
Explorer un corpus annoté avec Le Trameur / iTrameur

Paris, 15/11/2018

Serge Fleury

Université Sorbonne nouvelle, Paris 3

CLESTHIA / EA 73455

Le Trameur

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

Windows / MacOS X

iTrameur

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

En ligne dans un « bon navigateur »  

Contact : sergefleury@gmail.com

Le Trameur (ou *iTrameur*) est un programme d'analyse comportant de nombreuses fonctionnalités pour l'analyse automatique, statistique et documentaire de textes en vue de leur profilage sémantique, thématique et de leur interprétation.

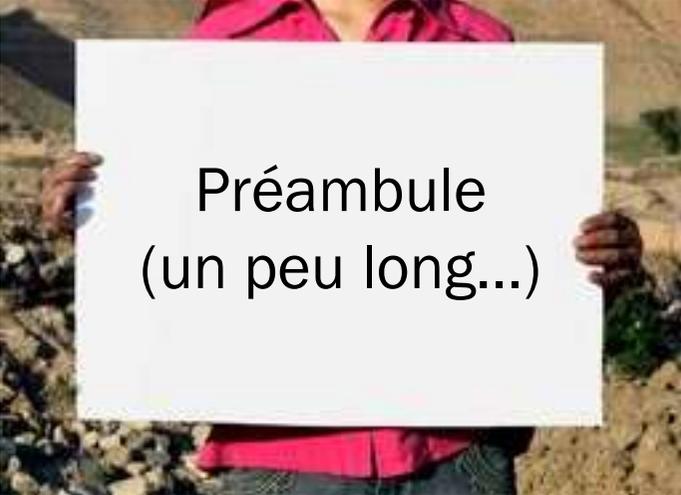
Ce logiciel est à l'origine un outil de **textométrie** : il intègre les fonctionnalités classiques de ce type d'outils dans ce domaine.

Il dispose aussi des fonctionnalités particulières qui permettent d'annoter dynamiquement des corpus ou d'explorer des ressources richement annotées (**treebanks** monolingues/multilingues ou des alignements).

Ressources pour l'atelier

■ En ligne ici :

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

A person wearing a bright pink jacket is holding a white rectangular sign in front of their chest. The sign has black text on it. The background is a natural, outdoor setting with some rocks and dry ground.

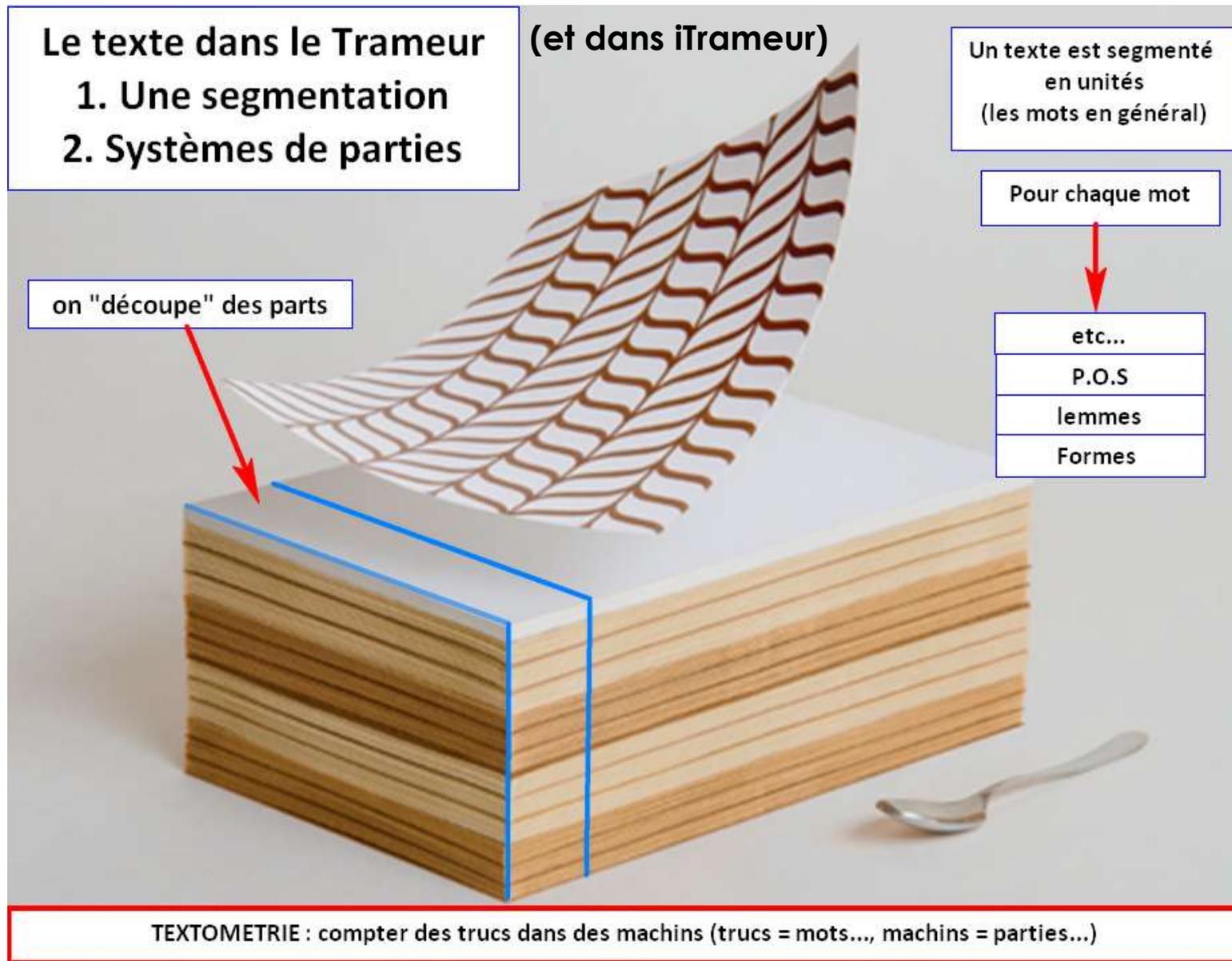
Préambule
(un peu long...)



Explorer un corpus annoté...

- Avant d'explorer un tel corpus, il faut en général :
 - Commencer par le « construire » (bon courage !)
 - On fera ce travail au S2 dans le cours projet
 - Ou le « récupérer » quelque part (coup de chance !)
 - Exemple pour SF : « le projet Rhapsodie »
- Construire un corpus annoté
 - Outils pour annoter dynamiquement un corpus
 - Gestion des « articulations » entre les informations intégrées (quelle architecture pour les données)
- Dans les 2 cas, adapter les données annotées aux outils d'analyse...

Explorer un corpus annoté... == « Explorer (i.e dévorez) » un « mille-feuilles »



Explorer un corpus annoté - iTrameur

□ Avec iTrameur

- On peut Créer une nouvelle base (sans annotations)

- Travail uniquement sur les formes graphiques

- (à ce jour) pas d'annotation dynamique

- Possibilité d'exporter et d'annoter en POS

- On peut Importer une base annotée

- Annotation dynamique (uniquement par fusion)

□ Avec Le Trameur

- On peut Créer une nouvelle base avec ou sans annotations en POS

- Annotation dynamique (fusion/ajout/modification...)

- On peut Importer une base annotée

- Annotation dynamique (fusion/ajout/modification...)

Explorer une nouvelle base - iTrameur

- ❑ **Chargement d'une base « brute »** (sans annotations)
 - ❑ Formats de fichiers
 - ❑ TXT brut, UTF8
 - ❑ partitionnement/balisage, découpage en sections
 - ❑ DEMO : « chargement d'un texte brut partitionné... »
- ❑ Exploration et calculs
- ❑ Export de la base → annotation en POS
 - ❑ Le format d'export est celui utilisé pour importer une base annotée...

Explorer un corpus annoté - iTrameur

□ Avec iTrameur

□ Importer une base annotée

□ Format de base (tabulaire)

□ Chaque item est décrit par ses annotations (une par colonne)

□ 2 types d'annotation...

□ Délimiteur de section (carte des sections)

□ Exploration et calculs

□ Fusion d'annotations

□ DEMO 1 : ressources en ligne (cf onglet AIDE)

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

Segmentation : liste de positions

1	forme	euh	I	euh	ROOT
2	delim		BLANK	BLANK	-
3	forme	bon	I	bon	ROOT
4	delim		BLANK	BLANK	-
5	forme	F pour P	Pre	L pour A	AD(33)
6	delim		BLANK	BLANK	-
7	forme	O aller O	V	E aller N	DEP(5)
8	delim		BLANK	BLANK	-
9	forme	R du S	Pre+D	de+le	OBJ(7)
10	delim		BLANK	BLANK	-
11	forme	M CRDT	N	CRDT	DEP(9)
12	delim		BLANK	BLANK	-
13	forme	E à	Pre	à	OBL(7)
14	delim		BLANK	BLANK	-
15	forme	E la	D	E le T	DEP(17)
16	delim		BLANK	BLANK	-
17	forme	gare	N	gare A	DEP(13)
18	delim		BLANK	BLANK	-
19	forme	euh	I	euh	ROOT
20	delim		BLANK	BLANK	T
21	forme	de	Pre	de	DEP(17)
22	delim		BLANK	BLANK	I
23	forme	Grenoble	N	Grenoble	DEP(21)
24	delim		BLANK	BLANK	-
25	forme	je	Cl	je O	PARA_DISFL(31)
26	delim		BLANK	BLANK	-
27	forme	euh	I	euh	ROOT
28	delim		BLANK	BLANK	N
29	forme	ben	I	ben	ROOT
30	delim		BLANK	BLANK	-
31	forme	je	Cl	je	SUB(33)
32	delim		BLANK	BLANK	-
33	forme	sors	V	sortir	ROOT
34	delim		BLANK	BLANK	-
35	forme	déjà	Adv	déjà	AD(33)
36	delim		BLANK	BLANK	-
37	forme	du	Pre+D	de+le	OBL(33)
38	delim		BLANK	BLANK	-
39	forme	CRDT	N	CRDT	DEP(37)
40	delim		BLANK	BLANK	-
41	delim	£	DELIM	£	-

TRAME

Liste de positions annotées

1/2

PARTITION: SUBGENRE	PARTIE: NARRATIVE	DEBUT: 2087	FIN: 4743
PARTITION: SUBGENRE	PARTIE: NARRATIVE	DEBUT: 56825	FIN: 59388
PARTITION: SUBGENRE	PARTIE: NARRATIVE	DEBUT: 70453	FIN: 73631
PARTITION: SUBGENRE	PARTIE: NARRATIVE	DEBUT: 84514	FIN: 89776
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 6407	FIN: 10350
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 13792	FIN: 16417
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 16822	FIN: 18008
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 18222	FIN: 18707
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 22415	FIN: 23277
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 23278	FIN: 24114
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 24406	FIN: 27618
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 32342	FIN: 32799
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 35404	FIN: 38698
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 38699	FIN: 39255
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 39256	FIN: 39904
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 39905	FIN: 40218
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 43916	FIN: 46353
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 46354	FIN: 47779
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 48365	FIN: 52390
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 59389	FIN: 60657
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 63650	FIN: 64113
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 65037	FIN: 67076
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 73632	FIN: 76113
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 83528	FIN: 84513
PARTITION: SUBGENRE	PARTIE: DESCRIPTION	DEBUT: 91379	FIN: 93955
PARTITION: SUBGENRE	PARTIE: ORATORY	DEBUT: 1248	FIN: 1866
PARTITION: SUBGENRE	PARTIE: ORATORY	DEBUT: 10351	FIN: 13791
PARTITION: SUBGENRE	PARTIE: ORATORY	DEBUT: 81508	FIN: 83527
PARTITION: SUBGENRE	PARTIE: ORATORY	DEBUT: 89942	FIN: 91378
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 4744	FIN: 6406
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 18708	FIN: 21988
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 27619	FIN: 30046
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 32800	FIN: 35403
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 40219	FIN: 43915
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 53688	FIN: 56824
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 60807	FIN: 63649
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 67459	FIN: 70452
PARTITION: SUBGENRE	PARTIE: ARGUMENTATION	DEBUT: 76114	FIN: 81507

CADRE

Ensemble de couples de positions sur la Trame structurant le texte en système de parties

Description des systèmes de partie

Explorer un corpus annoté – Le Trameur

□ Avec le Trameur

□ **Chargement d'une base « brute »** (sans annotations)

- Annotation dynamique (et correction)

- Fusion d'annotation

- Exploration et calculs

- DEMO2 : « chargement d'un texte brut partitionné... »

□ **Importer une base annotée**

- Format de base (XML) : similaire au format d'export

- Fusion d'annotation

- Exploration et calculs

- DEMO3 : « importation d'un treebank » (puis d'un bitexte aligné et annoté)

Entête d'une base textométrique

```
<?xml version="1.0" encoding="utf-8"?>
<baselexicometrie>
<teiheader>
<fileDesc>
<titleStmt>
<title>Modelisation XML de la base textometrique (le metier = le cadre et l
a trame), importable et echangeable avec d'autres plateformes</title>
</titleStmt>
<publicationStmt>
<p>Samedi 09 Avril 2016
18:23:42
    Ce document n'est pas encore publie.</p>
</publicationStmt>
<sourceDesc>
<title>
<br>Le Trameur 12.42. Samedi 09 Avril 2016
18:23:42 </br></title>
<content><l><br>Fichier traite </br><br> corpus-grece-utf8.txt</br></l>
<l><br>Encodage </br><br> utf-8</br></l>
<l><br>Nombre d'items </br><br> 963496</br></l>
<l><br>Nombre de délimiteurs </br><br> 520274</br></l>
<l><br>Nombre d'occurrences de forme </br><br> 443222</br></l>
<l><br>Nombre de formes </br><br> 34296</br></l>
<l><br>Nombre d'hapax </br><br> 14473</br></l>
<l><br>Fréquence maximale </br><br> 17299</br></l>
<l><br>Forme maximale </br><br> de</br></l>
<l><br>Délimiteurs </br><br><![CDATA[. , ; ! ? / _ - " ' ( ) [ ] { } $ % | * > < = +
« » ] ]></br></l>
<l><br>Etiquetage Treetagger </br><br> OUI</br></l>
<l><br>Langue pour Treetagger </br><br> français</br></l>
</content></sourceDesc>
</fileDesc>
</teiheader>
```

La Trame : les unités du texte

```
<trame>
<codage>utf-8</codage>
<delimiteur><![CDATA[. ,:;!/?/_-'()[]{}$$]*>=<+
«»    ]]></delimiteur>
<items>
    ....
    <item type="forme" pos="768"><f>Les</f><c>DET_ART</c><l>le</l></item>
    <item type="delim" pos="769"><f> </f><c>DELIM</c><l>BLANK</l></item>
    <item type="forme" pos="770"><f>âges</f><c>NOM</c><l>âge</l></item>
    <item type="delim" pos="771"><f> </f><c>DELIM</c><l>BLANK</l></item>
    <item type="forme" pos="772"><f>clés</f><c>NOM</c><l>clé</l></item>
    <item type="delim" pos="773"><f> </f><c>DELIM</c><l>BLANK</l></item>
    <item type="forme" pos="774"><f>de</f><c>PRP</c><l>de</l></item>
    <item type="delim" pos="775"><f> </f><c>DELIM</c><l>BLANK</l></item>
    <item type="forme" pos="776"><f>la</f><c>DET_ART</c><l>le</l></item>
    <item type="delim" pos="777"><f> </f><c>DELIM</c><l>BLANK</l></item>
    <item type="forme" pos="778"><f>vie</f><c>NOM</c><l>vie</l></item>
    <item type="delim" pos="779"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
    ..
    ....
```

Séquence textuelle : “Les âges clés de la vie”
avec lemmes et catégories grammaticales

Le cadre : les parties du texte

```
<cadre>
```

```
<aces>
```

```
<partition nom="page">
```

```
<p n="lefigaro.fr" d="3" f="8505" nd="2" nf="11"/>
```

```
<p n="lemonde.fr" d="8507" f="18557" nd="12" nf="19"/>
```

```
<p n="lopinion.fr" d="18559" f="20817" nd="20" nf="27"/>
```

```
<p n="liberation.fr" d="20819" f="28338" nd="28" nf="35"/>
```

```
<p n="leparisien.fr" d="28340" f="38192" nd="36" nf="43"/>
```

```
<p n="humanite.fr" d="38194" f="44445" nd="44" nf="51"/>
```

```
<p n="la-croix.com" d="44447" f="49669" nd="52" nf="59"/>
```

```
....
```

```
<partition nom="date">
```

```
<p n="20150109" d="1" f="280648" nd="1" nf="412"/>
```

```
<p n="20150202" d="280650" f="506187" nd="413" nf="696"/>
```

```
<p n="20150206" d="506190" f="729740" nd="697" nf="1074"/>
```

```
<p n="20150315" d="729744" f="963496" nd="1075" nf="1468"/>
```

```
</partition>
```



Le Trameur est ...

- Un logiciel de Textométrie
- Un outil d'annotation (démono1)
- Un outil d'exploration de *Treebank*
(corpus arboré, richement annoté) (démono2)
- Un outil de textométrie multilingue
(corpus alignés) (démono3)

Textométrie : préambule

- Objet traité :
 - Le texte (les textes)
- Objectif :
 - Définir et *compter* des unités dans les textes
 - ou dans les parties d'un texte
 - Contraster des parties ou des textes
- Moyens :
 - Outils / méthodes statistiques textuelles

Origines (du Trameur)

- Développement d'outils pour les linguistes
 - Traitements quantitatifs, textométrie
- Equipe CLA²T dirigée par André Salem
 - **Lexico3** | **Lexico5** : <http://lexi-co.com>
 - [Söze-Duval, 2008], Keyser Söze-Duval, *Pour une textométrie opérationnelle*
 - [Fleury, 2013], LE TRAMEUR. PROPOSITIONS DE DESCRIPTION ET D'IMPLÉMENTATION DES OBJETS TEXTOMÉTRIQUES
- Collaborations/discussions
 - Réseaux, partenaires

Le texte : une *Trame* et un *Cadre*

- La *Trame* = Le texte comme suite de positions
 - Une segmentation en unités
 - Annotation des unités et des relations entre unités
- Le *Cadre* = Ensemble des (couples de) positions structurant le texte initial en parties
 - *i.e* les systèmes de parties sur le texte
- *Sélection* :
 - Ensemble de positions de la *Trame* (ou du *Cadre*)

Interface du programme



Barre de navigation : accès aux fonctionnalités

Zone d'édition graphique du cadre

Zone d'édition d'une zone du cadre

Barre de progression

Zone d'écho

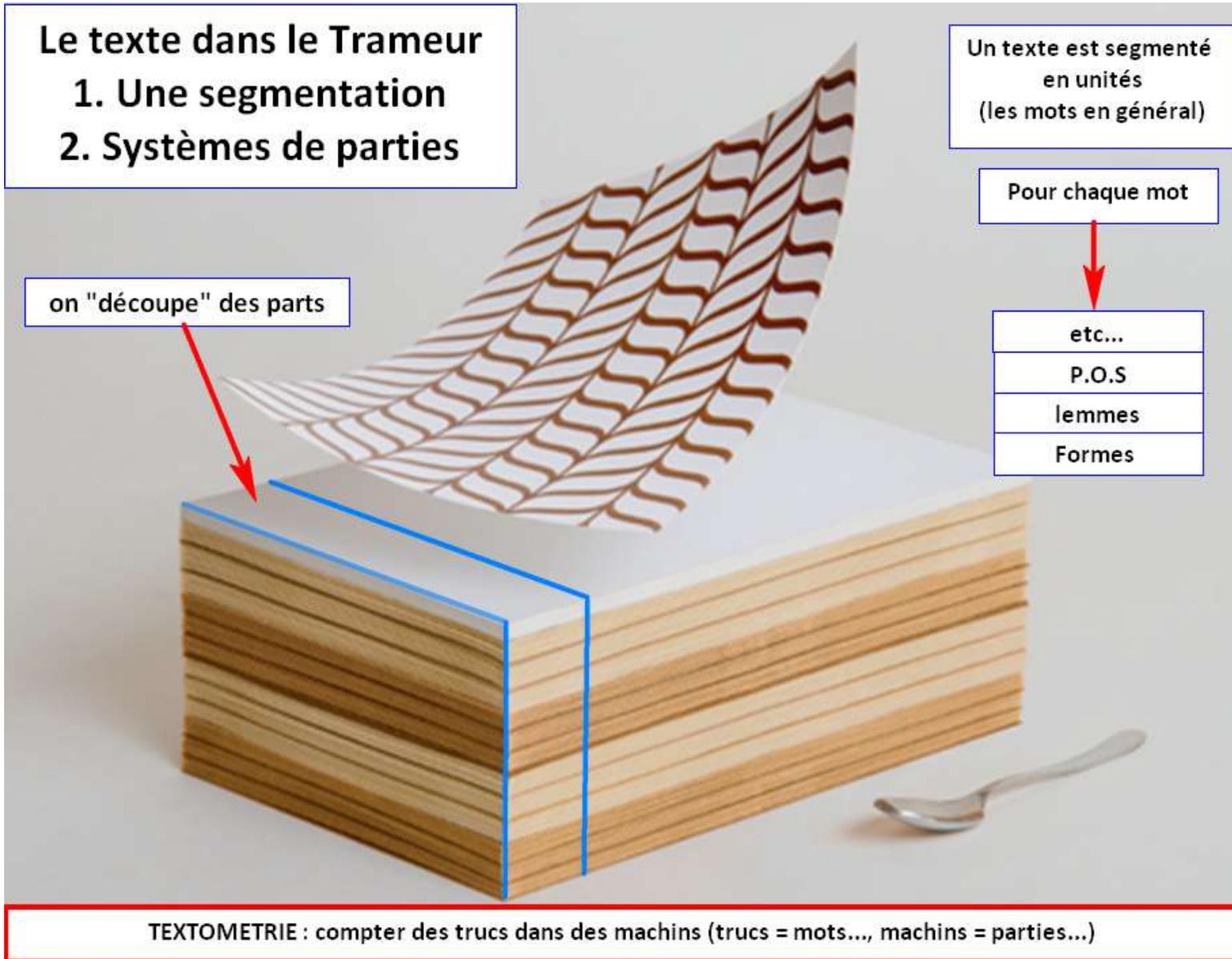
boutons



Comment on
entre dans le
Trameur ?

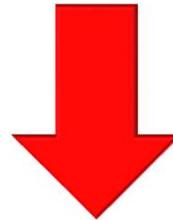


Le Trameur construit des « mille-feuilles »

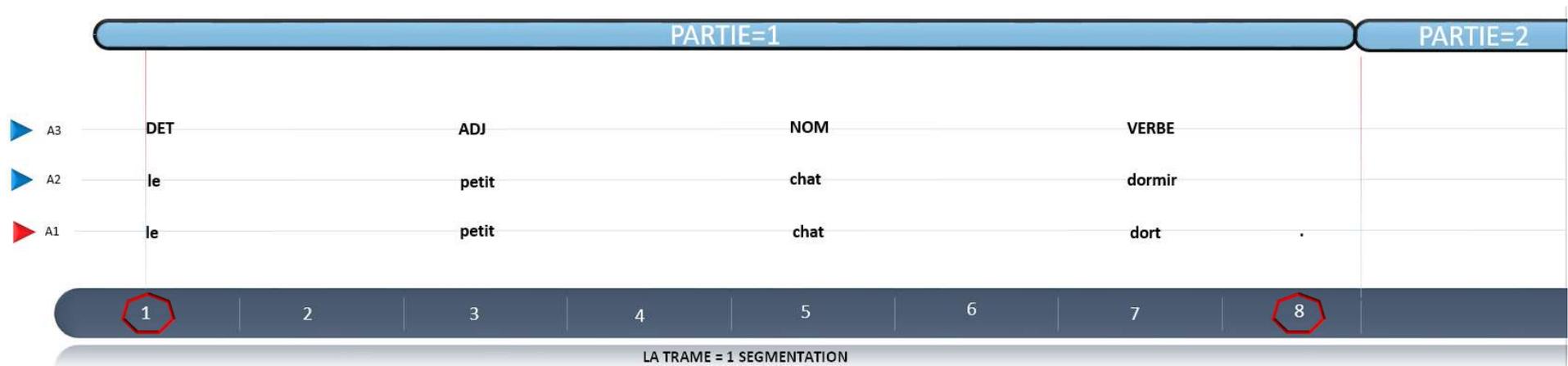


UN CORPUS

<partie=1>
Le petit chat dort.
<partie=2>
Le chien aussi !



1 TRAME (segmentation : suite de positions) + **1 CADRE** (partition : ensemble de couples de positions)



FORMATS d'entrée

- Créer une base

- Texte brut
- Texte semi-structuré (format Lexico3)
- Texte structuré (XML, TEI...)
- Corpus alignés (txt, xml)



- Importer une base

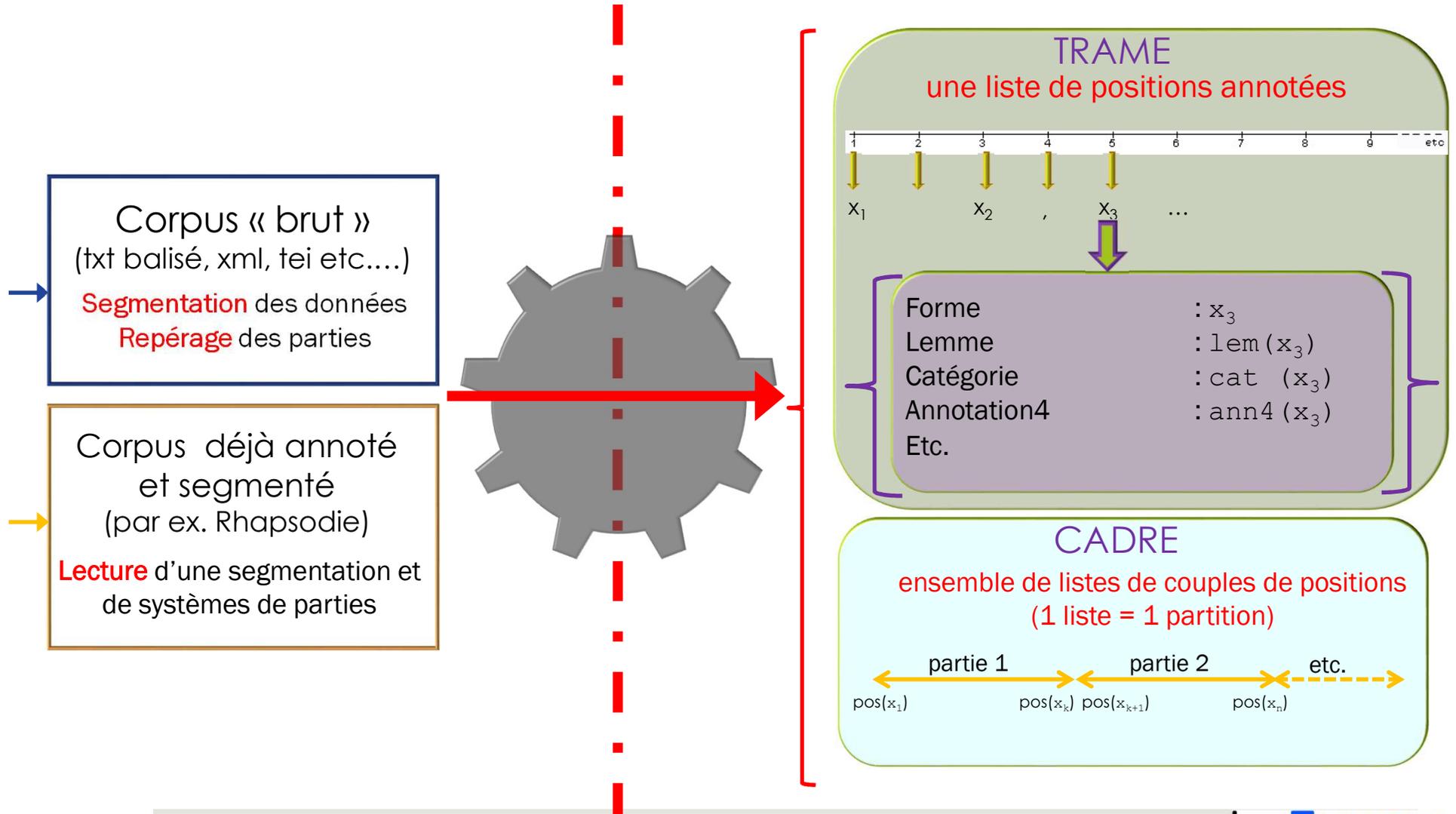
- Format XML prédéfini pour encoder une trame et un cadre (cf documentation)
 - ➔ Similaire au format d'export du logiciel

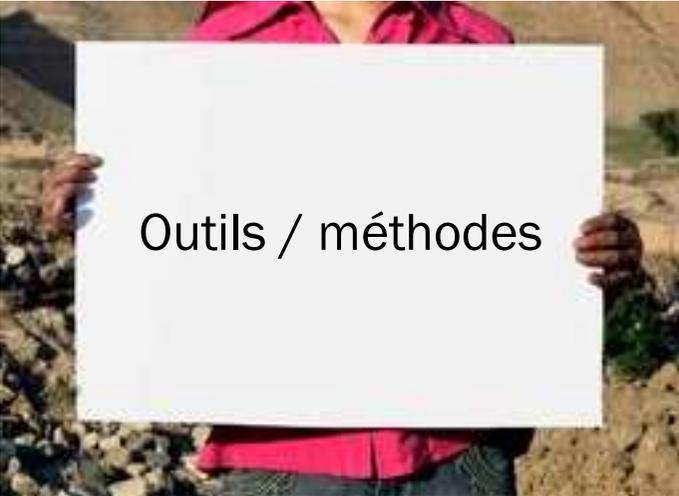


- Exemples sur « Le dormeur du val » / « Ulysse » etc.

Créer vs Importer une base

création vs lecture d'une segmentation





Outils/méthodes de la Statistique textuelle

- Méthodes inductives
 - Méthodes factorielles, classifications etc.
- Méthodes déductives
 - Concordance, ventilation/distribution
 - Cooccurrences
 - Spécificités
 - AFC

Outils / Méthodes dans le Trameur

- ❑ Fréquence, ventilation des unités
- ❑ Annotations des unités, correction des unités... 
- ❑ Concordance
- ❑ Moteur de recherche
- ❑ Sélection des unités
- ❑ Édition de textes alignés 
- ❑ « Poids » des unités dans les parties (spécificités)
- ❑ Contraster les parties (par leur vocabulaire (→ annotations))
- ❑ Proximités de certaines unités (cooccurrents)
- ❑ Au delà des unités primaires (les segments)
- ❑ Extraction de patron
- ❑ Traitement de relations de dépendance (syntaxique) 
- ❑ Méthodes factorielles : AFC...

cf Démo Trameur

Annoter/enrichir les unités

- Annoter les unités dans les textes
 - Etiquetage morphosyntaxique
 - Processus intégré dans *Le Trameur* via *TreeTagger* lors de la création d'une base
 - Importation possible d'autres étiquetages
 - Annotation (sémantique, dépendance, etc.)
 - Annoter, corriger (les annotations) dynamiquement
 - Importer des systèmes d'annotation sur une Trame donnée
- La Trame = une liste de positions annotées
- cf exemples *infra*

Annotations ECRISCOL

(corpus d'écrits d'élèves)

- Le mot « moustache » est composé de 4 annotations : (1) forme (initiale), (2) lemme, (3) catégorie, (4) forme (initiale : normée)

```
----- partie=EL6007 -----  
il faisait friser ses trois poils de moustache et se redressait, l'air bravache.
```

Position:<22999>
Forme:<moustache>|Freq:1
Lemme:<moustache>|Freq:1
Cat:<NOM>|Freq:2489
a-00004:<moustache>|Freq:1

- Le mot « déjà » est composé de 4 annotations dont certaines explicitent la transformation réalisée : (1) forme initiale (erronée), (2) transformation réalisée, (3) catégorie, (4) forme (finale : normée)

```
d'encre, l'homme depuis un moment {déjà} marchait quand il aperçut au loin la
```

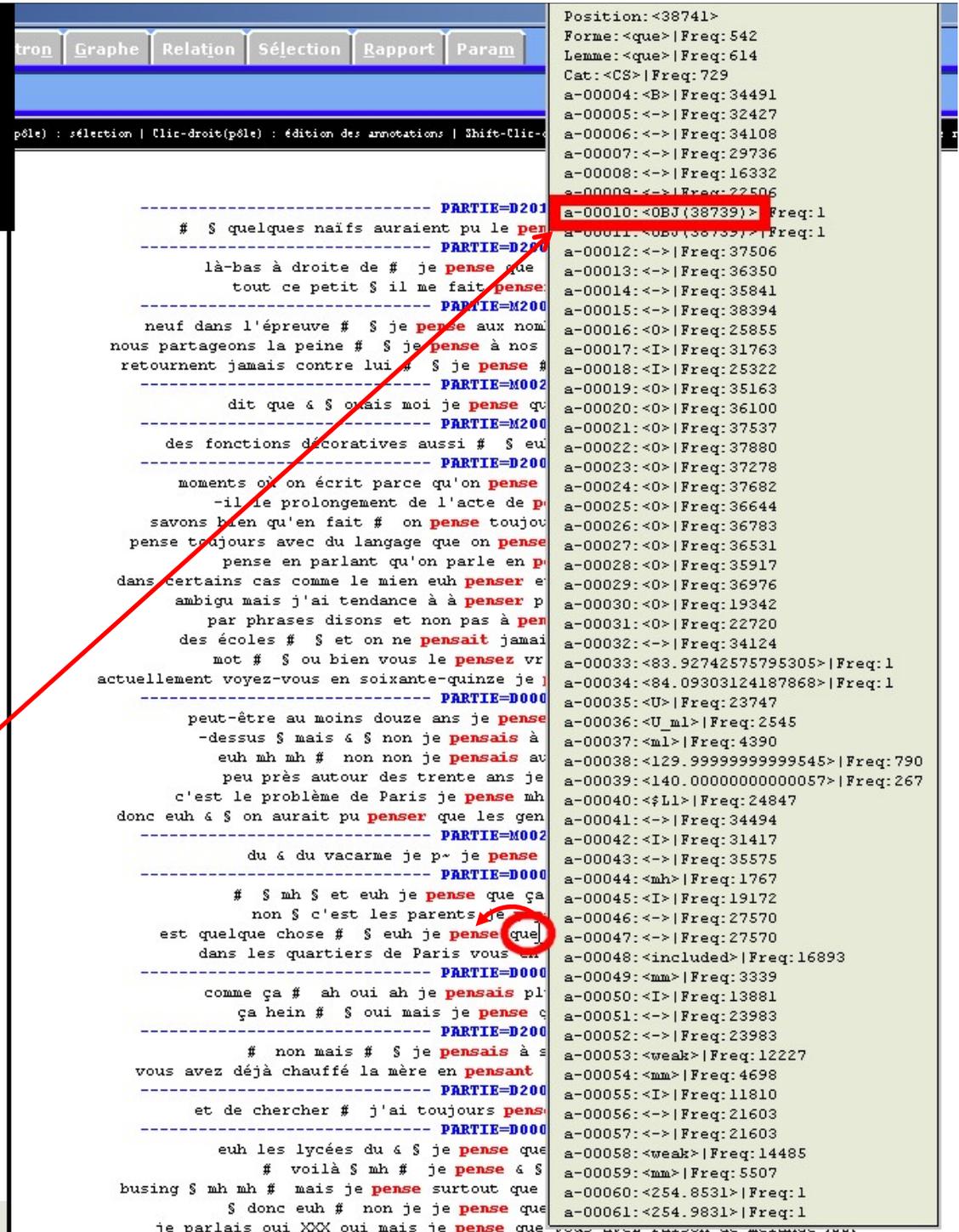
Position:<4420>
Forme:<{déjà}>|Freq:4
Lemme:<{déjà}_{déjà}>|Freq:4
Cat:<ABR>|Freq:675
a-00004:<{déjà}>|Freq:4

Annotations RHAPSODIE

Chaque item de la
Trame est associé à
une soixantaine
d'annotations

Micro/macro syntaxe
Prosodie

Traitement des
dépendances
entre items



The screenshot displays the Rhapsodie software interface. At the top, there are menu tabs: 'tron', 'Graphe', 'Relation', 'Sélection', 'Rapport', and 'Param'. Below the menu, a status bar indicates 'pôle) : sélection | Clic-droit(pôle) : édition des annotations | Shift-Clic-'. The main window shows a text document with several paragraphs of text. The text is annotated with various tags, including 'PARTIE=D201', 'PARTIE=M200', and 'PARTIE=D000'. The annotations are color-coded: 'PARTIE=D201' is blue, 'PARTIE=M200' is red, and 'PARTIE=D000' is green. The text includes phrases like 'quelques naifs auraient pu le penser', 'là-bas à droite de # je pense que', 'tout ce petit \$ il me fait penser', 'neuf dans l'épreuve # \$ je pense aux nom', 'nous partageons la peine # \$ je pense à nos', 'retournent jamais contre lui # \$ je pense #', 'dit que & \$ mais moi je pense qu', 'des fonctions décoratives aussi # \$ eu', 'moments où on écrit parce qu'on pense', '-il le prolongement de l'acte de p', 'savons bien qu'en fait # on pense toujou', 'pense toujours avec du langage que on pense', 'pense en parlant qu'on parle en p', 'dans certains cas comme le mien euh penser e', 'ambigu mais j'ai tendance à à penser p', 'par phrases disons et non pas à pen', 'des écoles # \$ et on ne pensait jamai', 'mot # \$ ou bien vous le pensez vr', 'actuellement voyez-vous en soixante-quinze je l', 'peut-être au moins douze ans je pense', '-dessus \$ mais & \$ non je pensais à', 'euh mh mh # non non je pensais au', 'peu près autour des trente ans je', 'c'est le problème de Paris je pense mh', 'donc euh & \$ on aurait pu penser que les gen', 'du & du vacarme je p~ je pense', '# \$ mh \$ et euh je pense que ça', 'non \$ c'est les parents je', 'est quelque chose # \$ euh je pense que', 'dans les quartiers de Paris vous', 'comme ça # ah oui ah je pensais pl', 'ça hein # \$ oui mais je pense q', '# non mais # \$ je pensais à s', 'vous avez déjà chauffé la mère en pensant', 'et de chercher # j'ai toujours pens', 'euh les lycées du & \$ je pense que', '# voilà \$ mh # je pense & \$', 'busing \$ mh mh # mais je pense surtout que', '\$ donc euh # non je je pense que', 'je parlais oui XXX oui mais je pense que

On the right side, there is a list of annotations for the selected text. The list includes the following items:

- Position: <38741>
- Forme: <que> | Freq: 542
- Lemme: <que> | Freq: 614
- Cat: <CS> | Freq: 729
- a-00004: | Freq: 34491
- a-00005: <-> | Freq: 32427
- a-00006: <-> | Freq: 34108
- a-00007: <-> | Freq: 29736
- a-00008: <-> | Freq: 16332
- a-00009: <-> | Freq: 22506
- a-00010: <OBJ(38739)> | Freq: 1
- a-00011: <OBJ(38739)> | Freq: 1
- a-00012: <-> | Freq: 37506
- a-00013: <-> | Freq: 36350
- a-00014: <-> | Freq: 35841
- a-00015: <-> | Freq: 38394
- a-00016: <O> | Freq: 25855
- a-00017: <I> | Freq: 31763
- a-00018: <I> | Freq: 25322
- a-00019: <O> | Freq: 35163
- a-00020: <O> | Freq: 36100
- a-00021: <O> | Freq: 37537
- a-00022: <O> | Freq: 37880
- a-00023: <O> | Freq: 37278
- a-00024: <O> | Freq: 37682
- a-00025: <O> | Freq: 36644
- a-00026: <O> | Freq: 36783
- a-00027: <O> | Freq: 36531
- a-00028: <O> | Freq: 35917
- a-00029: <O> | Freq: 36976
- a-00030: <O> | Freq: 19342
- a-00031: <O> | Freq: 22720
- a-00032: <-> | Freq: 34124
- a-00033: <83.92742575795305> | Freq: 1
- a-00034: <84.09303124187868> | Freq: 1
- a-00035: <U> | Freq: 23747
- a-00036: <U_ml> | Freq: 2545
- a-00037: <ml> | Freq: 4390
- a-00038: <129.999999999999545> | Freq: 790
- a-00039: <140.00000000000057> | Freq: 267
- a-00040: <fLl> | Freq: 24847
- a-00041: <-> | Freq: 34494
- a-00042: <I> | Freq: 31417
- a-00043: <-> | Freq: 35575
- a-00044: <mh> | Freq: 1767
- a-00045: <I> | Freq: 19172
- a-00046: <-> | Freq: 27570
- a-00047: <-> | Freq: 27570
- a-00048: <included> | Freq: 16893
- a-00049: <mm> | Freq: 3339
- a-00050: <I> | Freq: 13881
- a-00051: <-> | Freq: 23983
- a-00052: <-> | Freq: 23983
- a-00053: <weak> | Freq: 12227
- a-00054: <mm> | Freq: 4698
- a-00055: <I> | Freq: 11810
- a-00056: <-> | Freq: 21603
- a-00057: <-> | Freq: 21603
- a-00058: <weak> | Freq: 14485
- a-00059: <mm> | Freq: 5507
- a-00060: <254.8531> | Freq: 1
- a-00061: <254.9831> | Freq: 1

DEMO N°1



\$ la grande colère du *père *duchesne , de voir que les mouchards de *la-*fayette et tous les fripons soudoyés par la liste civile, veulent rétablir les compagnies de grenadiers et de chasseurs, pour égorger les *sans-culottes et les chasser des assemblées de *section .ses bons avis aux *lurons des *faubourgs pour qu' ils arrachent les moustaches postiches à ces grenadiers de la vierge *marie , qui veulent rétablir la royauté.
<S03=1>

\$ millions de tonnerre, nous ne mettrons donc jamais les fripons à la raison ? ils <Epg=2>ont laissé tomber leurs masques et nous les voyons à nu. serons nous encore dupes des fripons? quand je voulais faire la conduite de *grenoble à tous les talons rouges quand je disais, du soir au matin, que tous les ci-devant ne cesseraient de nous trahir, n' avais je pas raison, foutre?

\$ je me suis toujours plus défié des nobles convertis que des émigrés. c' est pour nous frapper de plus près que ces gredins sont restés au milieu de nous. ils ont fait les chiens couchants pour mieux nous tromper. jamais, foutre, ils n' ont cessé de s' entendre avec les ennemis du dehors. ce sont eux qui nous ont mis à chien et à chat, qui ont brouillé les cartes dans les trois assemblées nationales, et corrompu les représentants du peuple. si nous avions eu assez d' estoc pour les envoyer tous à *coblentz au commencement de la révolution, nous n' aurions pas acheté notre liberté par des flots de sang; nous aurions depuis longtemps une constitution; la paix et le bonheur régneraient dans notre république.

Guilhaumou J. (1986) - L'historien du discours et la lexicométrie. Etude d'une série chronologique : Le père Duchesne de Hébert, juillet 1793- mars 1794, Histoire & Mesure , Vol. I, n° 3-4.

DEMO N°2



<http://projet-rhapsodie.fr/>



Rhapsodie et Le Trameur en 2 mots

- Rhapsodie
 - Le TREEBANK RHAPSODIE (<http://www.projet-rhapsodie.fr/>)
 - de 33000 mots alignés au phonème a été développé avec comme objectif théorique la modélisation de l'interface prosodie, syntaxe, discours en français parlé...
- Le Trameur
 - *A Framework for Annotated Text Corpora Exploration*
 - <http://www.tal.univ-paris3.fr/trameur/>

La base



"Rhapsodie2Trameur" (v8) (màj : 11.05.2015). Chaque item de la *Trame* est associé à 61 niveaux d'annotation (prosodie, micro et macro-syntaxe).

```
nous ferons le point sur les violences en direct avec notre envoyé spécial Sébastien Paour dans quelques instants $
```

```
Position: <91>
Forme: <point> | Freq: 34
Lemme: <point> | Freq: 23
Cat: <B_N> | Freq: 6247
a-00004: <-> | Freq: 70852
a-00005: <-> | Freq: 72497
a-00006: <-> | Freq: 68133
a-00007: <sg> | Freq: 17353
a-00008: <masc> | Freq: 8816
a-00009: <OBJ(87)> | Freq: 1
a-00010: <-> | Freq: 74709
a-00011: <-> | Freq: 74145
a-00012: <-> | Freq: 75936
a-00013: <-> | Freq: 76812
```

SOURCES : [projet Rhapsodie](#)

Présentation



Descriptif et sources des annotations : [Annotations Rhapsodie pour le Trameur \(v8\)](#) (pdf)
Présentation du processus de transcodage des annotations Rhapsodie pour construire une base Textométrique.
Présentation des différents processus de traitements des annotations de dépendance.



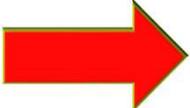
<http://www.tal.univ-paris3.fr/trameur>

Treebank Rhapsodie : une base « quasi » textométrique

Une segmentation en unités (des annotations sur les unités) → une *Trame*

Un système de parties (les échantillons du corpus) → un *Cadre*

Transcodage automatique :

Rhapsodie  Base textométrique

Données Rhapsodie

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	TextID	TreeID	TokenID	Token	Lemma	POS	Mode	Tense	Person	Number	Gender	Gov_rection	Type_rection	Gov_para	Type_para	Gov_inher	Type_inher	Gov_junc	Type_junc	Gov_juncinher	Type_juncinher
2	M2006	1	1	bonjour	bonjour	B_J						0	root								
3	M2006	1	2																		
4	M2006	1	3	Eric	Eric	B_N				sg	masc		0	root							
5	M2006																				
6	M2006	2	1	bonjour	bonjour	B_N				sg	masc		0	root							
7	M2006	2	2																		
8	M2006	2	3	Ä	Ä	B_Pre							1	dep							
9	M2006	2	4																		
10	M2006	2	5	tous	tous	B_Pro				3 pl	masc		3	dep							
11	M2006																				
12	M2006	3	1	nouvelle	nouveau	B_Adj				sg	fem		3	dep							
13	M2006	3	2																		
14	M2006	3	3	nuit	nuit	B_N				sg	fem		0	root							
15	M2006	3	4																		
16	M2006	3	5	de	de	B_Pre							3	dep							
17	M2006	3	6																		
18	M2006	3	7	pillage	pillage	B_N				sg	masc		5	dep							
19	M2006	3	8																		
20	M2006	3	9	et	et	B_J													5	junc	
21	M2006	3	10																		
22	M2006	3	11	d	de	B_Pre							5	para_coord		3	dep_inherited		9	junc	
23	M2006	3	12	'	^	I_Pre															
24	M2006	3	13	affrontement	affrontement	B_N				sg	masc		11	dep							
25	M2006	3	14																		
26	M2006	3	15	en	en	B_Pre							3	dep							
27	M2006	3	16																		
28	M2006	3	17	Guadeloupe	Guadeloupe	B_N				sg	masc/fem		15	dep							
29	M2006																				
30	M2006	4	1	trois	trois	B_D				sg	masc		3	dep							
31	M2006	4	2																		
32	M2006	4	3	policiers	policier	B_N				pl	masc		0	root							
33	M2006	4	4																		
34	M2006	4	5	blessés	blessés	B_Adj				pl	masc		3	dep							
35	M2006																				

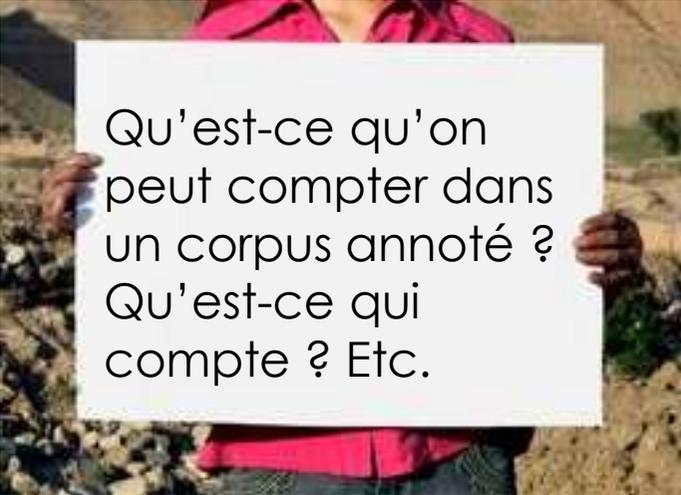
Ces données sont constituées par un certain nombre de textes (l'identifiant du texte est visible dans la première colonne), chacun d'eux est segmenté en « unité illocutoire » (UI, seconde colonne), chacune d'elle est segmentée en *token* (troisième colonne), chacun d'eux est annoté (les autres colonnes)

Transcodage des dépendances

```
<item type="delim" pos="46"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="47"><f>lance</f><c>B_V</c><l>lancer</l><a>indicative</a><a>present</a><a>3</a><a>sg</a><a>-</a><a>ROOT</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="48"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="49"><f>un</f><c>D</c><l>un</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="50"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="51"><f>appel</f><c>B_N</c><l>appel</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="52"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
```

Les annotations initiales de dépendance (identifiées dans 2 colonnes) sont « fusionnées » deux à deux sous la forme d'une relation marquée de la manière suivante :

- RELATION (CIBLE) :
 - Position 51 : « appel » OBJ(47)  « lance » (position 47)
- RELATION est une chaîne portant le nom de la relation visée
- CIBLE est une valeur numérique pointant vers une position de la *Trame*



Qu'est-ce qu'on
peut compter dans
un corpus annoté ?
Qu'est-ce qui
compte ? Etc.



Qu'est-ce qu'on peut compter ?

Annotations et dépendances

1. Quelle CATEGORIE peut être devant un VERBE ?
2. Quelles sont les MOTS qui peuvent être OBJET de "penser" ?
3. Quels sont les CATEGORIES connectées dans une relation de type OBJ ? De type SUB ? (longueur de la dépendance, postposition vs antéposition du dépendant...)
4. Quels sont les dépendants des VERBES ?
5. Les collocations de PENSER (cooccurrence + dépendance)

{ Comptages « simples » { Un peu plus de « stat »

Démo avec le Trameur (1)

Fq	Séquences de termes
2226	Cl V
1207	- V
874	V V
341	N V
337	Qu V
255	Pre V
230	Adv V
131	Pro V
111	I V
50	J V
42	Adj V
10	X V
9	CS V
4	D V

- V pour les verbes (et *voici* et *voilà* lorsqu'ils prennent des compléments, comme par exemple *voici un extrait...*)
- N pour les noms
- Adj pour les adjectifs
- Adv pour les adverbes
- Pre pour les prépositions
- CS pour les conjonctions de subordination
- J pour les joncteurs : il s'agit des traditionnelles conjonctions de coordinations et d'autres éléments qui lient les couches d'un entassement, comme *c'est-à-dire* ou *y compris*. Les éléments clôtureurs d'entassement comme *et caetera* sont classés comme joncteurs également.
- D pour les déterminants
- I pour les interjections, y compris des marqueurs de discours comme *bon*, *ben*, *euh*, *hein*, (les formes impératives comme *allons*, *écoute*, *tiens* sont traitées comme des verbes à l'impératif et non pas comme des interjections)
- Qu pour les mots *qu-* que sont les relatifs et les interrogatifs
- Cl pour les clitiques, y compris les clitiques sujets (*je*, *tu*, *il*, *on*, *ce*) et l'adverbe de négation *ne*.
- Pro pour les autres pronoms
- X pour les éléments dont on ne peut déterminer la catégorie syntaxique : partie inaudible (XXX), certaines amorces (quand on ne peut pas deviner le lexème et sa partie du discours), ainsi que les positions non instanciées marquées par &, uniquement quand on ne peut pas déterminer la catégorie attendue.

Retour en contexte

n°	Partie	ContexteGauche	Pôle	ContexteDroit
1	UI=0000000589		tu-Cl peux-V	le-D dire-V §
2	UI=0000001179		mais-J à-Pre	ce-D moment-V --Adv là-- il-Cl y-Cl a-V une-D dame-N un-D peu-N plu
3	UI=0000002414	I-D '- école-N c-Cl '- était-V &X est-V conçue-V maintenant-Adv comme-Pre	un-D	voilà-V pensum-N §
4	UI=0000002476	mais-J je-Cl crois-V que-CS j-Cl '- ai-V toujours-Adv oui-I eu-V eu-V cette-D	cette-D	disons-V cette-D faculté-N # §

Démo avec le Trameur (2)

The screenshot displays the 'Sélection de relation à afficher' window in the Trameur software. The window is divided into two main sections: a configuration panel on the left and a dependency graph on the right.

Configuration Panel (Left):

- Relation visée :** OBJ (highlighted with a red circle), 10
- n°Annotation :** Identifiant Item (Position) 0
- Filtrage sur les autres niveaux d'annotation:** A table with 'SOURCE' and 'CIBLE' columns. The 'CIBLE' column has 'Lemme' set to 'pense' (highlighted with a red circle).
- Buttons: Options, Annuler, Chercher (with a red arrow pointing to it).

Dependency Graph (Right):

- Shows a network of nodes representing linguistic items. Each node is a box containing a position and a lemma, e.g., 'Pos: <416> que' or 'Pos: <414> pense'.
- Nodes are connected by arrows labeled 'OBJ', indicating dependency relationships.
- The graph shows a complex web of dependencies, with 'que' nodes often depending on 'pense' nodes.

Ici, on cherche des items en relation de dépendance tels que :
SOURCE porte l'annotation cherchée **OBJ**
CIBLE doit porter en outre le lemme « penser »

Au final on obtient les items **SOURCE** qui sont **OBJET** de **CIBLE**
Par exemple « que » est **OBJET** de « penser »

PARTIE=D0002

les lycées du & § je **pense que** à l'intérieur de Paris #
voilà § mh # je **pense &** § moi je vois euh dans le
mh mh # mais je **pense** surtout **que** euh dans le sixième arrondissement les maternelles
donc euh # non je je **pense que** le problème & § je dis pas
parlais oui XXX oui mais je **pense que** vous avez raison de mélange XXX euh
que là # & § je **pense que** vous avez raison § mais mh #

PARTIE=D2011

est votre linge § qu'est-ce **que** vous en **pensez** de la boule magique

PARTIE=D2013

politique # § quelques naïfs auraient pu **le penser** # § eh bien non

PARTIE=D2007

avez déjà chauffé la mère en **pensant que** vous vous adressiez à la fille #

PARTIE=D0006

§ mh § et euh je **pense que** ça a été très bénéfique § vous
quelque chose # § euh je **pense que** la chirurgie c'est pareil hein #
les quartiers de Paris vous en **pensez quoi** # ici § euh # alors

PARTIE=M0023

que & § ouais moi je **pense qu'**en fait il voit d'abord la

PARTIE=D2005

de chercher # j'ai toujours **pensé qu'**on n'est pas sur terre pour

PARTIE=D2012

heure # § donc on peut **penser que** c'est une tradition euh ici qui
qui est représentée § on peut **penser que** euh à l'époque de Szymanoski on

PARTIE=D0009

ça # ah oui ah je **pensais plus** que ça oui # quarante quarante-
hein # § oui mais je **pense que** non ça va ça va pas très

PARTIE=D2008

pas de vaccins # § je **pense que** aujourd'hui euh s~ euh même une

PARTIE=M1001

§ mais en fait euh je **pense que** je vais peut-être me réorienter #
euh je réussis pas ben je **pense que** je vais en psychologie parce que c
passe l'année par contre je **pense que** je vais finir ma licence § s~
'est très intéressant § mais je **pense que** je me réorienterai en psychologie parce que
italien # § donc euh je **pense que** je vais euh je vais sûrement me

PARTIE=D2003

droite de # je **pense** que c'est Mascherano ou M~ euh peut~ peut-être

PARTIE=D2009

un mot # § ou bien vous **le pensez** vraiment § ou peut-être y
voyez-vous en soixante-quinze je **pense que** vous serez d'accord avec moi #

PARTIE=D0004

euh & § on aurait pu **penser que** les gens du quatorzième étaient privilégiés restaient

PARTIE=D2010

les musées hein # § je **pense que** si on s'était assis l'un

Démo avec le Trameur (3)

Relation	POS source	POS cible	Fq▼	LGmoy	POST	ANTE
OBJ	N	V	1081	1.64	1079	2
OBJ	Cl	V	346	0.62	19	327
OBJ	CS	V	199	0.79	182	17
OBJ	V	V	161	2.07	153	8
OBJ	Qu	V	141	2.10	19	122
OBJ	Pro	V	75	0.61	69	6
OBJ	Adv	V	40	0.47	39	1
OBJ	Adj	V	17	1.24	17	0
OBJ	Pre	V	14	2.07	13	1
OBJ	X	V	8	0.50	8	0
OBJ	I	V	8	0.13	8	0
OBJ	J	V	6	0.67	6	0
OBJ	Pre+D	V	2	0.00	2	0
OBJ	Qu	X	1	2.00	0	1

PARTIE=D2013

déroute \$ et rien n'**arrête** la **vague** rose écrivent Les Dernières Nouvelles d'Alsace de Libération qui ne **boude** pas son **plaisir** une caricature de Nicolas Sarkozy tout ratatiné UMP # s'obstinent à **nier** l'**évidence** dans un accès de mauvaise foi soviétique Figaro aussi la gauche a **transformé** l'**essai** du premier tour Le Figaro qui reconnaît tour Le Figaro qui **reconnaît** l'écrasante **victoire** de la gauche \$ le parti socialiste frappe territoriale considérable # \$ alors quelle **leçon tirer** de la vague rose \$ faut-il que le gouvernement pour **tenir compte** du vote des Français change sa politique compte du vote des Français **change** sa **politique** # \$ quelques naïfs auraient pu Figaro au contraire il faut **accélérer** le **rythme** du changement # \$ les réformes pas convaincu **balaye** d'un ricanement la **pensée** d'Étienne Mougeotte \$ les Français disent vite # \$ **voyons** maintenant l'autre **enjeu** largement évoqué avant les municipales \$ scrutin des municipales et des cantonales # un **message** fort envoyé à l'exécutif message national journaux Jean Christophe qui **soulignent** également la **faiblesse** de la mobilisation des électeurs hier # Roger Antech dans Midi Libre \$ **un sauvetage** miraculeux à bien des égards **écrit** La Provence # \$ Marseille **occupe** une **place** à part \$ et ce matin le il va sans doute **faire** la même **chose** qu'avant pronostique Francis Brochet mais autrement

PARTIE=D2003

avec Messi qui va **chercher** oh le **corner** et qui va trouver le corner ou le corner et qui va **trouver** le **corner** ou la touche ou la sortie de de but puisque il a **poussé** le **ballon** trop loin Messi \$ et il y il semblerait que Courcuff ait **perdu** le **ballon** ouais raison pour laquelle peut-être monsieur monsieur Erikson n'a pas **sifflé** le **penalty** # \$ alors Gallas a pris # \$ alors Gallas a **pris** un **carton** je vous signale tout de même \$ la tête de Mexès qui **repousse** ce **ballon** de la tête en corner # Agüero # \$ et il **expédie** ce **ballon** en corner \$ il n'y avait le côté droit qui va **tirer** ce **corner** \$ il ne le tire pas d gauche également Papa qui va **mettre** le **ballon** dans la petite boîte \$ surface de **frappe** à environ vingt-deux mètres le **ballon** dans la niche de Steve Mandanda \$ oh la la qui **marque** ce # **but** du pied droit \$ mais le côté \$ il a # pu **azmer** son **tir** \$ Mandanda a plongé \$ mais bon et Gutiérrez Gutiérrez donc Gutiérrez **marque** ce **but** # \$ alors que c'est quoique là il y **avait** un petit **relâchement** \$ eh bien voilà j~ j~ j~ oui et # commençait à **lâcher** son **emprise** sur l'adversaire # argentin \$ peu plus haut \$ mais on **voyait** **Maradona** faire de grands signes # euh mais on voyait Maradona **faire** de grands **signes** # euh demandant à ses joueurs à zéro \$ voilà qui **change** les **données** # euh du problème \$ parce

Relation	POS source	POS cible	Fq▼	LGmoy	POST	ANTE
SUB	Cl	V	2984	0.56	79	2905
SUB	N	V	455	2.18	45	410
SUB	Qu	V	419	0.54	0	419
SUB	Pro	V	189	0.72	1	188
SUB	Adj	V	7	1.00	0	7
SUB	V	V	4	7.25	0	4
SUB	Adv	V	3	2.33	0	3
SUB	X	V	1	0.00	0	1
SUB	J	V	1	1.00	0	1
SUB	Pre	V	1	5.00	0	1

Démo avec le Trameur (4)

Les
dépendants
de V

Recherche Relation (R1)				
Relation	.*			
n°Annot. Relation	10			
n° Ident. Item	0			
<input checked="" type="checkbox"/> (Position)				
<input type="button" value="Chercher"/>		<input type="button" value="Chercher Cible"/>		
Liste Relations (R2)				
Relation	.*			
n°Annot. Relation	10			
n° Ident. Item	0			
<input checked="" type="checkbox"/> (Position)				
POS source	.	POS cible	V	
<input type="button" value="Chercher"/>				

Relation	POS source	POS cible	Fq▼
SUB	Cl	V	2969
AD	Adv	V	1127
OBJ	N	V	1067
PRED	V	V	1019
AD	Pre	V	816
OBL	Pre	V	580
SUB	N	V	441
SUB	Qu	V	418
PRED	N	V	388
OBJ	Cl	V	347
PRED	Adj	V	340
DEP	Cl	V	326
AD	Cl	V	228
OBJ	CS	V	192
SUB	Pro	V	188
OBL	Cl	V	186
OBJ	V	V	161
PARA_DISFL	V	V	144
OBJ	Qu	V	138
AD	CS	V	125
PRED	Pre	V	120
AD	N	V	112
JUNC	J	V	98
AD	Qu	V	87
DEP	V	V	83
OBL	Pre+D	V	80
OBJ	Pro	V	75
PARA_REFORM	V	V	73

Les gouverneurs de V

Recherche Relation (R1)		R1	R2	R3	
Relation	.+	Relation	POS source	POS cible	Fq▼
n° Annot. Relation	10	PRED	V	V	1019
n° Ident. Item	0	DEP	V	CS	618
<input checked="" type="checkbox"/> (Position)		DEP	V	J	601
<input type="button" value="Chercher"/>	<input type="button" value="Chercher Cible"/>	DEP	V	N	423
Liste Relations (R2)		DEP	V	Pre	391
Relation	.+	OBJ	V	V	161
n° Annot. Relation	10	PARA_DISFL	V	V	144
n° Ident. Item	0	DEP	V	Pro	99
<input checked="" type="checkbox"/> (Position)		DEP	V	V	83
POS source	V	PARA_REFORM	V	V	73
POS cible	.+	PARA_COORD	V	V	53
<input type="button" value="Chercher"/>		JUNC	V	J	37
		DEP	V	Qu	24
		DEP	V	Cl	22
		OBJ	V	CS	19
		AD	V	V	15
		PARA_DFORM	V	V	11
		DEP	V	Adv	11
		DEP	V	Adj	8
		PARA_INTENS	V	V	6
		PARA_NEGOT	V	V	5
		SUB	V	V	4
		OBJ	V	N	4
		PARA_REFORM	V	Pre	3
		PARA_REFORM	V	N	2
		PARA_HYPER	V	V	2
		PARA_COORD	V	Adj	2
		AD	V	Pre	2
		PARA_COORD	V	Pre	2
		OBJ	V	Cl	1
		OBL	V	V	1
		AD	V	CS	1
		SUB	V	Pre	1
		PARA_COORD	V	N	1
		PARA_NEGOT	V	N	1
		DEP	V	D	1
		PARA_REFORM	V	Adj	1
		DEP	V	X	1
		JUNC	V	Adv	1
		PRED	V	Pre	1
		PARA_DISFL	V	Pre	1

Démo avec le Trameur (5)

Paramètres :

Pôle : penser

Co-Freq : 4

Seuil : 4

Contexte : \$

RELATION

n°Annot. Relation : *

n° Ident. Item : 0

(Position)

RegExp Fq Partie Filtrage

Sélection Annotation :

Forme Lemme Catégorie

Stop-liste (édition/import)

Cooccurrences :

+3R Layout

Poly-Cooccurrences :

(*) Collocation - Relation :

Ajouter au Rapport (liste)

Sauvegarde Graphe Poly-Cooc

Clic-droit sur noeud : concordance | ctrl-clic sur noeud : sélection

Cooccurents (Lemme (annotation 2) / Relation)

Pole : penser | Relation : (PARA_COORD PRED OBJ SUB PARA_DISFL PARA_NEGOT PARA_REFORM PARA_INTENS OBL PARA_DFORM DEP JUNC PARA_HYPER AD)

fq : 53

Co-Freq : 4 | Seuil : 4

ARC Couleur => bleu: (4<Sp<=6) | vert: (6<Sp<=12) | orange: (12<Sp<=50) | rouge: (50<Sp)

ARC Epaisseur -> 1: (1<=NbContexte<=20) | 2: (20<NbContexte<=40) | 3: (40<NbContexte<=60) | 4: (60<NbContexte)

ARC label : cofreq(Sp)(nbContexte)

Cooccurrences: Forme pôle co-présence co-pr. + pôle

Pôle	Cooccurrent	Fq (Cooccurrent)	co-freq	specif	contextes
penser//OBL	à+le	187	5	8.2	5
penser//SUB	on	305	4	4.3	3
penser//OBJ	que	614	23	41.7	23
penser//OBL	à	545	5	8.2	5
penser//AD	par	113	4	5.5	3
penser//SUB	je	787	31	**	31

DEMO N°3



Explorations textométriques multilingues

- Corpus comparable
 - Base bilingue BBC-EN-RU (corpus comparable aligné anglais/russe)
- Corpus « bruts » alignés
 - Textes bruts alignés « verticalement »
 - Corpus **Convention** (3 volets), corpus **Investiture Obama** (5 volets)
 - Corpus **Freud IntroPsy** (3 volets)
- Base textométrique multilingue annotée
 - Textes alignés et annotés
 - Corpus **Investiture Obama** annoté : Base bilingue "Investiture Obama" (français/anglais)
 - Corpus **Freud** annoté : **Base bilingue « IntroPsy »** (allemand/français)
 - Alignement de treebanks
 - Corpus **PARTUT** : Base bilingue ParTUT2Trameur (français/anglais)

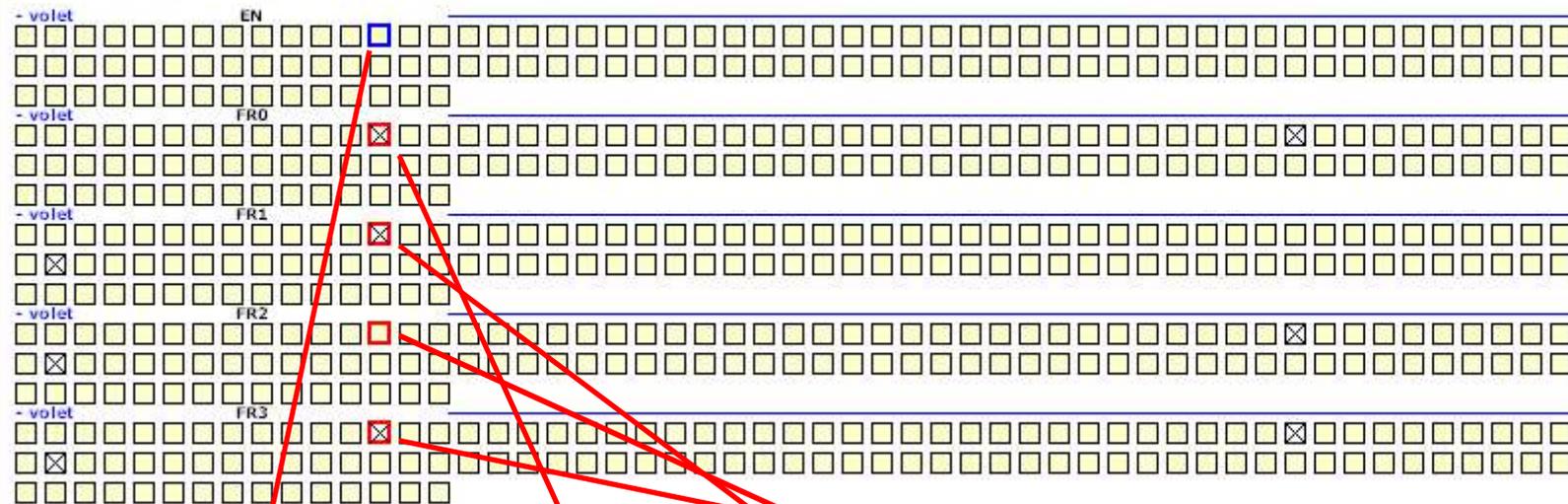
Corpus *Investiture Obama* annoté

- 30.000 mots, 5 volets
- Discours prononcé par B. Obama le 20/01/2009
- + 4 traductions en français
- Annotation morpho-syntaxique via *treetagger*
- Alignement en phrases (caractère §)

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 le seuillage :

(+/-) : masquage/affichage des sections en cliquant sur la partie



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

Nb L. Sections sélectionnées : 0 N° Sect. : 13:(463,484) Annotation : 1 Aperçu : 50 Nb Volets BiText

Homes have been lost, jobs shed, businesses shuttered. §	Des domiciles ont été perdus ; des emplois ont été éliminés ; des entreprises ont été brisées. §	Des maisons ont été perdues; des emplois ont été détruits; des entreprises ont fait faillite. §	Des gens ont perdu leur maison ou leur emploi, des entreprises ont dû fermer leurs portes. §	Des gens n'ont plus de domicile. Des emplois sont perdus. Des entreprises sont brisées. §
---	--	---	---	---

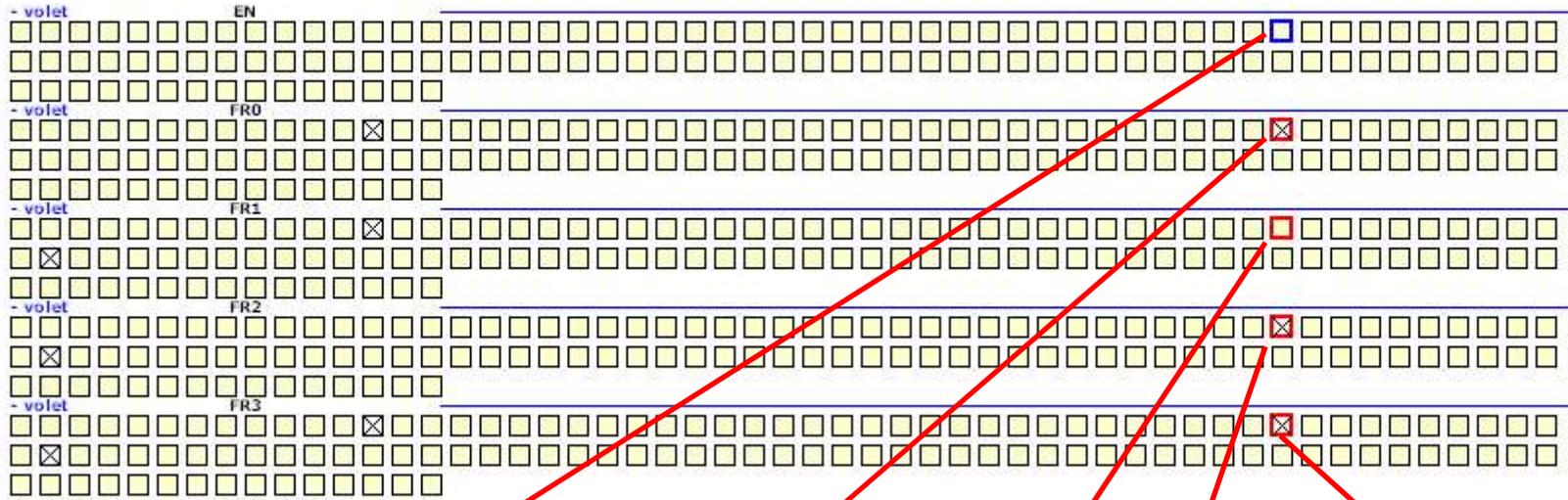
Shift-Clic : sélection | Clic-droit : édition | Ctrl-Clic : noeud | 2-Clic : graphe | Shift-Clic-droit : relation | Control-Clic-droit : recherche relation

Annotations : 1 2 3

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 le seuillage :

(+/-) : masquage/affichage des sections en cliquant sur la partie



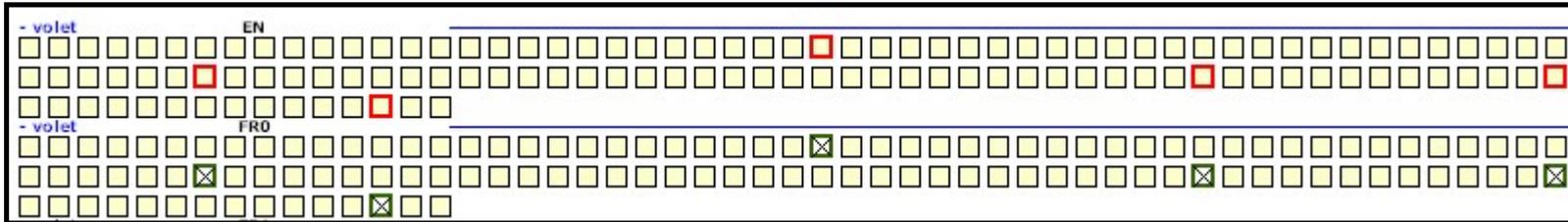
Control-clic sur marqueur de page : sélection 5 sections | Shift-control-clic sur marqueur de page : sélection 25 sections (1 ligne)

Mb L. Sections sélectionnées : 0 N° Sect. : 44: (1735,1773) Annotation : 1 Aperçu : 50 Mb Volet 5

And we will act not only to create new jobs but to lay a new foundation for growth. §	et nous agirons - non seulement pour créer des emplois , mais pour poser une nouvelle fondation de la croissance. §	et nous allons les prendre - pas seulement pour créer de nouveaux endroits, mais pour poser des jalons en vue de faire redémarrer la croissance. §	Et nous agirons - non seulement pour créer de nouveaux emplois mais pour jeter les fondations d'une nouvelle croissance. §	et nous n'agirons pas seulement pour créer de nouveaux emplois , mais pour poser les fondations d'une nouvelle croissance. §
--	---	---	--	--

Shift-Clic : sélection | Clic-droit : édition | Ctrl-Clic : noeud | 2-Clic : graphe | Shift-Clic-droit : relation | Control-Clic-droit : recherche relation

Annotations : 1 2 3



Specificites du vocabulaire (VOLET FR0 - Annotation Forme) sur les sections SOURCE contenant le motif : liberté

Nombre d'occurrences du texte global = 2845
 Nombre d'occurrences dans la partie visee = 188
 Seuil : 10
 (Specificites positives en haut de liste, negatives en bas)

Item	Ind-Specif	Fq-Totale	Fq-Partie	
liberté	6.9	5	5	S
sens	3.4	2	2	S
qui	3.0	38	7	S
sans	2.6	4	2	S
hommes	2.6	4	2	S
toute	2.6	4	2	S
femmes	2.6	4	2	S
peuvent	2.4	5	2	S
des	2.3	42	6	S
race	2.2	1	1	S
agissent	2.2	1	1	S
yeux	2.2	1	1	S
mène	2.2	1	1	S
joyeusement	2.2	1	1	S
répandre	2.2	1	1	S
laquelle	2.2	1	1	S
mais	2.2	14	3	S
horizon	2.2	1	1	S
porté	2.2	1	1	S
vigilant	2.2	1	1	S
pareil	2.2	1	1	S
futures	2.2	1	1	S
raison	2.2	1	1	S
disposition	2.2	1	1	S
les	2.2	54	7	S
ombre	2.2	1	1	S
credo	2.2	1	1	S
honorons	2.2	1	1	S
grâce	2.2	1	1	S
prennent	2.2	1	1	S
fabriquent	2.2	1	1	S
choses	2.2	1	1	S
transmis	2.2	1	1	S
travaillent	2.2	1	1	S

Specificites du vocabulaire (VOLET EN - Annotation Forme) sur les sections CIBLE m

Nombre d'occurrences du texte global = 2411
 Nombre d'occurrences dans la partie visee = 154
 Seuil : 10
 (Specificites positives en haut de liste, negatives en bas)

Item	Ind-Specif	Fq-Totale	Fq-Partie	
freedom	4.6	3	3	C
meaning	3.4	2	2	C
liberty	3.4	2	2	C
carried	2.9	3	2	C
men	2.7	4	2	C
women	2.7	4	2	C
unmatched	2.2	1	1	C
delivered	2.2	1	1	C
watchful	2.2	1	1	C
eyes	2.2	1	1	C
risk	2.2	1	1	C
eye	2.2	1	1	C
race	2.2	1	1	C
safely	2.2	1	1	C
horizon	2.2	1	1	C
fixed	2.2	1	1	C
embody	2.2	1	1	C
creed	2.2	1	1	C
labor	2.2	1	1	C
Its	2.2	1	1	C
themselves	2.2	1	1	C
join	2.2	1	1	C
grace	2.2	1	1	C
great	2.2	1	1	C
guardians	2.2	1	1	C
doers	2.2	1	1	C
control	2.2	1	1	C
celebrated	2.2	1	1	C
rugged	2.2	1	1	C
expand	2.2	1	1	C
magnificent	2.2	1	1	C
obscure	2.2	1	1	C
generate	2.2	1	1	C
takers	2.2	1	1	C

Corpus Freud annoté

- Le corpus *IP3* regroupe 3 textes :
 - *Einführung in die Psychoanalyse*, le texte original de Sigmund Freud (dorénavant : **EinfPsy**) ;
 - *Introduction à la psychanalyse*, première traduction en français, réalisée par Serge Jankélévitch et publiée en 1921, (dorénavant : **Intropsy**) ;
 - *Leçons d'introduction à la psychanalyse*, traduction publiée en 2000, dans le cadre de la publication des œuvres complètes de Freud par un groupe de psychanalystes français, sous la direction scientifique de Jacques Laplanche (dorénavant : **Leçons**).
- Corpus aligné et annoté (500 000 mots env.)

Einleitung.
Meine Damen und Herren! Ich weiß nicht, wieviel die einzelnen von Ihnen aus ihrer Lektüre oder vom Hörensagen über die Psychoanalyse wissen. Ich bin aber durch den Wortlaut meiner Ankündigung - Elementare Einführung in die Psychoanalyse - verpflichtet, Sie so zu behandeln, als wüßten Sie nichts und bedürften einer ersten Unterweisung. §

Introduction.
J'ignore combien d'entre vous connaissent la psychanalyse par leurs lectures ou par ouï-dire. Mais le titre même de ces leçons : Introduction à la Psychanalyse, m'impose l'obligation de faire comme si vous ne saviez rien sur ce sujet et comme si vous aviez besoin d'être initiés à ses premiers éléments. §

Introduction.
Mesdames et Messieurs, je ne sais pas ce que vous pouvez savoir, les uns et les autres, sur la psychanalyse par vos lectures ou par ouï-dire. Mais je suis tenu par l'intitulé annoncé - Introduction élémentaire à la psychanalyse - de vous traiter comme si vous ne saviez rien et que vous ayez besoin d'une première initiation. §

Corpus PARTUT

- *ParTUT (Multilingual Turin University TreeBank)*
 - Création et alignement phrastique de *treebanks* parallèles en italien, anglais et français à l'aide d'un seul schéma unifié
 - Les données *ParTUT* utilise le format CoNLL (*Conference on Computational Natural Language Learning*)
- Seules les ressources en français et en anglais ont été utilisées dans notre corpus.
 - Elles ont été transcodées dans un format permettant de définir une *base textométrique* intégrant une segmentation annotée
 - La *Trame* de cette base **ParTUT2Trameur** intègre 10 niveaux d'annotation et met en contraste les volets anglais et français au sein du *Cadre* organisé en deux parties (**EN** et **FR**)
 - L'alignement des phrases est matérialisé par le délimiteur «§»