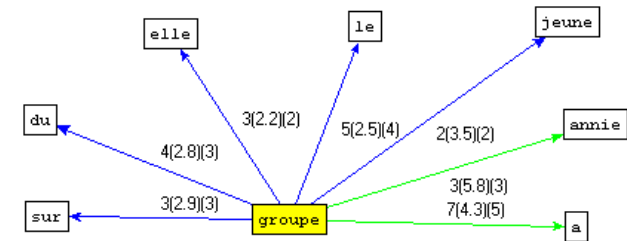
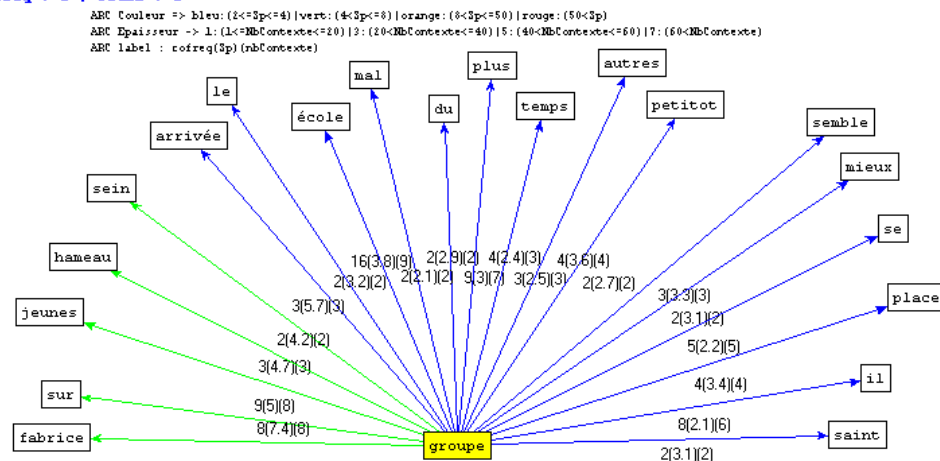
Pole : groupe

fq : 15

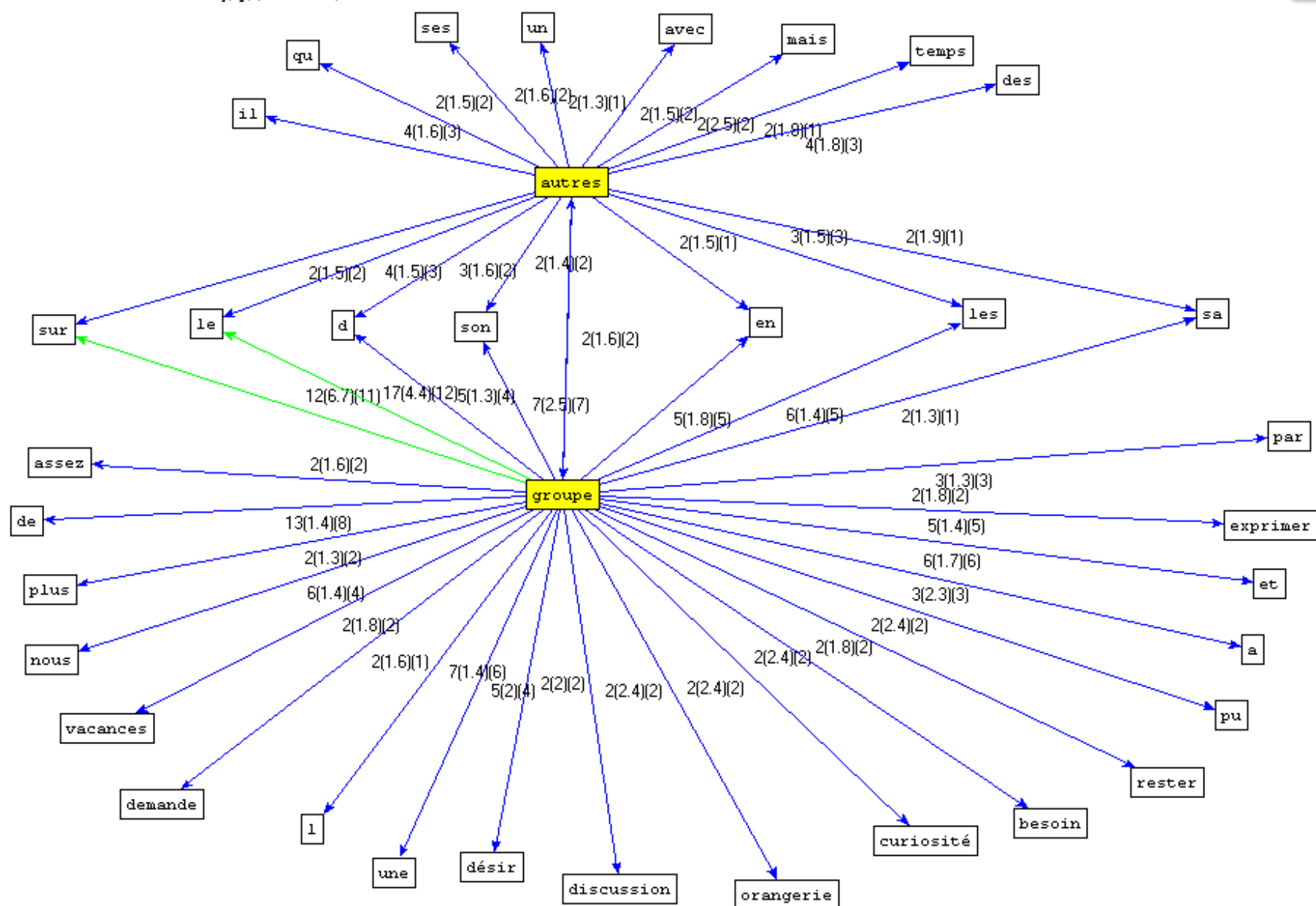
fq-globale: 51

Co-Freq : 2 | Seuil : 2

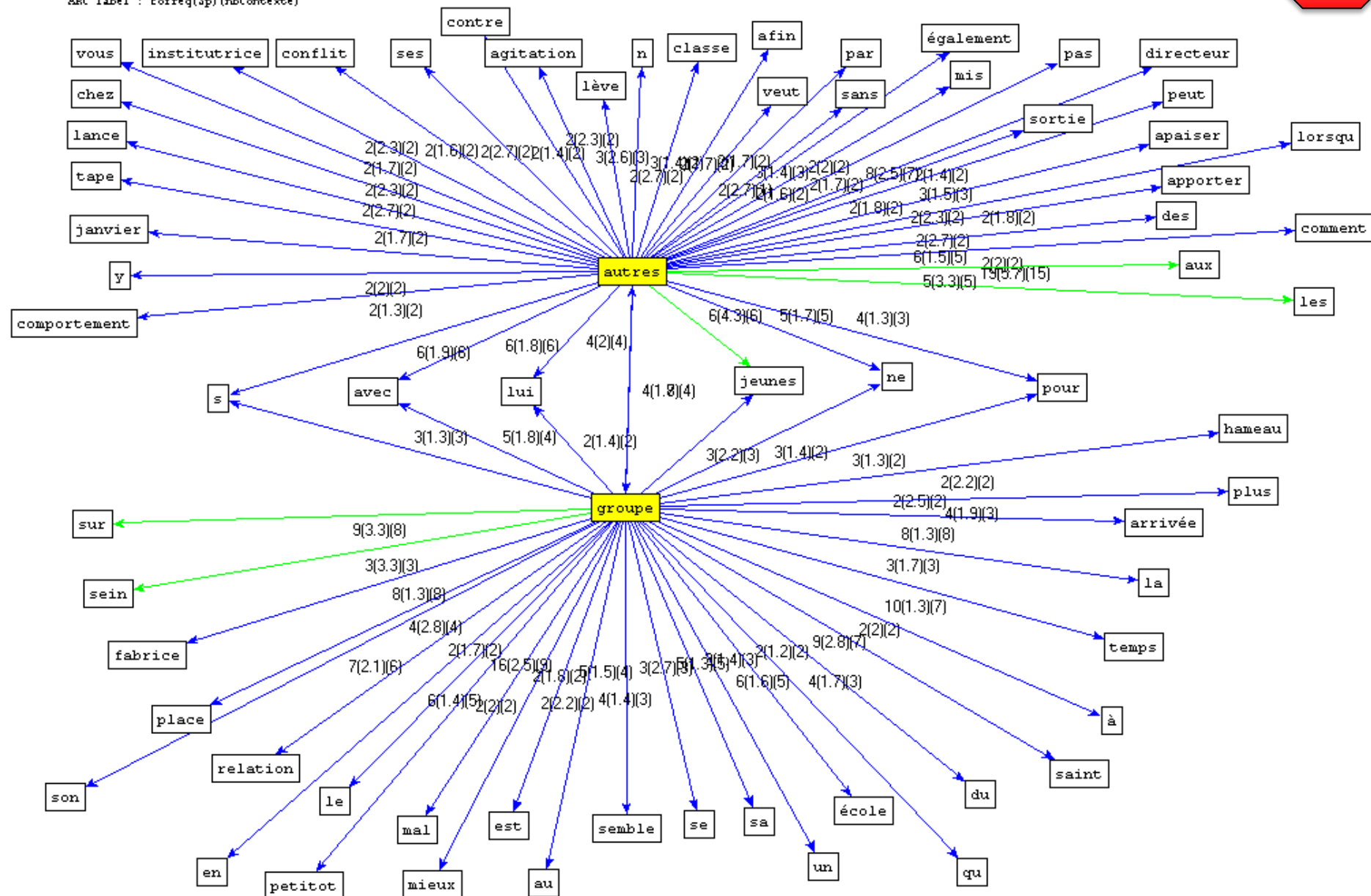
(en tenant compte du corpus complet)



Cooccurents (Forme (annotation 1)) sur parties selectionnees : dossier_viti
Pole : autres|groupe
fq : autres(5)|groupe(13)
fq-globale: autres(55)|groupe(51)
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 : coifreq(Sp)(nbContexte)



```
ARC label : cofreq(Sp)(nbContexte)
```



Cooccurents (Forme (annotation 1)) sur parties selectionnees : dossier_pauty

Pole : autres|groupe

fq : autres(7)|groupe(7)

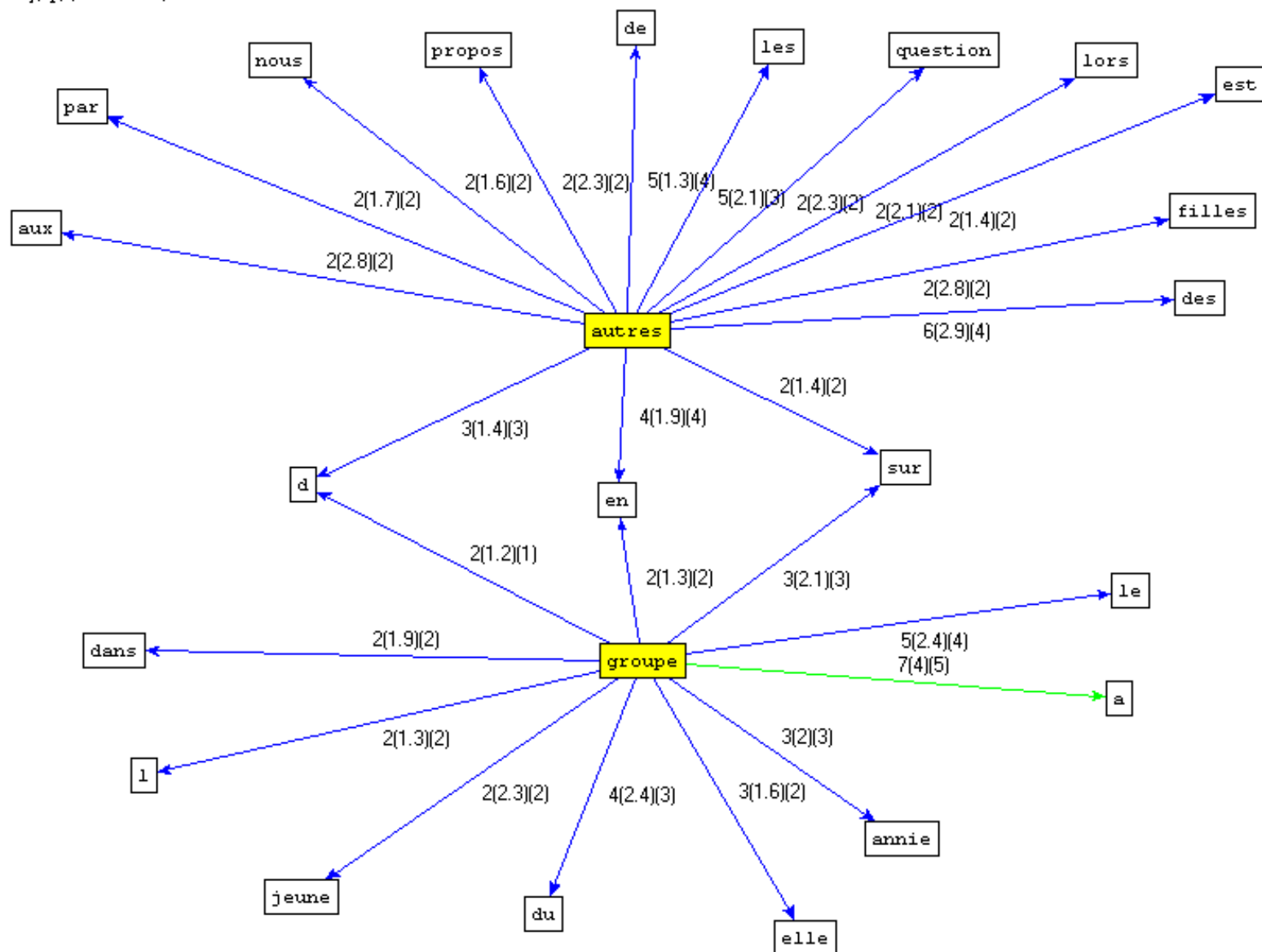
fq-globale: autres(55)|groupe(51)

Co-Freq : 2 | Seuil : 1

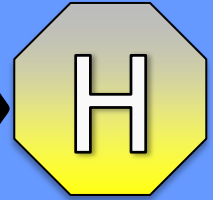
ARC Couleur => bleu: (1<=3p<=3) | vert: (3<3p<=12) | orange: (12<3p<=50) | rouge: (50<3p)

ARC Epaisseur -> 1: (1<=NbContexte<=20) | 3: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)

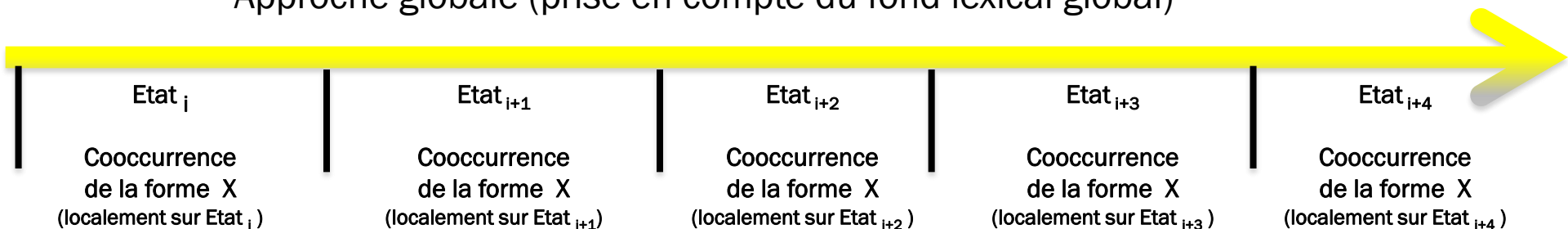
ARC label : cofreq(3p)(nbContexte)



Cooccurents « dynamiques »



- Chantier en cours au sein de CLA²T (Martinez 2003, MacMurray 2012)
- Evolution chronologique du réseau de cooccurents d'une forme dans chacun des états d'un même dossier
 - Les différents états constituent un corpus chronologique
 - Approche locale : La partie sélectionnée (le dossier) peut devenir le « corpus de travail » (calcul dynamique des propriétés lexicométriques de cette partie)
 - Approche chronologique (prise en compte du fond lexical antérieur)
 - Approche globale (prise en compte du fond lexical global)



Cooccurents (Forme (annotation 1)) sur parties selectionnees : **etat_24**

Pôle : groupe

fg : 13

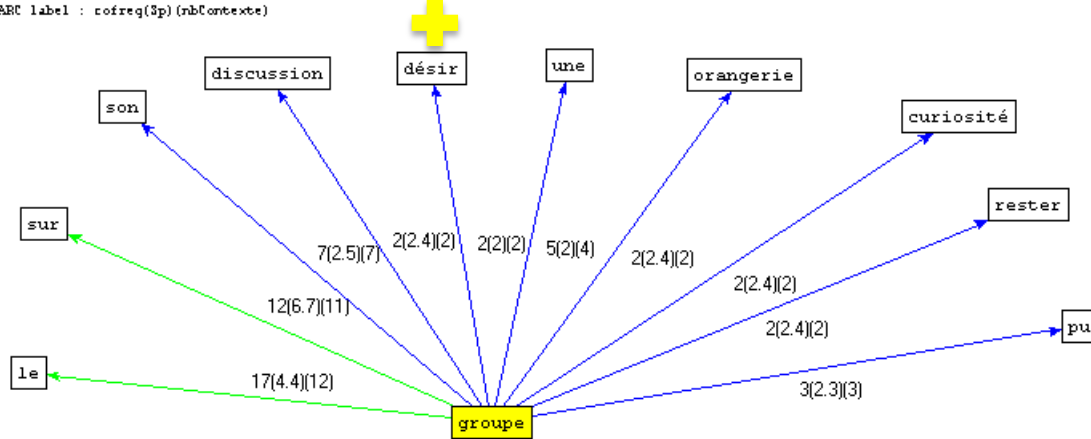
fg-globale : 246

Co-Freq : 2 | Seuil : 2

ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<=Sp<=8) | orange: (8<=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)



Cooccurents du pôle « groupe » sur une sélection d'état de rédaction d'un dossier

Pôle	groupe
Co-Freq	2
Seuil	4
Contexte	.

Cooccurents (Forme (annotation 1)) sur parties selectionnees : **etat_20**

Pôle : groupe

fg : 16

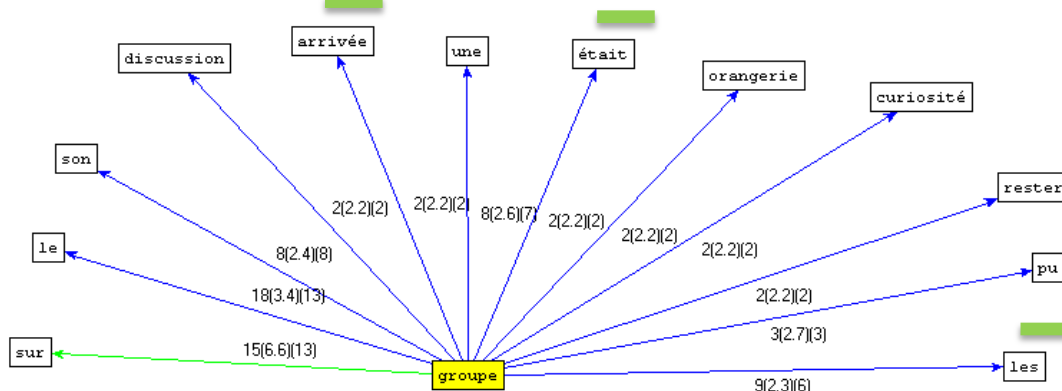
fg-globale : 246

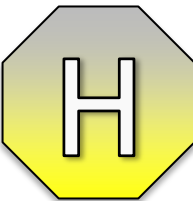
Co-Freq : 2 | Seuil : 2

ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<=Sp<=8) | orange: (8<=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)





local

Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_10 (option : local)

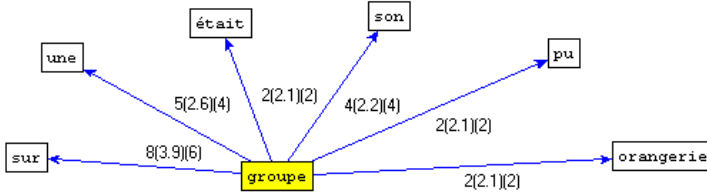
Pole : groupe

fq : 8

fq-globale: 246

Co-Freq : 2 | Seuil : 2

ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<=Sp<=8) | orange: (8<=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)



Etat 10 (en oubliant le reste du corpus)

Etat 15 (en oubliant le reste du corpus)

Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_15 (option : local)

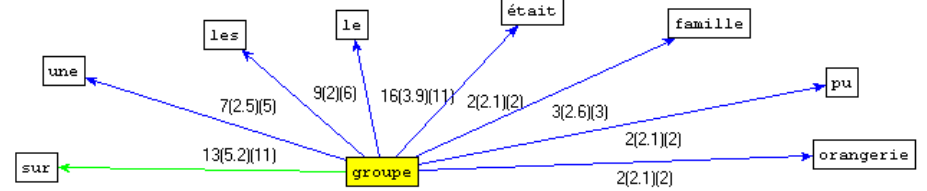
Pole : groupe

fq : 13

fq-globale: 246

Co-Freq : 2 | Seuil : 2

ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<=Sp<=8) | orange: (8<=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)



Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_20 (option : local)

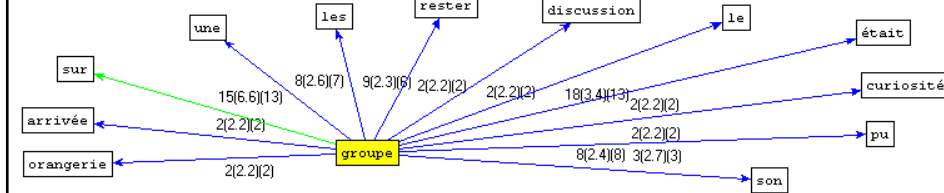
Pole : groupe

fq : 16

fq-globale: 246

Co-Freq : 2 | Seuil : 2

ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<=Sp<=8) | orange: (8<=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)



Etat 20 (en oubliant le reste du corpus)

Etat 24 (en oubliant le reste du corpus)

Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_24 (option : local)

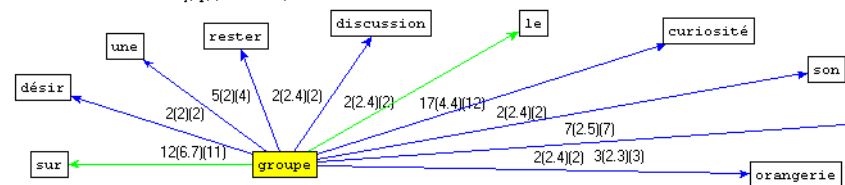
Pole : groupe

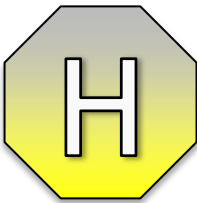
fq : 13

fq-globale: 246

Co-Freq : 2 | Seuil : 2

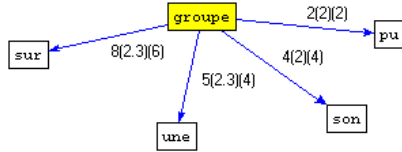
ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<=Sp<=8) | orange: (8<=Sp<=50) | rouge: (50<=Sp)
ARC Epaisseur -> 1: (1<=NbContexte<=20) | 3: (20<=NbContexte<=40) | 5: (40<=NbContexte<=60) | 7: (60<=NbContexte)
ARC label : cofreq(Sp) (nbContexte)





chrono

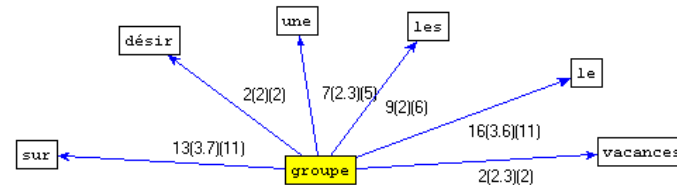
Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_10 (option : chrono)
Pole : groupe
fq : 8
fq-globale: 246
Co-Freq : 2 | Seuil : 2
ARC Couleur => bleu: (2<=5p<=4) | vert: (4<5p<=8) | orange: (8<5p<=50) | rouge: (50<5p)
ARC Epaisseur -> 1: (1<=NbContexte<=20) | 2: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)
ARC label : cofreq(5p) (nbContexte)



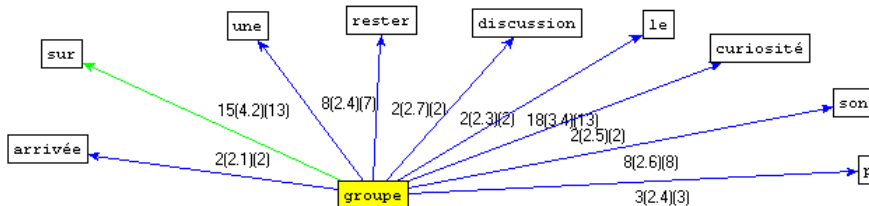
Etat 10 (en tenant compte des états précédents)

Etat 15 (en tenant compte des états précédents)

Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_15 (option : chrono)
Pole : groupe
fq : 13
fq-globale: 246
Co-Freq : 2 | Seuil : 2
ARC Couleur => bleu: (2<=5p<=4) | vert: (4<5p<=8) | orange: (8<5p<=50) | rouge: (50<5p)
ARC Epaisseur -> 1: (1<=NbContexte<=20) | 3: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)
ARC label : cofreq(5p) (nbContexte)



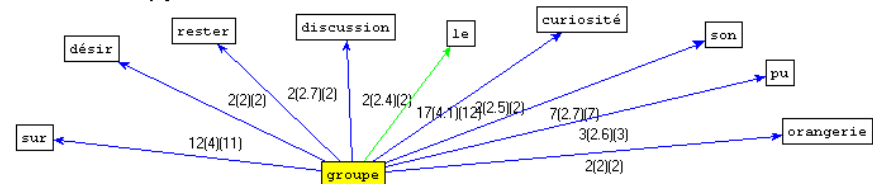
Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_20 (option : chrono)
Pole : groupe
fq : 16
fq-globale: 246
Co-Freq : 2 | Seuil : 2
ARC Couleur => bleu: (2<=5p<=4) | vert: (4<5p<=8) | orange: (8<5p<=50) | rouge: (50<5p)
ARC Epaisseur -> 1: (1<=NbContexte<=20) | 3: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)
ARC label : cofreq(5p) (nbContexte)

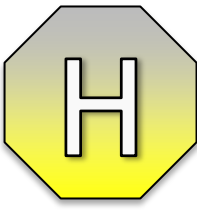


Etat 20 (en tenant compte des états précédents)

Etat 24 (en tenant compte des états précédents)

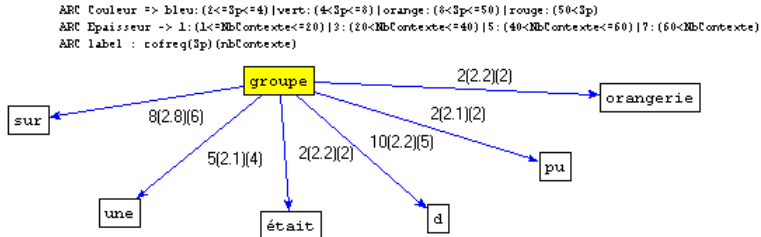
Cooccurrents (Forme (annotation 1)) sur parties selectionnees : etat_24 (option : chrono)
Pole : groupe
fq : 13
fq-globale: 246
Co-Freq : 2 | Seuil : 2
ARC Couleur => bleu: (2<=5p<=4) | vert: (4<5p<=8) | orange: (8<5p<=50) | rouge: (50<5p)
ARC Epaisseur -> 1: (1<=NbContexte<=20) | 3: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)
ARC label : cofreq(5p) (nbContexte)





global

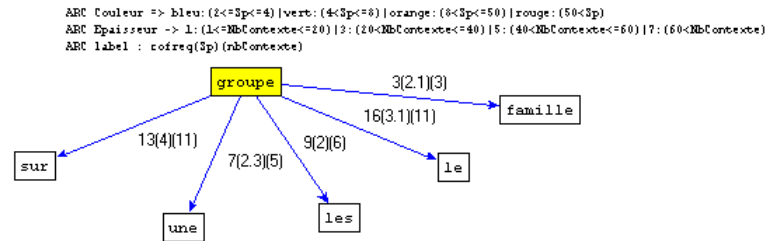
Cooccurents (Forme (annotation 1)) sur parties selectionnees : etat_10 (option : global)
Pole : groupe
fq : 8
fq-globale: 246
Co-Freq : 2 | Seuil : 2



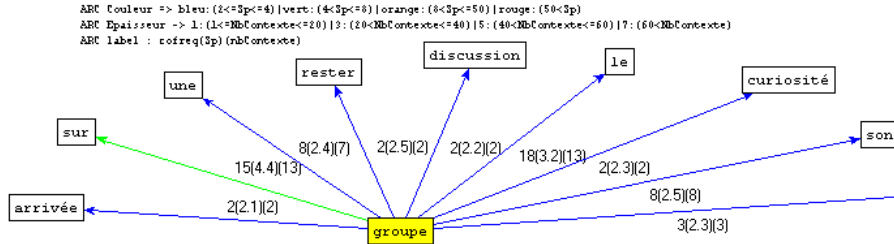
Etat 10 (en tenant compte du corpus complet)

Etat 15 (en tenant compte du corpus complet)

Cooccurents (Forme (annotation 1)) sur parties selectionnees : etat_15 (option : global)
Pole : groupe
fq : 13
fq-globale: 246
Co-Freq : 2 | Seuil : 2



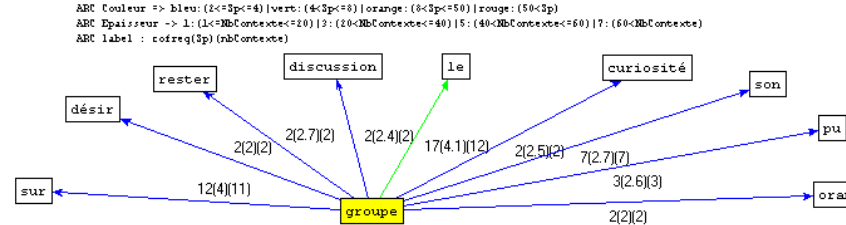
Cooccurents (Forme (annotation 1)) sur parties selectionnees : etat_20 (option : global)
Pole : groupe
fq : 16
fq-globale: 246
Co-Freq : 2 | Seuil : 2



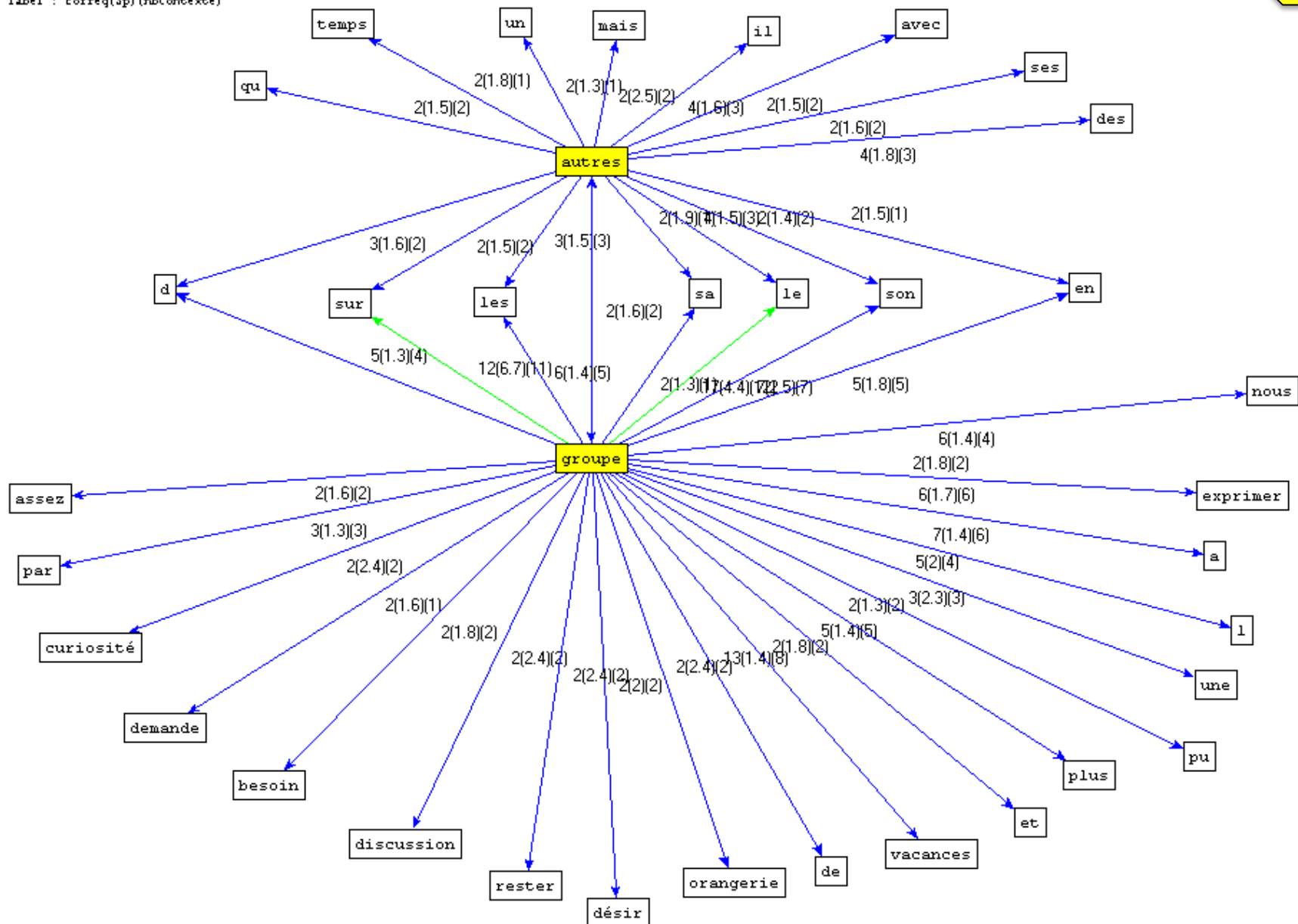
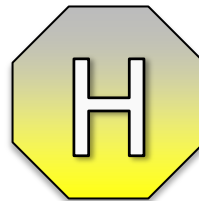
Etat 20 (en tenant compte du corpus complet)

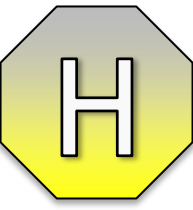
Etat 24 (en tenant compte du corpus complet)

Cooccurents (Forme (annotation 1)) sur parties selectionnees : etat_24 (option : global)
Pole : groupe
fq : 13
fq-globale: 246
Co-Freq : 2 | Seuil : 2



Cooccurents (Forme (annotation 1)) sur parties selectionnees : etat_24
 Pole : groupe|autres
 fq : autres(5)|groupe(13)
 fq-globale: autres(58)|groupe(246)
 Co-Freq : 2 | Seuil : 1
 ARC couleur => bleu: (1<=3p<=3) |vert: (3<3p<=12) |orange: (12<3p<=50) |rouge: (50<3p)
 ARC Epaisseur -> 1: (1<=NbContexte<=20) |3: (20<NbContexte<=40) |5: (40<NbContexte<=60) |7: (60<NbContexte)
 ARC label : cofreq(3p)(nbContexte)





Cooccurents (Forme (annotation 1)) sur parties selectionnees : etat_24

Pole : ^le groupe\$|^les autres\$

fg : le groupe(10)|les autres(2)

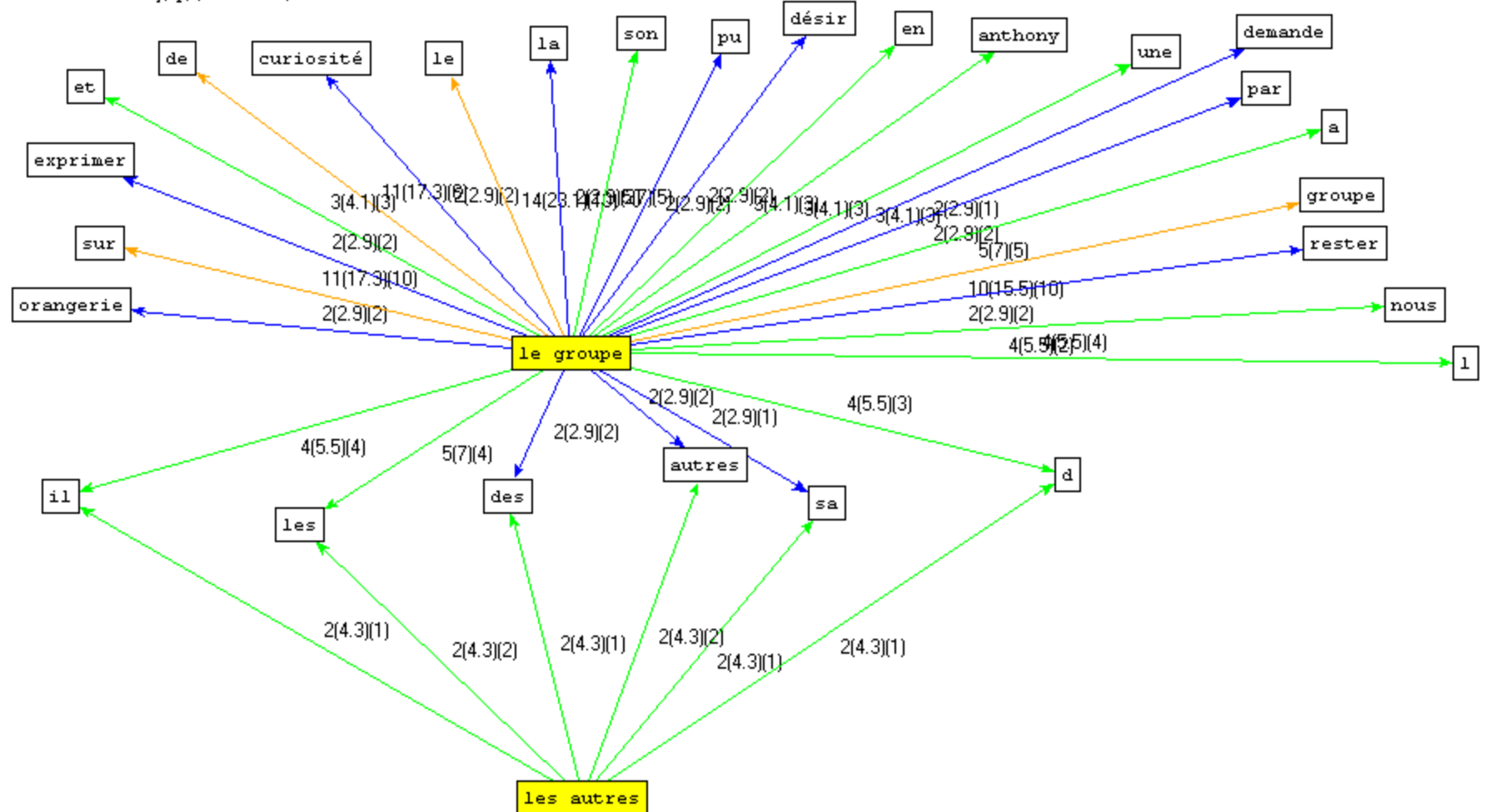
fg-globale: le groupe(10)|les autres(2)

Co-Freq : 2 | Seuil : 2

ABC Couleur => bleu: (2<Sp<=4)|vert: (4<Sp<=8)|orange: (8<Sp<=50)|rouge: (50<Sp)

ABC Epaisseur -> 1: (1<NbContexte<=20)|3: (20<NbContexte<=40)|5: (40<NbContexte<=60)|7: (60<NbContexte)

ABC label : nofreq(Sp)(nbContexte)

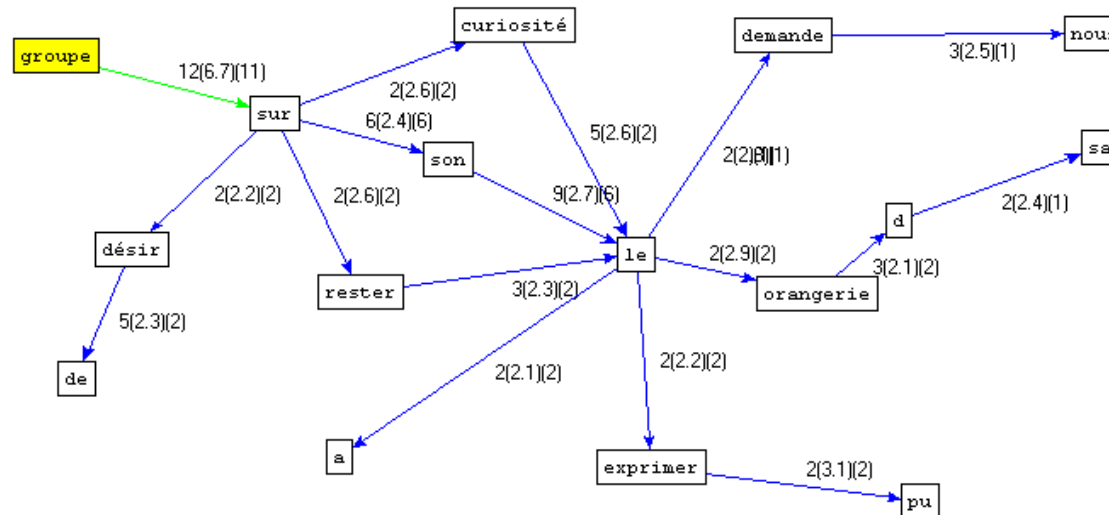


Poly-cooccurents

- Le terme *poly-cooccurrence* désigne les attractions lexicales au-delà de la cooccurrence binaire [Martinez, 2006]
- On calcule pour le pôle A les cooccurents spécifiques A_1 , A_2 et A_3
 - Dans leurs contextes communs, on calcule pour les pôles $A+A_1$ les cooccurents spécifiques A_{11} et A_{12}
 - Les pôles $A+A_1+A_{11}$ ont pour cooccurent spécifique A_{111}
 - On examine les cooccurents de $A+A_1+A_{11}+A_{111}$:
Les pôles $A+A_1+A_{11}+A_{111}$ n'ont pas de cooccurent spécifique : l'exploration s'interrompt pour ce chemin
 - Les pôles $A+A_1+A_{12}$ ont pour cooccurent spécifique A_{121}
 - On examine les cooccurents de $A+A_1+A_{11}+A_{121}$:
Etc.
 - Dans leurs contextes communs, on calcule pour les pôles $A+A_2$ les cooccurents spécifiques A_{21} et A_{22}
 - Etc.
- Durant l'exploration, différents filtrages conditionnent l'épuisement des explorations contextuelles et réduisent le bruit dans les résultats pour privilégier l'information la plus spécifique : seuils maximaux de fréquence et de spécificité du cooccurent

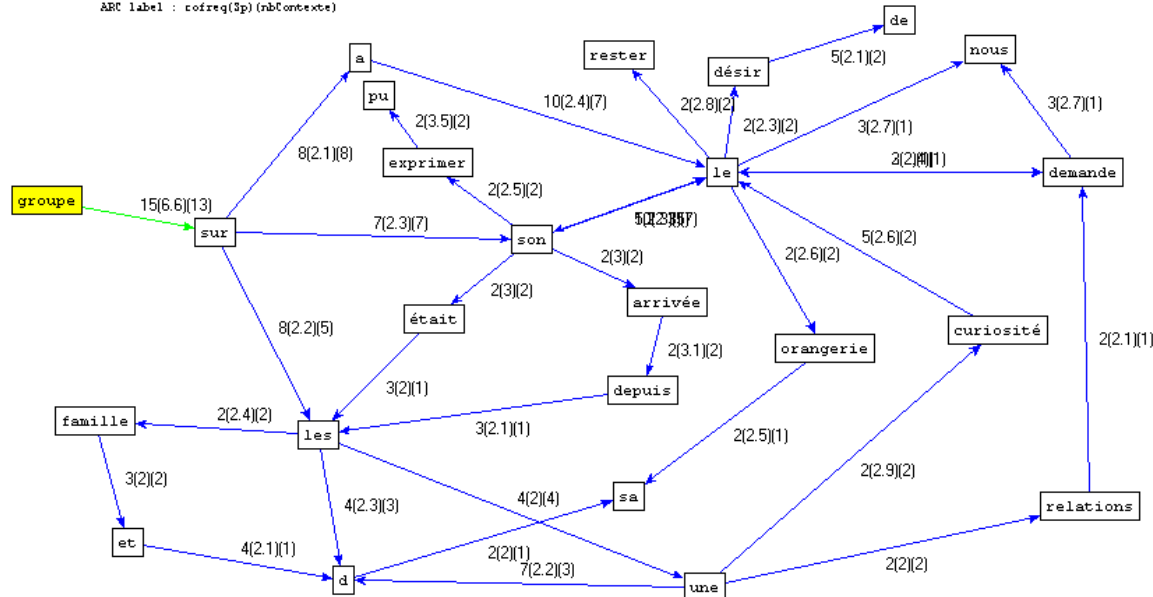
Co-Freq : 2 | Seuil : 2

ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<Sp<=8) | orange: (8<Sp<=50) | rouge: (50<Sp)
 ARC Epaisseur -> 1: (1<=NbContexte<=20) | 2: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)
 ARC label : cofreq(Sp) (nbContexte)



Co-Freq : 2 | Seuil : 2

ARC Couleur => bleu: (2<=Sp<=4) | vert: (4<Sp<=8) | orange: (8<Sp<=50) | rouge: (50<Sp)
 ARC Epaisseur -> 1: (1<=NbContexte<=20) | 2: (20<NbContexte<=40) | 5: (40<NbContexte<=60) | 7: (60<NbContexte)
 ARC label : cofreq(Sp) (nbContexte)



Le Trameur

<http://tal.univ-paris3.fr/trameur>

Nous appelons **trameur** l'outil informatique qui permet de construire une ressource textométrique trame/cadre, à partir d'un ensemble de textes.

→ La partie **Trame** de la ressource textuelle produite par le trameur est constituée par la suite des items isolés lors de l'opération de segmentation.

→ La partie **Cadre** rassemble les données relatives aux différents découpages réalisés sur le corpus ainsi que les annotations projetées sur les contenants ou les contenus.

La transmission d'une ressource textuelle constituée sous la forme **Trame/Cadre** constitue une solution suffisante pour servir de base à toute exploration textométrique.

[Söze-Duval, 2008], Keyser Söze-Duval. *Pour une textométrie opérationnelle*



Origines

- Développement d'outils pour les linguistes
 - textométrie
- CLA2T André Salem
 - Lexico
 - [Söze-Duval, 2008], Keyser Söze-Duval. *Pour une textométrie opérationnelle*
- Collaborations/discussions
 - Réseaux, partenaires

Le texte = une trame [+ un cadre]

A partir d'un texte segmenté, la numérotation des items découpés dans le texte de départ permet de constituer un système de coordonnées sur le texte dans lequel chaque item est repéré par son numéro d'ordre. Nous appelons ce système de coordonnées sur la séquence textuelle : la **Trame** textométrique. Ce même système de coordonnées permet de définir et de localiser, au sein du corpus, des zones textuelles (zones formées par une suite d'items consécutifs, entre la position x_1 la position x_2 , réunion d'un certain nombre de zones de ce type, etc.). La définition d'une **Trame** textométrique sur un corpus de textes permet de décrire les systèmes de zones qui correspondent aux contenants de l'analyse textométrique (parties, paragraphes, phrases, sections, chapitres etc.). On peut rassembler les descriptions relatives aux systèmes de contenants dans une structure de données particulière le **Cadre** textométrique. (Söze-Duval 2008)



Le texte : une suite de position

- Un texte : séquence ordonnée d'items
 - Une liste de positions
 - A chaque position un item
 - 2 types d'item : les délimiteurs et les non-délimiteurs
 - Un item = une pile d'annotations
 - forme, lemme, catégorie etc.

Forme		Le		donneur		du		val	RC		C	...
Positions	1	2	3	4	5	6	7	8	9	10	11	12

Le texte : une suite de positions organisées

<STRUCTURE="TITRE"> Le dormeur du val

<STRUCTURE="TEXTEPOEME">

[LIGNE="VERS1"> C' est un trou de verdure où chante une rivière,

[LIGNE="VERS2"> Accrochant follement aux herbes des haillons

[LIGNE="VERS3"> D' argent où le soleil ; de la montagne fière,

Annotation importée n°1	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	
Catégorie		Cat(Le)		Cat(dormeur)		Cat(du)		Cat(val)	...
Lemme		Lemme(Le)		Lemme(dormeur)		Lemme(du)		Lemme(val)	...
Forme		Le		dormeur		du		val	...
positions	1	2	3	4	5	6	7	8	...

[LIGNE="VERS10"> Sourirait un enfant malade, il fait un somme :

[LIGNE="VERS11"> Nature, berce-le chaudement : il a froid.

[LIGNE="VERS12"> Les parfums ne font pas frissonner sa narine ;

[LIGNE="VERS13"> Il dort dans le soleil, la main sur sa poitrine,

[LIGNE="VERS14"> Tranquille. Il a deux trous rouges au côté droit.

<STRUCTURE="AUTEUR">Arthur Rimbaud

Le texte : une trame et un cadre

- La **Trame** = Le texte comme suite de positions
- Le **Cadre** = Ensembles de positions structurant le texte initial en parties
- Sélection : ensemble de positions de la trame
 - Le **Cadre** est une instance particulière de « sélection »

```

<STRUCTURE="TITRE"> Le dormeur du val
<STRUCTURE="TEXTEPOEME">
<LIGNE="VERS1"> C' est un trou de verdure où chante une rivière,
<LIGNE="VERS2"> Accrochant follement aux herbes des haillons
<LIGNE="VERS3"> D' argent ; où le soleil ; de la montagne fière,
<LIGNE="VERS4"> Luit : c' est un petit val qui mousse de rayons.

```

La *Trame* du texte correspond à toutes les positions des **items** (délimiteurs et formes graphiques). Sur cet exemple :

En position 1, le caractère « espace » (considéré comme un délimiteur), en position 2, la forme graphique « Le », en position 3, le caractère « espace » etc.

Le *Trameur* produira une segmentation que l'on peut schématiser ainsi sur le début du texte :

Forme		Le		dormeur		du		val	RC		C	...
Positions	1	2	3	4	5	6	7	8	9	10	11	...

(RC = retour à la ligne)

Le *Cadre* du texte est matérialisé ici par des balises indiquant le début d'une section (sous la forme <partie=valeur>). Ce balisage traduit une structuration en parties du texte (1 titre et 1 poème contenant des vers). Plus précisément, la partie STRUCTURE=TITRE recouvre les items de la position 1 à la position 8 et la partie STRUCTURE=POEME couvre les items à partir de la position 9 jusqu'à la dernière.

Le Trameur-tk

- Le *Trameur* : Programme de génération puis de gestion de la **Trame** et du **Cadre** d'un texte (*le métier textométrique*) pour des opérations textométriques.
 - Le *Trameur* intègre le programme *treetagger* : système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation.
 - <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

**Chargement de la base :
La Trame + Le Cadre****Options chargement :**☐ Etiquetage ☐ Traceur ☐ Dics**Exporter la base :
La Trame + Le Cadre****Représentation du cadre :**

(choisir la/les partie(s))

**Parties**

(Ctrl-Clic : désélection partie)

**Recherche Forme dans
le Cadre (regex) :**

...

(Ecrire motif supra puis <Entrée>)

regex

CADRE ACTIF : Clic sur partie : les formes | Clic-droit sur partie : les lemmes | Control+Clic sur partie : vocabulaire spécifique

ARBRE XML : Clic sur noeud : les formes

Nb L. Shift-Clic : sélection | Clic-droit : édition | Ctrl-Clic : noeud | 2-Clic : graphe

Partie : ?

Annotation : 1

Aperçu : 50



1

DÉMO

Chargement d'un texte (avec étiquetage)

Exploration du *Cadre* et de la *Trame*

- le cadre comme accès à la *Trame*
- édition/modification de la *Trame*

Principales fonctionnalités

- concordance, ventilation, carte des sections, segments répétés
- cooccurrents, polycooccurrents
- extraction de patron
- spécificités, AFC



<http://tal.univ-paris3.fr/trameur/demo/trameur-demo.htm>

CONCLUSION



MERCI

Des questions ?  sergefleury@gmail.com