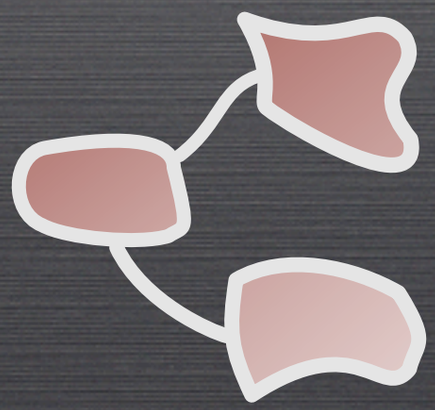
La plate-forme LinguaStream

Frédérik Bilhaut
GREYC - CNRS - Université de Caen



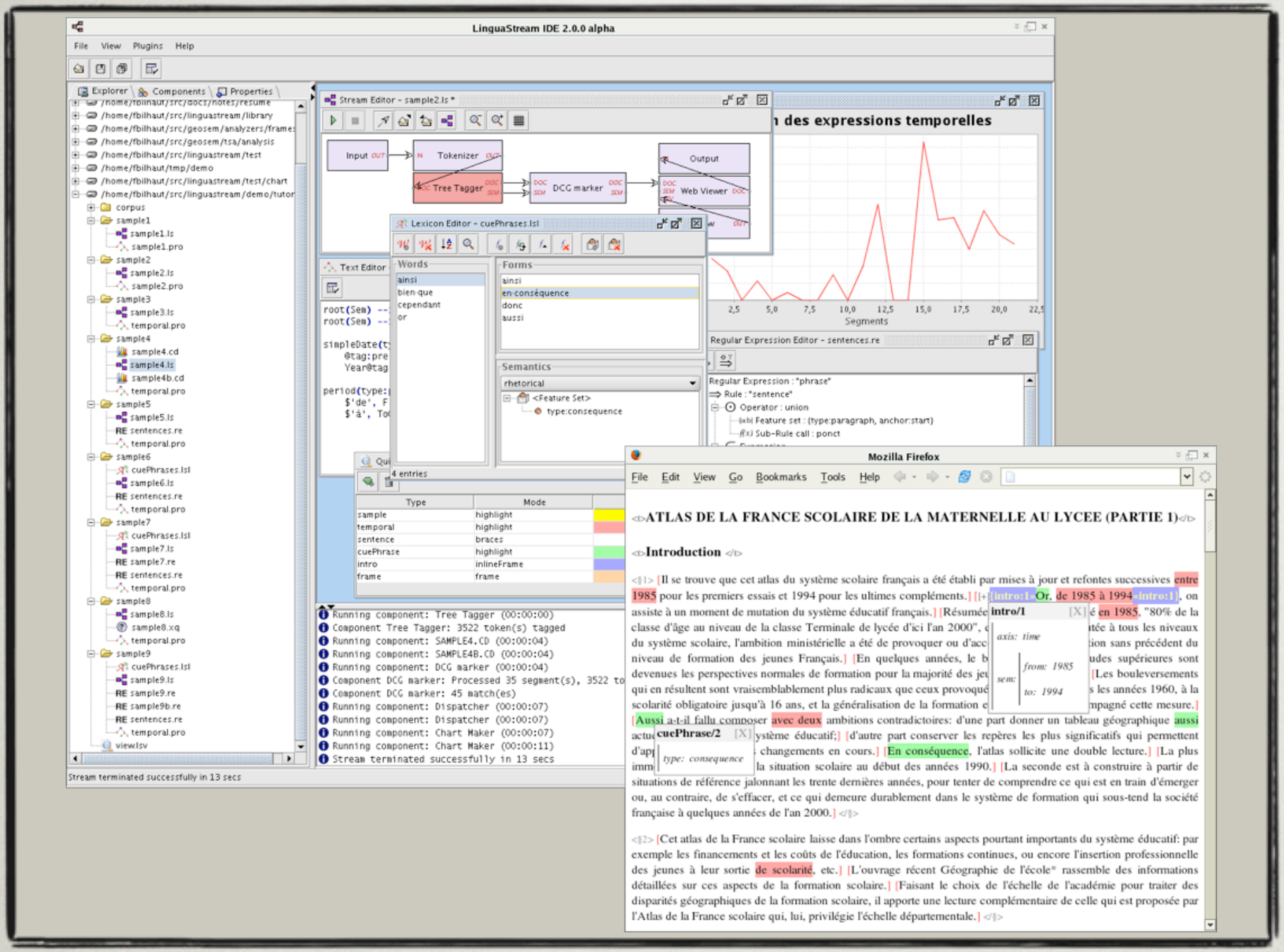
Problématique

- En TAL comme en linguistique, besoin de **“projeter”** automatiquement des **modèles** sur des **corpus**
  - A des fins de modélisation, observation, automatisation, etc.
- **Besoin d’outils** conceptuels et techniques pour :
  - Exprimer les modèles de façon purement formelle
  - Les appliquer automatiquement sur un corpus
  - Visualiser, mesurer, évaluer les résultats

“Nouveaux” besoins

- **Résultats probants** sur certaines tâches de TAL, que l’on souhaite pouvoir exploiter pour d’autres tâches
  - Articulation des traitements (cf. journée ATALA)
- De plus en plus de **documents** numériques **structurés**
  - Prise en compte de la structure du document
  - Transparence relativement aux différents formats existants
  - Gestion de différentes langues, alphabets, ...
- Besoin de **capitaliser et partager les “ressources”**
  - Lexiques, règles, annotations, chaînes de traitements, ...

- Volonté de réduire la distance entre **description** linguistique et **implémentation** informatique
  - Objectif : exprimer la plus grande part possible d'un modèle linguistique avec formalisme exploitable automatiquement
  - Exprimer **uniquement** des règles et des contraintes linguistiques
  - Les règles ont un rôle à la fois **descriptif** et **prescriptif**
- **Perspectives** "nouvelles" en TAL
  - Traitements sémantiques, niveau discursif, ...
  - Besoin d'utiliser et produire des annotations riches, d'injecter des connaissances de domaine, ...



La plate-forme LinguaStream

Un atelier d'expérimentation pour le TAL et d'exploration linguistique de corpus, développée au GREYC en collaboration avec l'ERSS.

Contributeurs

- GREYC (Caen) : Antoine Widlöcher (plate-forme, DSDL), Gérard Bécher, Arnaud Soulet, Stéphane Ferrari, Djamal Benallen, Patrice Enjalbert (dir), ...
- ERSS (Toulouse) : Marion Laignelet, Mai Ho-Dac, Christophe Pimm, Gérard Tozere, Marianne Vergez, Marie-Paule Pery-Woodley (dir), ...
- LIUPPA (Pau) : Mauro Gaio (dir) & co.

Plan de l'exposé

- Description générale de la plate-forme
  - Principes "fondateurs"
  - Modèle d'annotation
  - Articulation de traitements
- Modèles d'analyse
  - La notion de "perspective d'analyse"
- Visualisation et évaluation des résultats
- Gestion des chaînes de traitement
- Exemple d'application : cadres de discours temporels

Principes méthodologiques

- Elaboration de **chaînes de traitement**
  - Décomposition de traitements complexes en tâches élémentaires
  - Choix et enchaînement visuels des différents composants dédiés à ces tâches
- **Modularité et réutilisabilité** : isoler et capitaliser les traitements ± génériques et réutilisables
  - Composants dont la tâche est fixée (mais paramétrable)
  - Composants dont le comportement est à définir
    - Ecriture de règles dans différents formalismes

- Privilégier les **formalismes déclaratifs**
  - Leur pouvoir descriptif permet de capitaliser des connaissances linguistiques opérationalisables
- **Simplifier et automatiser** le processus de traitement
  - Sans coût de développement informatique
  - Pour se concentrer sur les modèles et traitements linguistiques
- Exploiter la **complémentarité des modèles d'analyse**
  - Différents modèles pour différents points de vue sur le texte
  - Permettre (et encourager) la cohabitation des formalismes
- Partage d'un **modèle d'annotation commun**

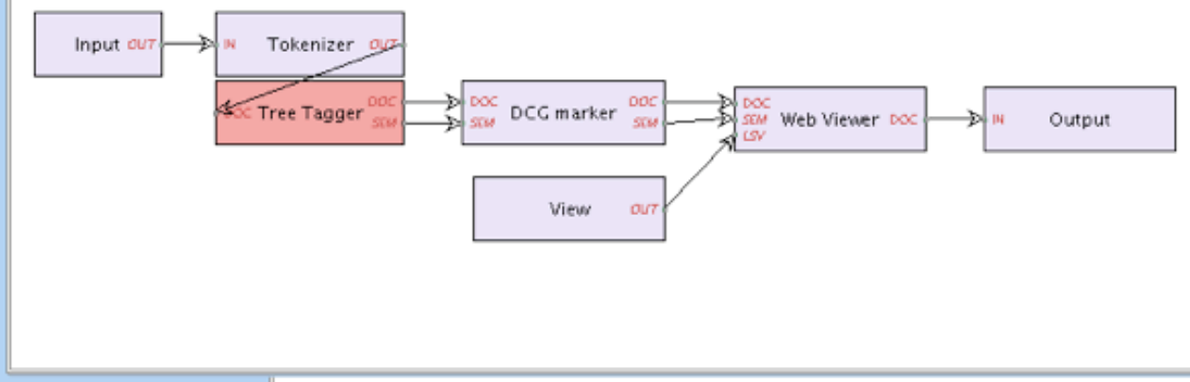
LinguaStream IDE 2.0.0 alpha

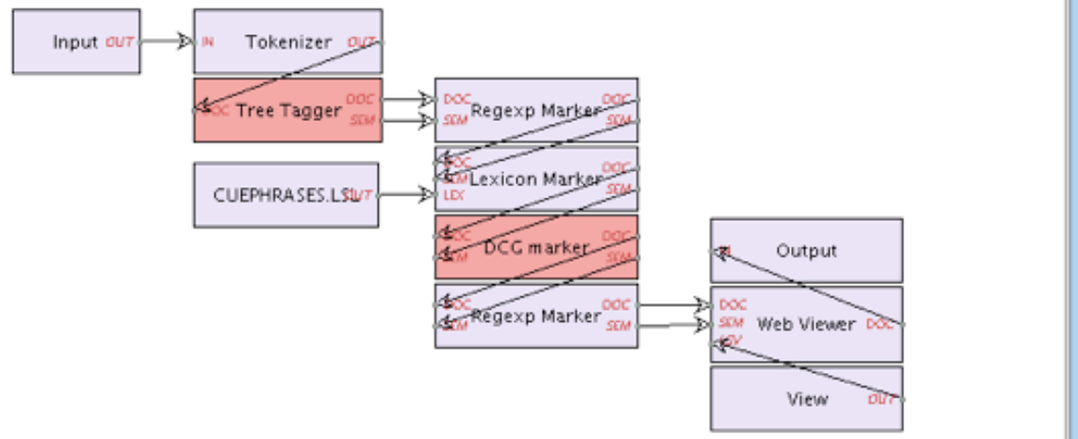
File View Plugins Help

Explorer Components Properties

Explorer
/home/fbilhaut/src/linguastream
/home/fbilhaut/src/geosem
/home/fbilhaut/tmp
/home/fbilhaut/src/docs/notes/resume
/home/fbilhaut/src/linguastream/library
/home/fbilhaut/src/geosem/analyzers/frames
/home/fbilhaut/src/geosem/tsa/analysis
/home/fbilhaut/src/linguastream/test
/home/fbilhaut/tmp/demo
/home/fbilhaut/src/linguastream/test/chart
/home/fbilhaut/src/linguastream/demo/tutorial
corpus
sample1
sample2
sample3
sample3.ls
temporal.pro
sample4
sample4.cd
sample4.ls
sample4b.cd
temporal.pro
sample5
sample5.ls
sentences.re
temporal.pro
sample6
cuePhrases.lsl
sample6.ls
sentences.re
temporal.pro
sample7
cuePhrases.lsl
sample7.ls
sample7.re
sentences.re
temporal.pro
sample8
sample9
view.lsv

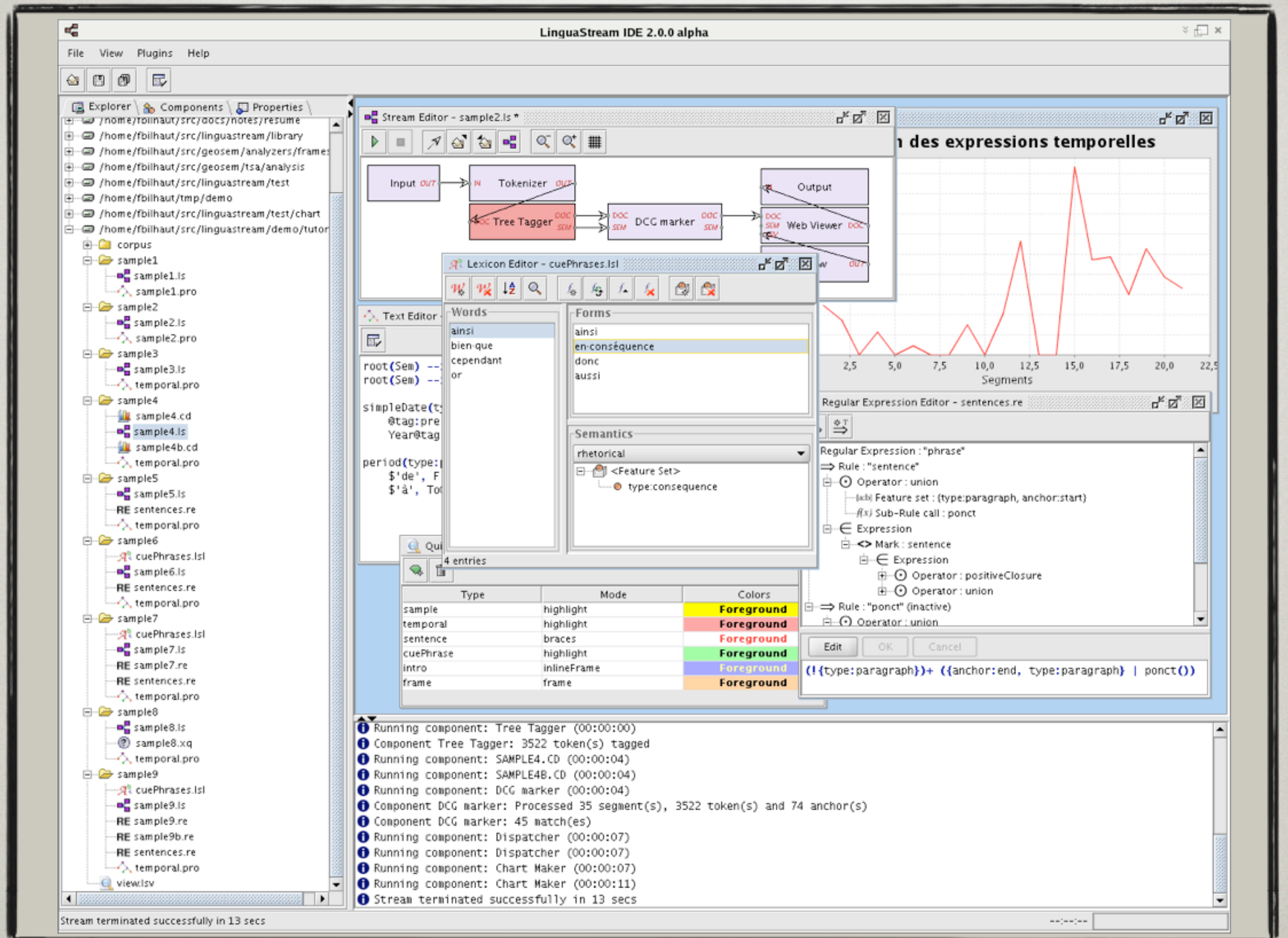
Stream Editor - sample3.ls *





Plugin 'SWI Prolog' v. 2.0.0 successfully loaded
Plugin 'Tree Tagger' v. 2.0.0 successfully loaded
Plugin 'GeoSem' v. 0.6.1 successfully loaded
Plugin 'Lexicon Marker' v. 2.0.0 successfully loaded
Plugin 'RegExp' v. 2.0.0 successfully loaded
Plugin 'Markup' v. 2.0.0 successfully loaded
Plugin 'Python Marker' v. 2.0.0 successfully loaded
Plugin 'Token Marker' v. 2.0.0 successfully loaded
Plugin 'Rule Marker' v. 1.0.0 successfully loaded
Plugin 'Markup Charter' v. 2.0.0 successfully loaded
File /home/fbilhaut/src/linguastream/demo/tutorial/sample3/sample3.ls saved.
File /home/fbilhaut/src/linguastream/demo/tutorial/sample3/sample3.ls saved.

File /home/fbilhaut/src/linguastream/demo/tutorial/sample3/sample3.ls saved.


LinguaStream IDE 2.0.0 alpha
File View Plugins Help
Explorer Components Properties
/home/fbilhaut/src/docs/notes/resume
/home/fbilhaut/src/linguastream/library
/home/fbilhaut/src/geosem/analyzers/frames
/home/fbilhaut/src/geosem/tsa/analysis
/home/fbilhaut/src/linguastream/test
/home/fbilhaut/tmp/demo
/home/fbilhaut/src/linguastream/test/chart
/home/fbilhaut/src/linguastream/demo/tutor
corpus
sample1
sample1.ls
sample1.pro
sample2
sample2.ls
sample2.pro
sample3
sample3.ls
temporal.pro
sample4
sample4.cd
sample4.ls
sample4b.cd
temporal.pro
sample5
sample5.ls
sentences.re
temporal.pro
sample6
cuePhrases.lsl
sample6.ls
sentences.re
temporal.pro
sample7
cuePhrases.lsl
sample7.ls
sample7.re
sentences.re
temporal.pro
sample8
sample8.ls
sample8.xq
temporal.pro
sample9
cuePhrases.lsl
sample9.ls
sample9.re
sample9b.re
sentences.re
temporal.pro
view.lsv
Stream Editor - sample2.ls *
Input
Tokenizer
Tree Tagger
DCG marker
Output
Web Viewer
des expressions temporelles
2,5 5,0 7,5 10,0 12,5 15,0 17,5 20,0
Segments
Lexicon Editor - cuePhrases.lsl
Words
ainsi
bien-que
cependant
or
Forms
ainsi
en-conséquence
donc
aussi
Semantics
rhetorical
<Feature Set>
type:consequence
4 entries
Type Mode Colors
sample highlight Foreground
temporal highlight Foreground
sentence braces Foreground
cuePhrase highlight Foreground
intro inlineFrame Foreground
frame frame Foreground
Regular Expression Editor - sentences.re
Regular Expression : "phrase"
Rule : "sentence"
Operator : union
Feature set : (type:paragraph, anchor:start)
Sub-Rule call : ponct
Expression
Mark : sentence
Expression
Operator : positiveClosure
Operator : union
Rule : "ponct" (inactive)
Operator : union
Edit OK Cancel
(!{type:paragraph})+ ({anchor:end, type:paragraph} | ponct())
Running component: Tree Tagger (00:00:00)
Component Tree Tagger: 3522 token(s) tagged
Running component: SAMPLE4.CD (00:00:04)
Running component: SAMPLE4B.CD (00:00:04)
Running component: DCG marker (00:00:04)
Component DCG marker: Processed 35 segment(s), 3522 token(s) and 74 anchor(s)
Component DCG marker: 45 match(es)
Running component: Dispatcher (00:00:07)
Running component: Dispatcher (00:00:07)
Running component: Chart Maker (00:00:07)
Running component: Chart Maker (00:00:11)
Stream terminated successfully in 13 secs
Stream terminated successfully in 13 secs

LinguaStream IDE 2.0.0 alpha

File View Plugins Help

Properties
Explorer Components

- Components
  - Structure
  - Transformation
  - Query
  - Markup
    - Lexical
    - Analysis
      - DCG Marker
      - BeanShell Marker
      - Groovy Marker
      - Lexicon Marker
      - Regexp Marker
      - Rule Marker
    - Tools
    - Statistics
    - Scripting
  - External
  - Statistics
  - View
  - Templates
  - GeoSem
  - TTSA

Lexicon Editor - small.lsl

Words

Nord-Pas-de-Calais
Haute-Normandie
Basse-Normandie
Picardie
Ile-de-France
Champagne
Lorraine
Alsace
Bretagne
Pays-de-Loire
Poitou-Charentes
Aquitaine
Midi-Pyrénées
Languedoc-Roussillon
Alpes-Provence-Côte-d'Azur
Corse
Rhône-Alpes
Auvergne
Limousin
Centre
Bourgogne
France-Comté
Aisne

Forms

Basse-Normandie
Basse-Normandie

Semantics

np

- <Feature Set>
  - type_sem:egn
  - type_zone:région
  - nom:Basse-Normandie

105 entries

Plugin 'Token Marker' v. 2.0.0 successfully loaded
Plugin 'RegExp' v. 2.0.0 successfully loaded
Plugin 'Python Marker' v. 2.0.0 successfully loaded
Plugin 'Markup' v. 2.0.0 successfully loaded
Plugin 'GeoSem' v. 0.6.1 successfully loaded
Plugin 'Rule Marker' v. 1.0.0 successfully loaded
Plugin 'Markup Charter' v. 2.0.0 successfully loaded
File /home/fbilhaut/src/fac/data/geosem/lexiques/small.lsl saved.

File /home/fbilhaut/src/fac/data/geosem/lexiques/small.lsl saved. --:--:--

LinguaStream IDE 2.0.0 alpha

File View Plugins Help

Properties

Explorer Components

Colors

backgroundColor

color

General

mode Highlight (blocks only

type sn

Standard Expert

Quick View Editor - view.lsv

Type	Mode	Colors
sn	highlight	Foreground
phrase	braces	Foreground
introducer	inlineFrame	Foreground
mse	highlight	Foreground
introducer	inlineFrame	Foreground
ttsa	frame	Foreground
ttsa		
ttsa		
ttsa		
ttsa		
ttsa		
cuePhrase		
temporal		
ttsa		
coord		

Stream

Running co
Running co
Running co
Running co
Running co
Running co
Stream ter

Stream terminated successfully in 13 secs

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

file:///home/fbilhaut/

aux concours.] [La tendance [a] été inversée au début des années 1980, entraînant la diminution du ta du recrutement [expliquent] la surreprésentation actuelle d

temporal:10 [X]

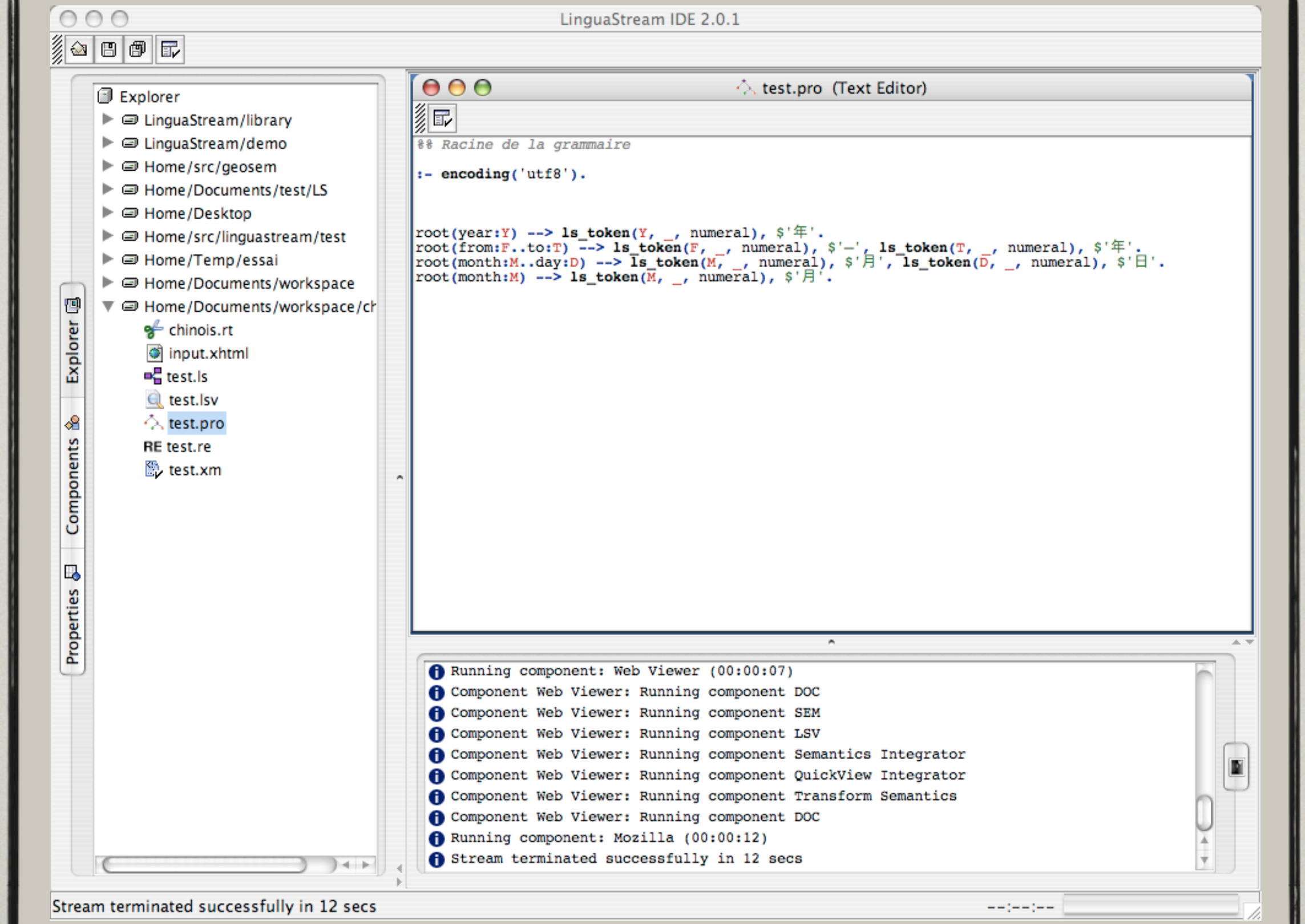
- *type: d*
- *periode:*
  - *debut:* *annees:*
    - *type: debut*
    - *annee: 1960*

980<intro:1], l'offre de recrutement (bien qu'en forte a ndidats se présentant aux concours, ne [sont] plus à la h fait, la proportion de maîtres auxiliaires, personnels non ti lans les lycées professionnels, [qui [attirent] le moins], [ ];] [[puis] dans les lycées, alors [que la stagnation des e présence dans les collèges].] [Ce [sont] actuellement les lycées [qui [absorbent] l'essentiel de l'augmentation des effectifs enseignants].]

[(*)[intro:2>En trente ans<intro:2], le corps enseignant du second degré s'[est] féminisé, particulièrement dans les collèges et dans les grades les moins élevés.] [Le taux de féminité s'[est] stabilisé autour de 55% depuis une dizaine d'années: 42% en lycée professionnel, 50% en lycée, 61% en collège.]

[(*)[intro:3>Depuis le début des années 1960<intro:3], la composition du corps enseignant [a] été diversifiée: les disciplines, catégories indiciaires et grades [ont] été multipliés, avec l'apparition de CAPES et de CAPET artistiques et techniques et la création de certains corps enseignants (PEGC et plus récemment, PLP1 et PLP2).] [Même si la tendance actuelle [est] à la simplification, les grades

javascript:toggleSemantics('temporal_10')


LinguaStream IDE 2.0.1
Explorer
LinguaStream/library
LinguaStream/demo
Home/src/geosem
Home/Documents/test/LS
Home/Desktop
Home/src/linguastream/test
Home/Temp/essai
Home/Documents/workspace
Home/Documents/workspace/ch
chinois.rt
input.xhtml
test.ls
test.lsv
test.pro
RE test.re
test.xm
Explorer
Components
Properties
test.pro (Text Editor)
%% Racine de la grammaire
:- encoding('utf8').
root(year:Y) --> ls_token(Y, _, numeral), $'年'.
root(from:F..to:T) --> ls_token(F, _, numeral), $'–', ls_token(T, _, numeral), $'年'.
root(month:M..day:D) --> ls_token(M, _, numeral), $'月', ls_token(D, _, numeral), $'日'.
root(month:M) --> ls_token(M, _, numeral), $'月'.
Running component: Web Viewer (00:00:07)
Component Web Viewer: Running component DOC
Component Web Viewer: Running component SEM
Component Web Viewer: Running component LSV
Component Web Viewer: Running component Semantics Integrator
Component Web Viewer: Running component QuickView Integrator
Component Web Viewer: Running component Transform Semantics
Component Web Viewer: Running component DOC
Running component: Mozilla (00:00:12)
Stream terminated successfully in 12 secs
Stream terminated successfully in 12 secs
--:--:--

Document en chinois

1981年

6月，中共十一届六中全会通过邓小平主持起草的《关于建国以来党的若干历史问题的决议》。决议彻底否定了“文化大革命”，全面评价了毛泽东 的历史地位，提出必须坚持和发展毛泽东思想。会议选举邓小平为中央军委主席。

7月2日，在中共省、自治区、直辖市委员会书记座谈会上讲话提出，老干部第一位的任务是选拔中青年干部。

8月，视察新疆。

9月19日，在华北某地检阅军事演习部队，讲话时提出，要建设强大的现代化正规化的革命军队。

1982年

4月10日，在中共中央政治局会议上讲话，提出坚持社会主义道路的四项必要保证：体制改革；建设社会主义精神文明；打击经济犯罪活动； 整顿党的作风和党的组。强调一手坚持对外开放和对内搞活经济的政策，一手坚决打击经济犯罪活

date/7 [X]

month: 4

day: 10

5月 列比里亚国家元首多伊。谈话时说，我们一方面实行开放政策，一方面仍坚持自力更生为主的方针。

8月21日，会见联合国秘书长德奎利亚尔。谈话时重申，中国是第三世界的一员。反对霸权主义，维护世界和平是中国对外政策的纲领。

9月1日，在中国共产党第十二次全国代表大会上致开幕词，提出建设有中国特色社会主义的主题。

9月12日至13日，中共十二届一中全会召开，选举邓小平为中央政治局常务委员，决定他任中央军委主席。

9月13日，在中共中央顾问委员会第一次全体会议上，当选为中央顾问委员会主任。

9月18日，陪同朝鲜劳动党中央委员会总书记金日成去四川访问。

9月24日，会见英国首相撒切尔夫人，阐述中国对香港问题的基本立场，为以后中英两国政府的谈判定了基调。

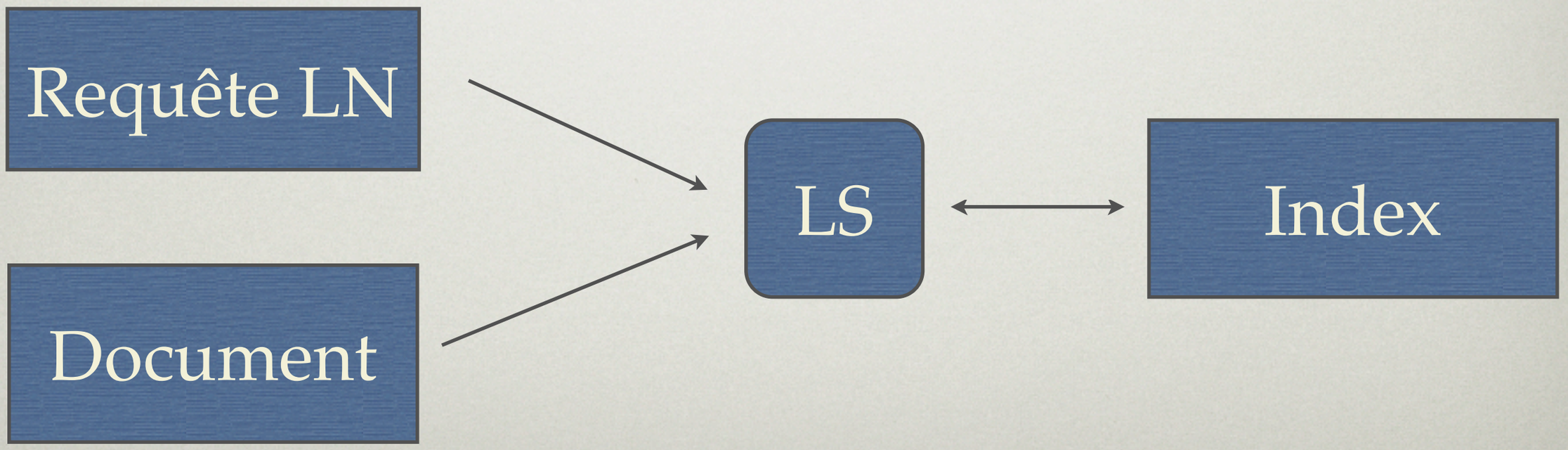
1983年

1月12日，同国家计委、国家经委和农业部门负责人谈话时指出，各项工作都要有助于建设有中国特色的社会主义，并强调农业是根本，不忘掉。

Done

Applications

- La plate-forme est avant tout un “laboratoire” pour le TAL, et un outil d’aide à l’observation linguistique
- Elle peut également être utilisée dans un cadre applicatif (Ex : prototype moteur de recherche GeoSem)



Modèle d'annotation

Modèle d'annotation

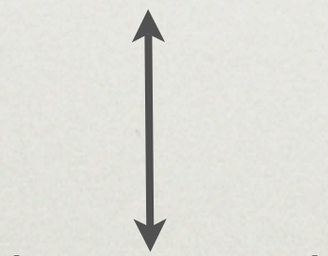
accuser . Les croissances ont été générales dans la France du Nord et du Nord-Ouest , ainsi que dans le quart sud-est du pays , Rhône-Alpes et régions méditerranéennes ; dans une quinzaine de départements , la grande banlieue parisienne , la Bretagne orientale et ses marges , la région Rhône-Alpes , etc. , l' augmentation de la population scolaire en vingt ans a été de plus du quart des effectifs de 1965 . En revanche , les chiffres sont stables , ou même en diminution dans la France des petits effectifs scolaires . Sont également en baisse les départements lorrains . Ces évolutions reflètent certes très directement celles de l' ensemble de la population au cours de la même période , mais en les accentuant . L' augmentation des effectifs scolaires a été sensiblement plus marquée que celle de la

$$\begin{bmatrix} graphie : 'effectifs' \\ tag : nom \\ lemme : 'effectif' \end{bmatrix}$$

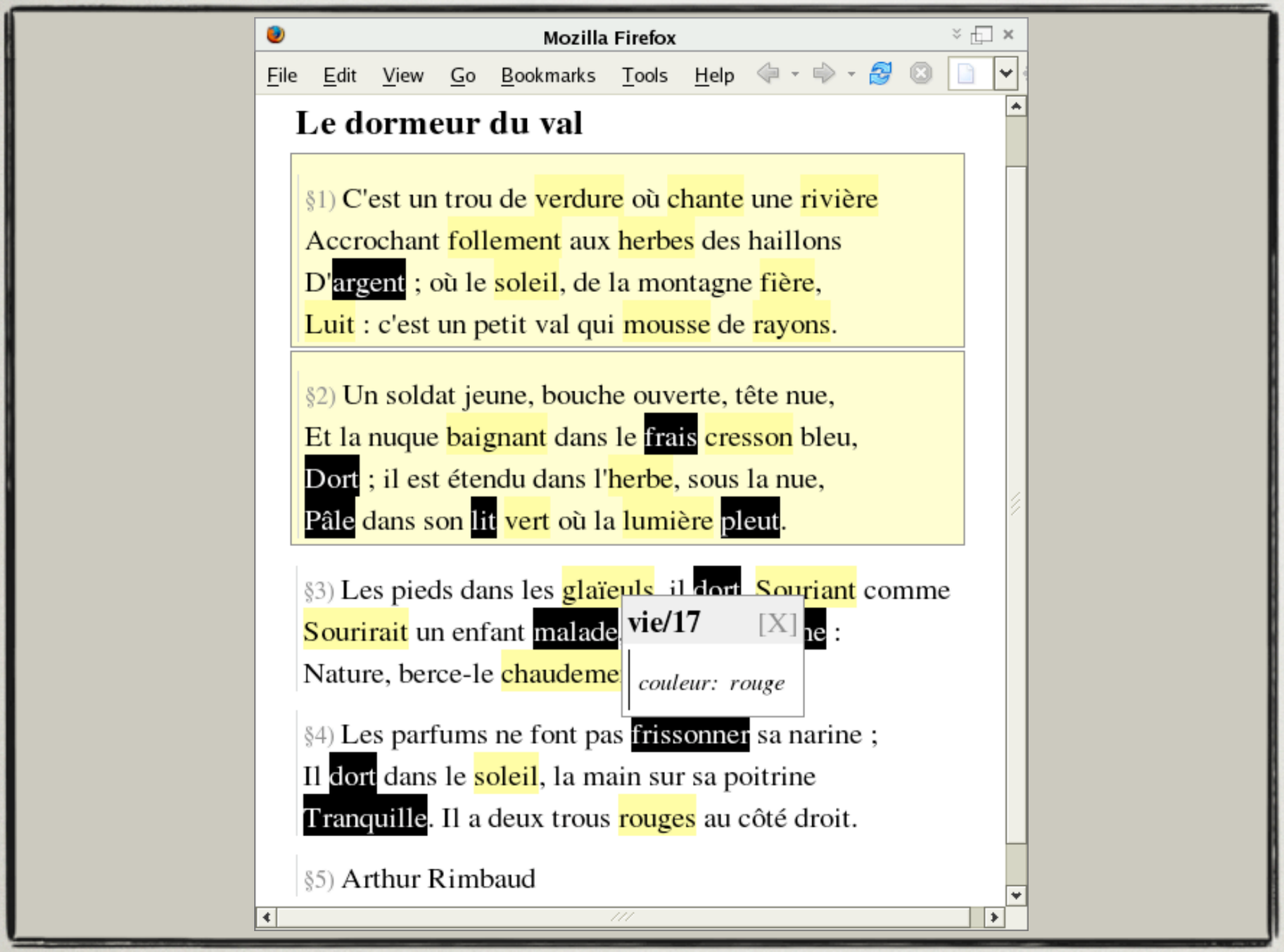
$$\begin{bmatrix} enum : \begin{bmatrix} zone : \begin{bmatrix} egn : \begin{bmatrix} type : banlieue \\ clef : Paris \end{bmatrix} \end{bmatrix} \\ zone : \begin{bmatrix} egn : \begin{bmatrix} type : région \\ clef : Bretagne \end{bmatrix} \\ loc : interne \\ geo : oriental \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Relations

accuser . Les croissances ont été générales dans la France du Nord et du Nord-Ouest , ainsi que dans le quart sud-est du pays , Rhône-Alpes et régions méditerranéennes ; dans une quinzaine de départements , la grande banlieue parisienne , la Bretagne orientale et ses marges , la région Rhône-Alpes , etc. , l' augmentation de la population scolaire en vingt ans a été de plus du quart des effectifs de 1965 . En revanche , les chiffres sont stables , ou même en diminution dans la France des petits effectifs scolaires . Sont également en baisse les départements lorrains . Ces évolutions reflètent certes très directement celles de l' ensemble de la population au cours de la même période , mais en les accentuant . L' augmentation des effectifs scolaires a été sensiblement plus marquée que celle de la

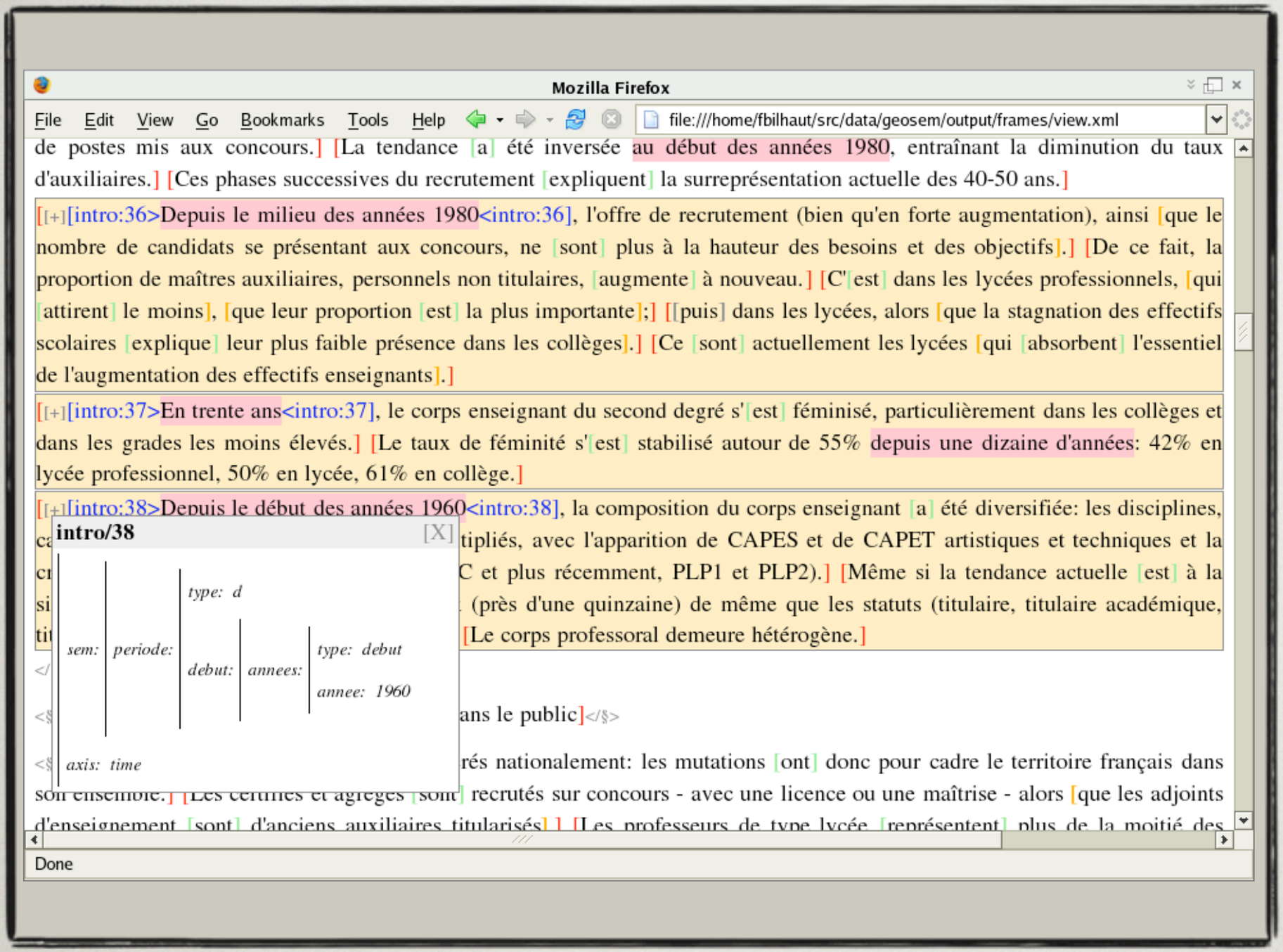


[*type :'sujet '*]



Exemple de document annoté

“Le dormeur du val” : annotation à partir d’un lexique sémantique.

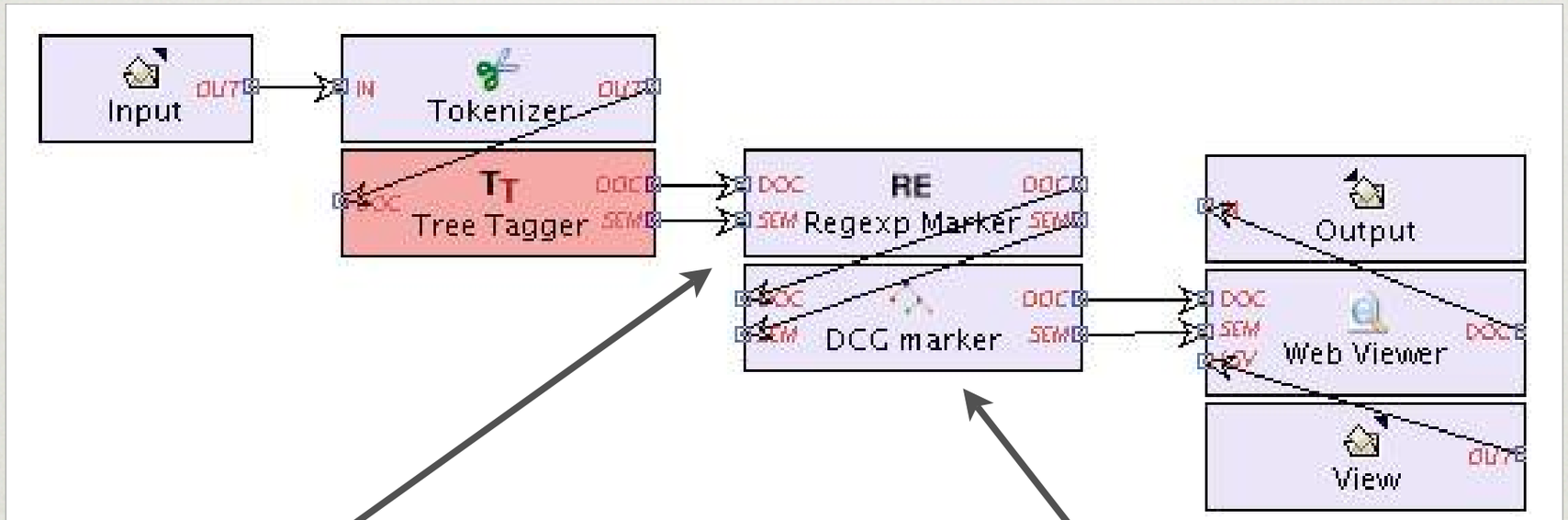


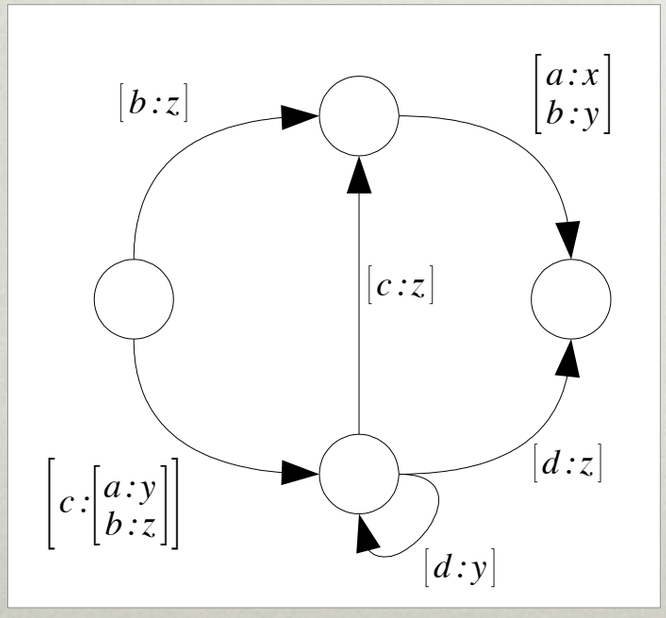
Exemple de document annoté

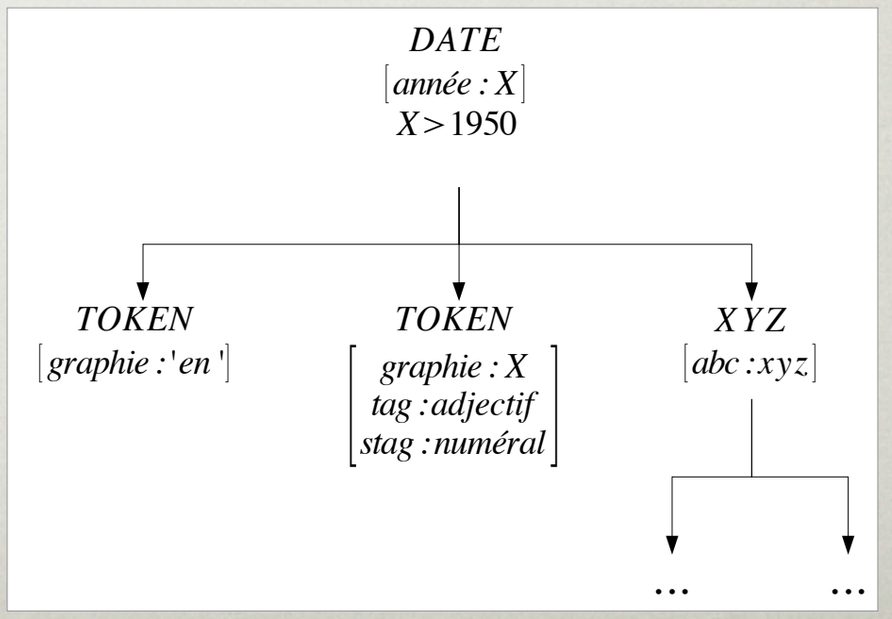
“Atlas scolaire de la maternelle au lycée” : annotation des cadres de discours temporels.

Articulation des traitements

Chaînes de traitement







Principes d'articulation

- Tous les composants partagent la même représentation...
  - ... des documents annotés :
    - Le texte est balisé par différents marquages
    - Un marquage contient un nombre arbitraire de marques
    - Chaque marquage est caractérisé par un type
    - Chaque marque identifie un fragment textuel de taille arbitraire
    - Les chevauchements sont autorisés
  - ... des annotations elles-mêmes :
    - À chaque marque peut être associée une structure de traits

- Tous les modèles d'analyse les exploitent uniformément
- Chaque composant peut exploiter les marquages antérieurs
- Abstraction progressive des formes de surface
- Par expression de contraintes :
  - Sur la forme des unités (marques) préalablement balisées
  - Sur les structures de traits qui leur sont associées (par unification)

Technologies XML

- Utilisation systématique du format XML
  - Pour les documents et les annotations
  - Pour tous les autres objets manipulés (chaînes de traitement, règles, etc.)
- Tout document XML peut être traité par la plate-forme
  - Un espace de nom spécifique est utilisé
  - Utilisation d'outils standards pour la visualisation, l'annotation manuelle, etc.

- Mode d'annotation mixte :
  - Le balisage est inséré dans le document (*inline*)
    - Les espaces de noms garantissent la transparence pour les autres applications
    - Si nécessaire, les chevauchements sont gérés à l'aide de jalons (*milestones*)
  - Les annotations sont stockées de façon externe (*stand-off*)
    - Représentation XML des structures de traits
    - Possibilité d'export RDF
  - Tout l'outillage XML "standard" est disponible sous forme de composants XSLT, XQuery, XSL:FO, etc...

Modèles d'analyse

Définition

- Modèle d'analyse :
  - Formalisme (± générique) de représentation de règles
  - Algorithme d'application des règles aux textes
- Des notions sur l'aspect algorithmique sont facultatives, mais utiles...

Grammaires d'unification (EDCG)

- Grammaires à Clauses Définies Étendues (EDCG)
  - Grammaires locales
  - Chaque non-terminal définit un patron tout en lui associant une représentation symbolique
- Expression de contraintes par unification :
  - Spécification de contraintes sur les unités du texte
  - Construction de représentations symboliques
- Basées sur Prolog
  - Les fonctionnalités logiques du langage sont disponibles
    - Ex : calculs sémantiques complexes

EDCG

```
date(type:date..mois:M) --> @tag:pre, mois(M).
date(type:date..année:A) --> ...
...
mois(1) --> $'janvier'.
mois(2) --> $'février'.
...
```

Macro-expressions régulières (MRE)

- Automates finis déterministes
  - Les patrons à reconnaître sont spécifiés à l'aide d'un langage analogue aux expressions régulières
  - Spécification de contraintes sur les marquages antérieurs par le biais de structures de traits
  - Spécification libre de la ou des zones à marquer
  - Plus performant (temps linéaire en pratique) mais moins "expressis" que les grammaires EDCG

MRE

```
{type:phrase, anchor:start}
<introducteur>
    {type:connecteur}? {tag:pre} {type:temporel} /as $t
</introducteur> /sem {axe:temps, valeur:$t}
" ,"
```

Discourse Structure Description Language

- Conçu par Antoine Widlöcher (GREYC)
- Formalisme dédié à l'analyse du discours
  - Niveau de granularité potentiellement élevé
  - Variété combinatoire importante
  - Descriptif : représentation formelle d'un phénomène discursif
  - Prescriptif : paramétrer un analyseur, interroger le texte
- Définition de « grammaires de discours »
  - Permettant l'expression de contraintes :
    - non séquentielles (la notion d'ordre ne prédomine pas)
    - non linéaires (acceptant des discontinuités)

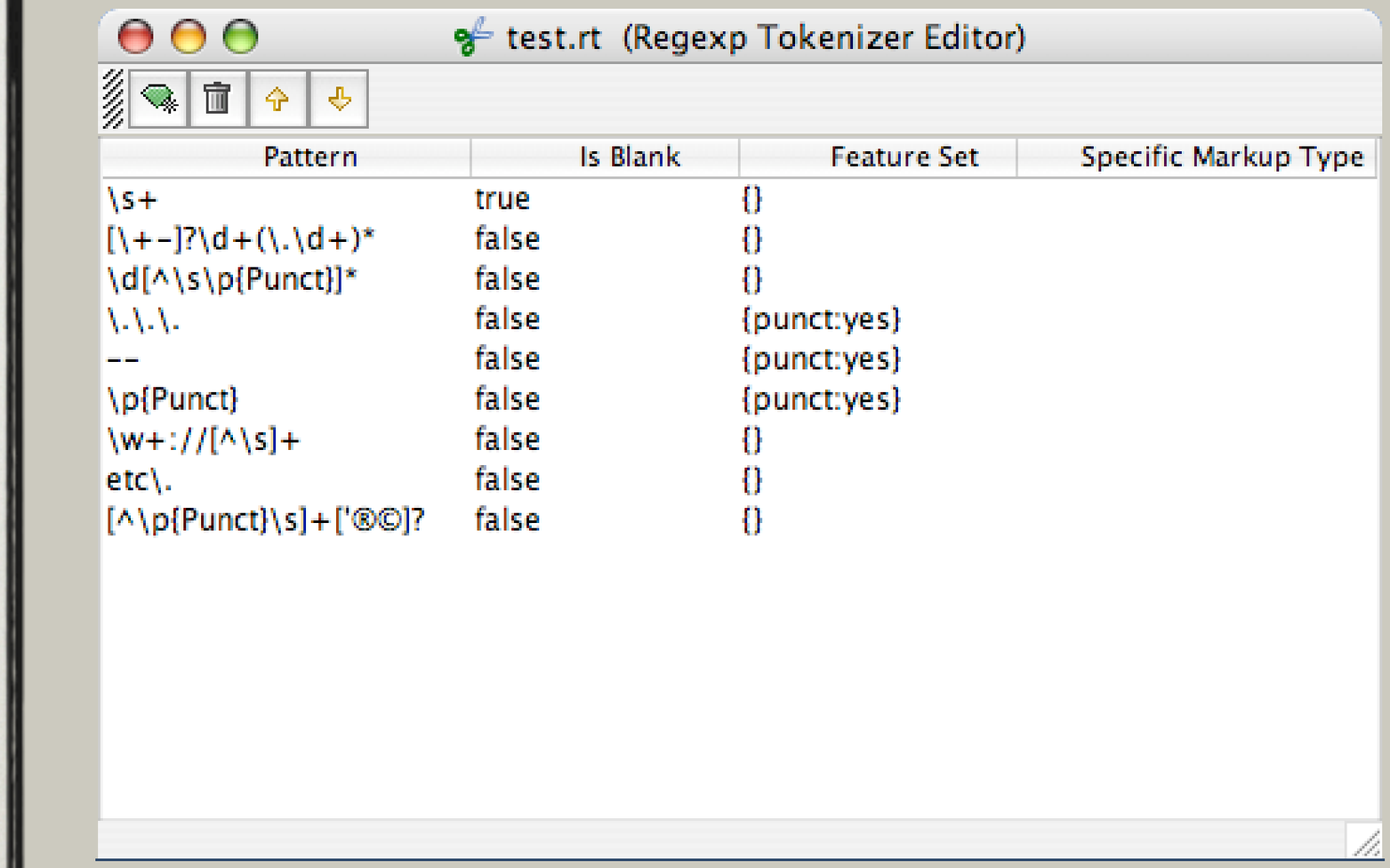
- Exemples de contraintes DSDL :
  - le segment recherché contient plus de X% d'expressions de type Y
  - il doit être homogène selon un critère donné
  - il doit commencer / terminer par un élément de type X
  - il doit être le plus long / court possible
  - il doit contenir X éléments en relation de type Y avec un autre segment de type Z
  - ...

DSDL

```
Rule {type:"cadre"}
{
    start({type: "introducteur"})
    end({type : "phrase"})
    homogeneity(comparator:portée)
    size(mode:"longest")
    not presence(pattern : {type : "introducteur"}, amount : 2)
    size(mode : #LONGEST)
}

Comparator portée ({type: "verbe"} as $v1, {type: "verbe"} as $v2)
{
    $v1/temps = $v2/temps
}

Comparator portée ({type: "intro"} as $i, {type: "tempo"} as $t)
{
    (($i/debut >= $t/debut) and ($i/debut <= $t/fin))
  or
    (($i/fin >= $t/debut) and ($i/fin <= $t/fin))
}
```

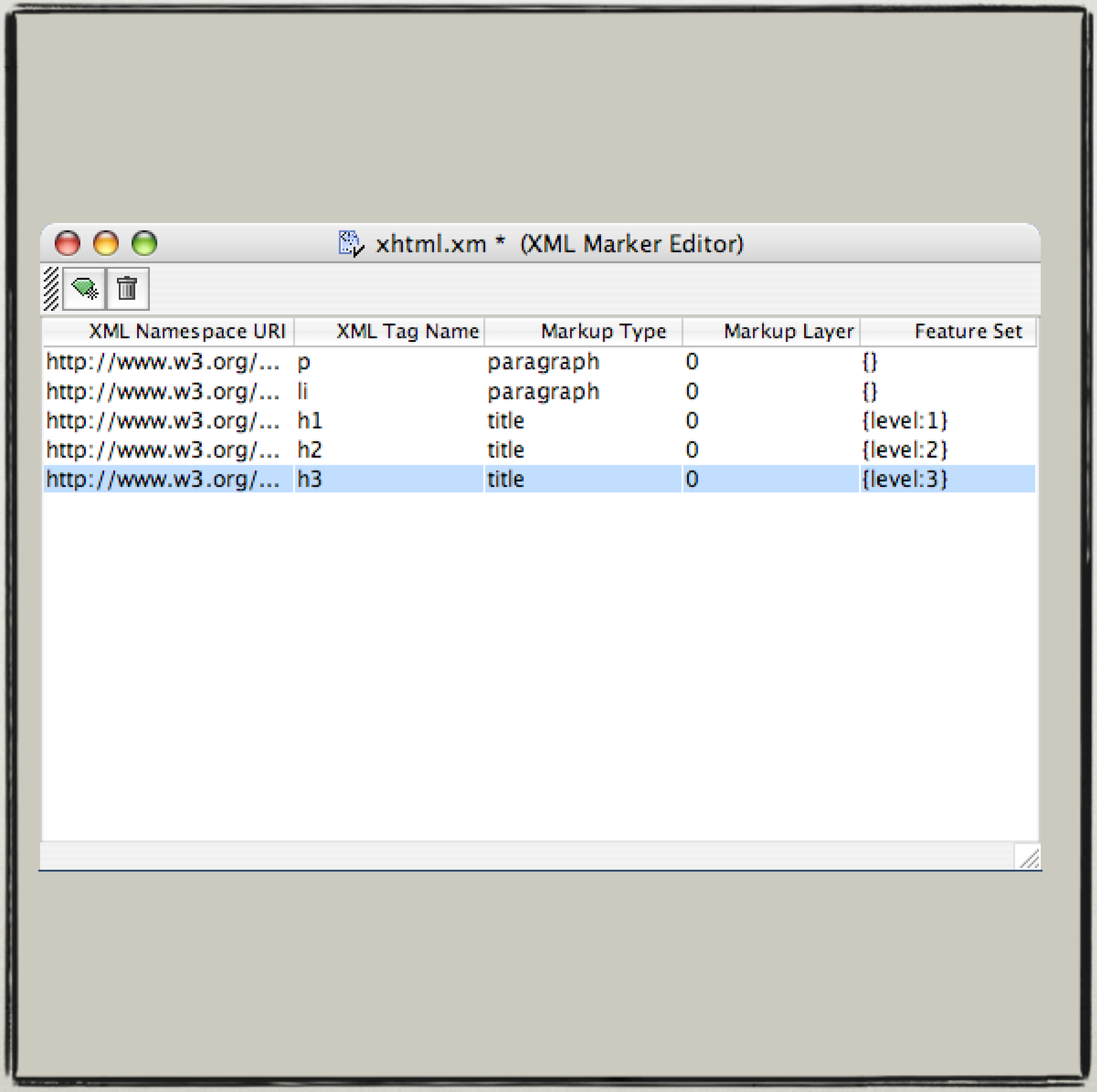
Expressions régulières "classiques"

- Patrons au niveau des caractères
  - À partir d'expressions régulières (type Perl)
- Deux modes d'utilisation :
  - Marquage d'unités élémentaires (ex : mots)
    - Permet de remplacer le tokeniseur "par défaut"
  - Attribution de traits sémantiques à des marquages antérieurs, selon des critères purement formels


test.rt (Regexp Tokenizer Editor)
Pattern
Is Blank
Feature Set
Specific Markup Type
\s+ true {}
[\+-]?\d+(\.\d+)* false {}
\d[^\s\p{Punct}]* false {}
\.\.\. false {punct:yes}
-- false {punct:yes}
\p{Punct} false {punct:yes}
\w+://[^\s]+ false {}
etc\. false {}
[^\p{Punct}\s]+['®©]? false {}

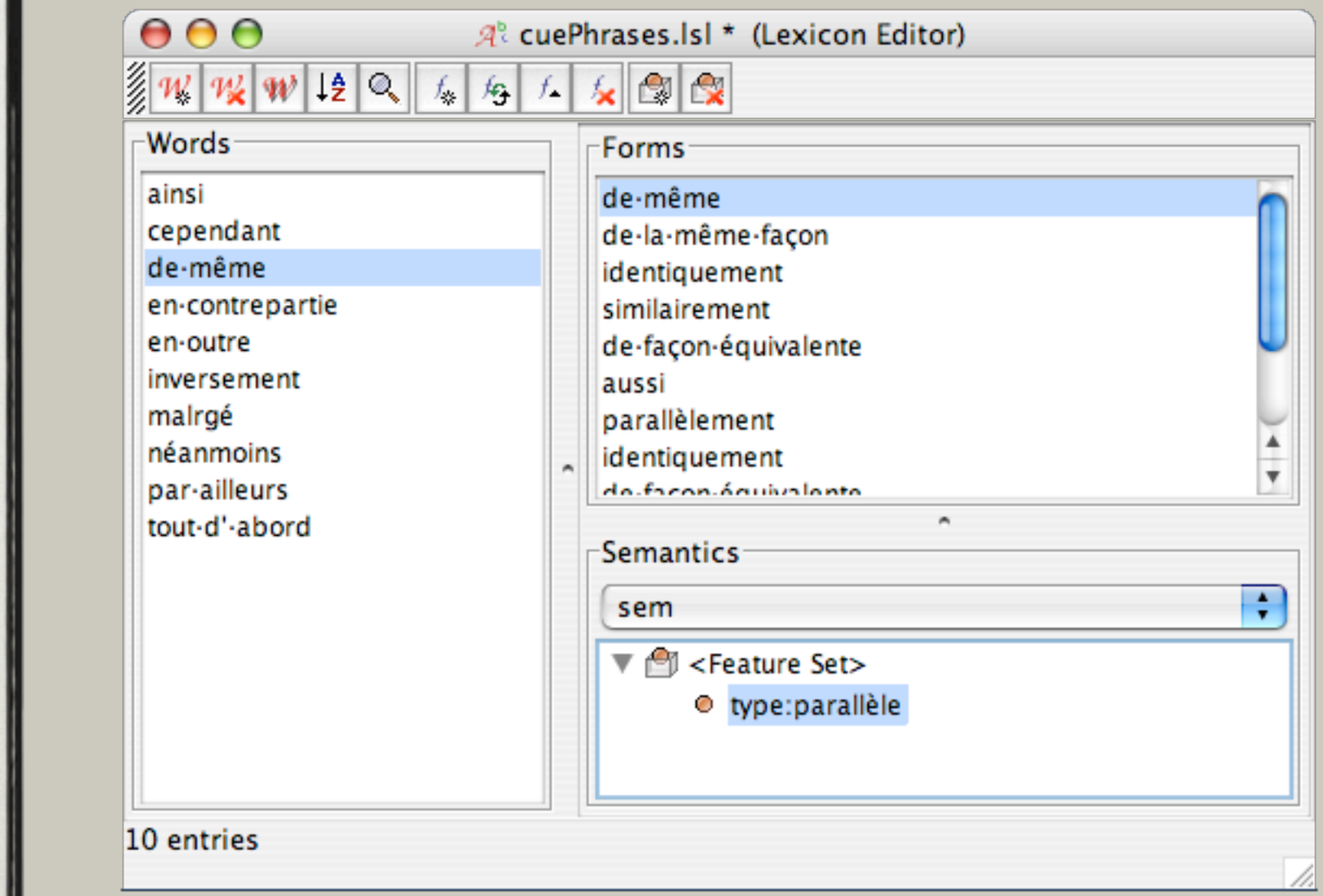
Exploitation de la structure XML

- Les analyseurs ne “voient” que les annotations LS
  - Mais il est possible d’écrire des règles permettant d’ajouter ces marques à partir de règles portant sur la structure XML
    - Ex. en XHTML : p/li → paragraphe, h1 → titre [niveau : 1]
- Le passage de balises XML à des annotations correspond à un mode d’analyse parmi d’autres
- Il permet de s’abstraire du schéma XML effectivement utilisé (la suite du traitement ne prendra en compte que les annotations LS)


xhtml.xm * (XML Marker Editor)
XML Namespace URI
XML Tag Name
Markup Type
Markup Layer
Feature Set
http://www.w3.org/... p paragraph 0 {}
http://www.w3.org/... li paragraph 0 {}
http://www.w3.org/... h1 title 0 {level:1}
http://www.w3.org/... h2 title 0 {level:2}
http://www.w3.org/... h3 title 0 {level:3}

Autres modèles d'analyse

- Projection de lexiques sémantiques
  - Interface graphique d'édition de lexiques
  - Composant de marquage dans les documents
  - En développement : schémas et ontologies
- Analyseur de type « système expert »
  - Pseudo-grammaires formées de règles de déduction
  - Basé sur un moteur moteur d'inférence type CLIPS


cuePhrases.lsl * (Lexicon Editor)
Words
ainsi
cependant
de-même
en-contrepartie
en-outre
inversement
malrgé
néanmoins
par-ailleurs
tout-d'-abord
Forms
de-même
de-la-même-façon
identiquement
similairement
de-façon-équivalente
aussi
parallèlement
identiquement
Semantics
sem
<Feature Set>
type:parallèle
10 entries

- Langages de scripts
  - Plusieurs langages sont disponibles pour créer des analyseurs répondant à un tâche très spécifique
  - Python, BeanShell, Groovy
- Développement de composants tiers
  - La plate-forme offre une API permettant de développer ses propres composants.
  - Un système de plugins permet de les distribuer séparément et de les charger dynamiquement

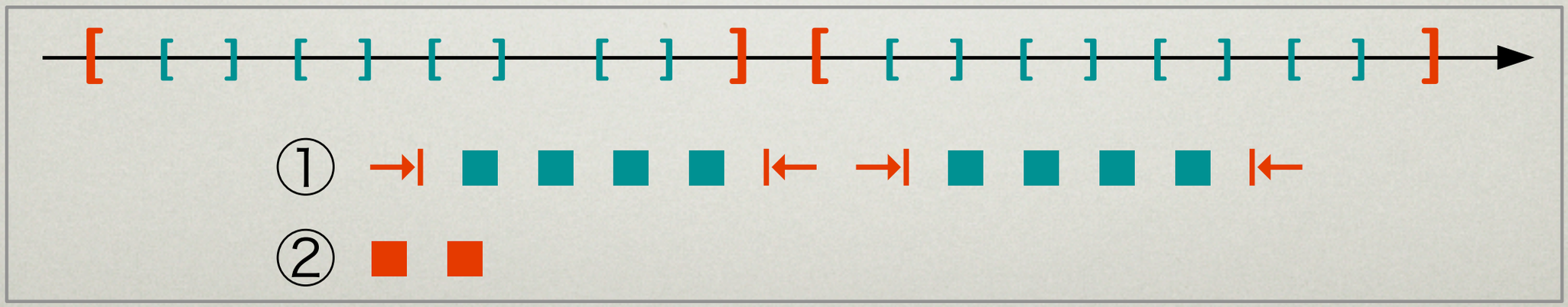
Perspectives d'analyse

La notion de "perspective d'analyse"

- Selon les modèles d'analyse différents types de "grain" doivent être définis :
  - Unité minimale (jeton ou *token*)
  - Espace de recherche (modèles non déterministes)
  - Unités à ignorer, etc.
- La définition de ces grains forme une perspective (ou point de vue) particulière sur un document
- La plate-forme permet de spécifier, **pour chaque module d'analyse**, la perspective à adopter

Définition du grain minimal d'analyse

- Unité minimale sur laquelle les patrons agissent (token)
  - Typiquement : le mot
- Chaque composant définit localement le(s) type(s) de marquages qui doivent être considérés comme tokens
  - Chaque occurrence de l'un des types spécifiés forme un token
  - Les autres marquages seront vus comme de simples « jalons »



Autres composantes d'une perspective

- Domaine d'analyse
  - La plupart des modèles permettent de spécifier un espace de recherche auquel ils doivent se limiter
    - Obligatoire pour les modèles non déterministes
  - Donné par un ou plusieurs types de marquage
- Filtrage des marquages
  - Chaque composant peut choisir de ne voir qu'une partie des marquages présents dans le document

Exemple

- Extraction d'information dans des constats d'accidents
  - Identification des événements de type "collision entre véhicules"
  - Analyses préliminaires : mots et phrases

- **Phase 1** : identification des unités pertinentes
  - Elle procède au marquage :
    - des entités « véhicules » (grammaire EDCG)
    - des verbes de collision (lexique sémantique LSL)
  - Sa perspective est définie par :
    - *token* = mot
    - *filtre* = aucun
    - *espace de recherche* : quelconque

- **Phase 2** : identification des relations de collision
  - Elle procède au marquage :
    - des séquences entité$_1$ – verbe collision – entité$_2$ (expression MRE)
  - Sa perspective est définie par :
    - *token* = véhicule, verbe-collision
    - *filtre* = ne conserver que phrase, véhicule, et verbe-collision
    - *espace de recherche* : phrase

MRE

```
<evnt>
    {type:véhicule} /as $p1
    {type:prédicat-collision}
    {type:véhicule} /as $p2
</evnt> /sem {protag1:$p1, protag2:$p2}
```

Mozilla Firefox

File Edit View Go Bookmarks Tools Help file:///home/fbilhaut/maif.xml

[§1] [Texte A1.] [Me rendant à Beaumont sur Oise depuis Cergy, je me suis retrouvée à un carrefour juste après la sortie Beaumont sur Oise.] [J'étais à un stop avec 2 voitures devant moi tournant à droite vers Mours.] [Alors que la première voiture passait ce stop je fis mon contrôle à gauche et je démarrais mais [+]{event~je percutais la deuxième voiture~event} qui n'avait pas encore passé le stop.] [/§]

[§2] [Texte A2.] [Voulant dépasser un semi-remorque clignotant à droite, ce dernier tourna à gauche m'obligeant à braquer à gauche pour l'éviter.] [La voiture a dérapé sur la chaussée mouillée et a percuté un trottoir puis un mur de clôture en face.] [Le conducteur du camion avait bien mis son clignotant à gauche mais sa remorque inversait le signal sur la droite.] [Ne m'ayant pas touché le conducteur s'est déclaré hors de cause et n'a pas voulu établir de constat.] [Ayant quitté ma voiture pour appeler un dépanneur j'ai retrouvé celle-ci avec la portière arrière droite enfoncée sans coordonnées du responsable.] [/§]

[§3] [Texte A3.] [Fort trafic à 17 h 15 Bd Sébastopol.] [Je roulais entre deux files de voitures arrêtées quand l'une des voitures à ma gauche a ouvert sa porte avant droite.] [Pour l'éviter, [+]{event~j'ai fait un écart qui m'a fait toucher le véhicule B~event} avec l'arrière de ma moto ce qui a provoqué ma chute.] [Vu l'importance du trafic à échangé nos assurances et noms ce qui explique que mon constat amiable ne soit signé que par moi.] [/§]

evnt/2 [X]

- *prota1:*
  - *type_ent: personne*
  - *role: redacteur*
- *prota2:*
  - *type_ent: vehicule*
  - *type_vehic: véhicule*
  - *spec:* *nomAB: B*

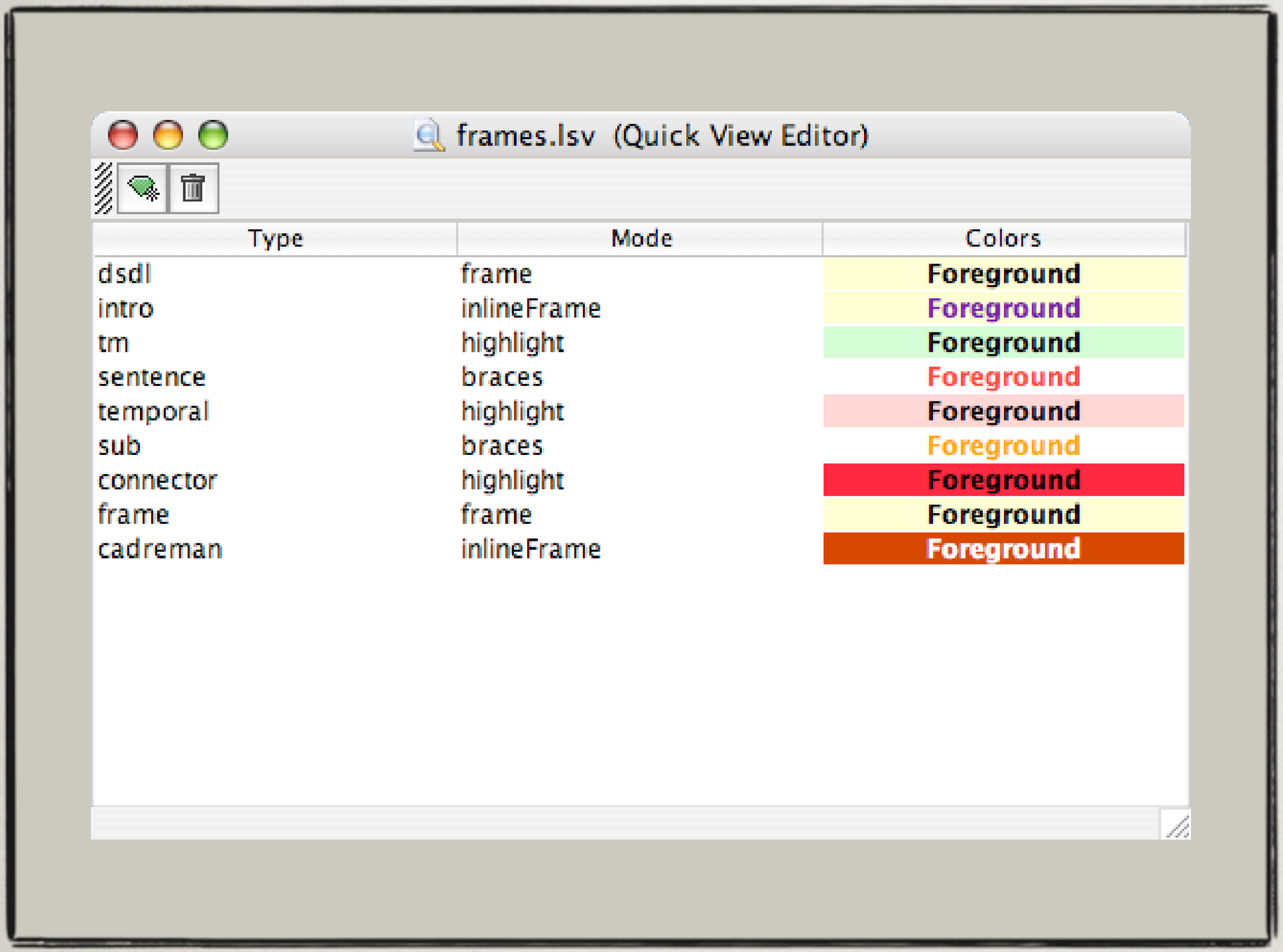
[§4] [Texte A4.] [Véhicule B venan dans le carrefour, à faible vitesse environ 40 km/h, quand [+]{event~le véhicule B, percute mon véhicule~event}, et me premier choc atteint mon aile arrière gauche, sous le choc, et à cause de la chaussée glissante, mon véhicule dérape, et pe un arbre, d'où un second choc frontal.] [/§]

[§5] [Texte A5.] [J'étais arrêté à l'int route où la circulation intense s'effectue à sens unique sur deux voies;] [lorsque le dernier véhicule du flot arrivait, j'ai v file, lui laissant libre la première.] [Au moment où je démarrais, j'ai entendu le choc arrière;] [je ne m'attendais pas à ce qu car il n'y avait pas deux voies matérialisées sur la portion de route où je me trouvais à l'arrêt.] [/§]

[§6] [Texte A6.] [Mr C.][Delon, abordant le carrefour, laissait le passage aux véhicules roulant sur la voie abordée, car d'ordinaire se trouve un feu à ce carrefour(hors fonctionnement ce jour-là).] [Venant de derrière moi, roulant dans le même sens, dans la même file, Mr Oms n'a pas vu que [+]{event~j'étais arrêté et a percuté fortement mon véhicule~event}, l'abîmant gravement.] [De ce fait, j'ai subi (C.][Delon) "le coup du lapin";] [le siège conducteur a été endommagé;] [les gendarmes se sont rendus sur place;] [j'ignore s'ils ont établi un rapport.] [/§]

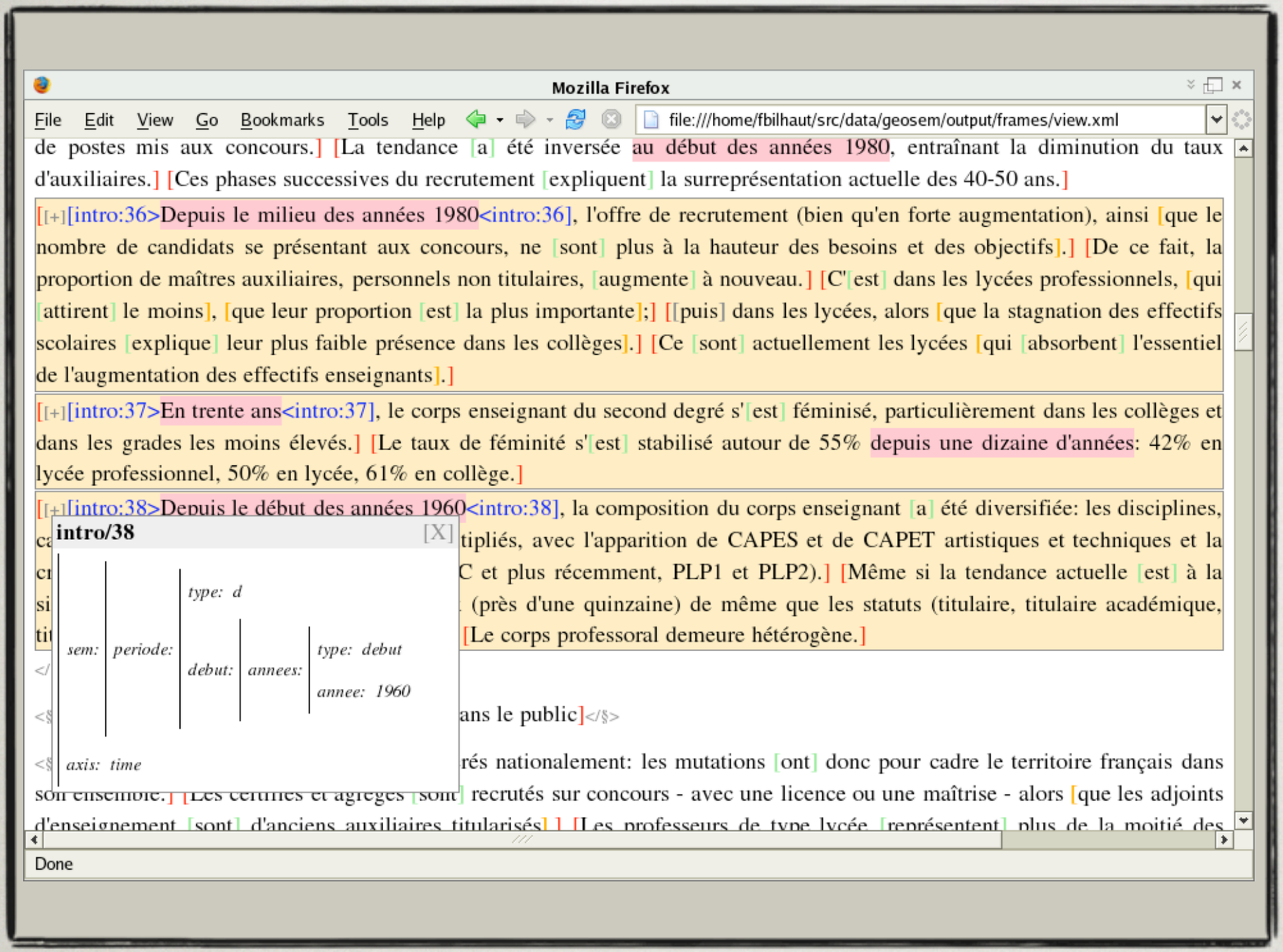
Done

Visualisation et évaluation des résultats



Propriétés de visualisation

Chaque type d'annotation peut se voir attribuer des propriétés d'affichages spécifiques.



Visualisation dans le document

Les annotations (structures de traits, relations) peuvent être affichées.
La structure originelle est préservée.

World Wide Web Consortium

W3C WORLD WIDE WEB consortium

Leading the Web to Its Full Potential...

Activities | Technical Reports | Site Index | New Visitors | About W3C | Join W3C | Contact W3C

The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential. W3C is a forum for information, commerce, communication, and collective understanding. On this page, you'll find W3C news, links to W3C technologies and ways to get involved. New visitors can find help in *Finding Your Way at W3C*. We encourage organizations to learn more about W3C and about W3C Membership.

Mobile Web Initiative

W3C has launched the Mobile Web Initiative (MWI) to make Web access from a mobile device as simple, easy, and convenient as Web access from a desktop device. Read about MWI and MWI sponsors and learn how to become one of them.

W3C Membership News

W3C has lowered fees for non-profit and small commercial organizations based in developing countries (see the press release) and for intermediate-sized companies. Read about W3C fees.

W3C A to Z

- Accessibility
- Amaya
- Annotea
- CC/PP
- Compound Document Formats
- CSS

News

W3C Office Opens in Australia

2005-10-06: W3C is pleased to announce the CSIRO ICT Centre in Canberra hosts the W3C Australian Office effective 10 October. Ross Ackland is Office Manager. "W3C considers Australia a key to global adoption of Web technologies, and we welcome CSIRO as an Office host," said Ivan Herman, W3C Head of Offices. W3C wishes to thank DSTC in Brisbane and staff members Liz Armstrong and Hoylen Sue for hosting the previous Australian Office. Read about W3C Offices. (News archive)

Working Group Note: Time Zones

2005-10-13: Based on discussions with the XQuery and XSL Working Groups, the Internationalization Core Working Group has released Working with Time Zones as a Working Group Note. The document discusses problems

Search

Google

Search W3C

Go

Search W3C Mailing Lists

Members

Fundación CTIC (Centro Tecnológico para el Desarrollo en Asturias de las Tecnologías de la Información y la Comunicación)

FUNDACIÓN CTIC

The mission of the Information and Communication Technology Centre is the promotion and development of the Information Society within the social, institutional, and business communities. The Centre is commited to use W3C

Done

World Wide Web Consortium

W3C® WORLD WIDE WEB consortium

[Leading the Web to Its Full Potential...]

Activities | Technical Reports | Site Index | New Visitors | About W3C | Join W3C | Contact W3C

[The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential.] [W3C is a forum for information, commerce, communication, and collective understanding.] [On this page, you'll find W3C news, links to W3C technologies and ways to get involved.] [New visitors can find help in *Finding Your Way at W3C*.] [We encourage organizations to learn more about W3C and about W3C Membership.]

[Mobile Web Initiative]

[W3C has launched the Mobile Web Initiative (MWI) to make Web access from a mobile device as simple, easy, and convenient as Web access from a desktop device.] [Read about MWI and MWI sponsors and learn how to become one of them.]

[W3C Membership News]

[W3C has lowered fees for non-profit and small commercial organizations based in developing countries (see the press release) and for intermediate-sized companies.] [Read about W3C fees.]

[W3C A to Z]

- Accessibility
- Amaya
- Annotea
- CC/PP
- Compound Document Formats
- CSS

[News]

▸ W3C Office Opens in Australia

[2005-10-06: W3C is pleased to announce the CSIRO ICT Centre in Canberra hosts the W3C Australian Office effective 10 October.] [Ross Ackland is Office Manager.] ["W3C considers Australia a key to global adoption of Web technologies, and we welcome CSIRO as an Office host," said Ivan Herman, W3C Head of Offices.] [W3C wishes to thank DSTC in Brisbane and staff members Liz Armstrong and Hoylen Sue for hosting the previous Australian Office.] [Read about W3C Offices.] [(News archive)]

▸ Working Group Note: Time Zones

[2005-10-13: Based on discussions with the XQuery and XSL Working Groups, the Internationalization Core Working Group has released Working with Time Zones as a Working Group Note.] [The document discusses problems

[Search]

Google™

Search W3C

Go

[Search W3C Mailing Lists]

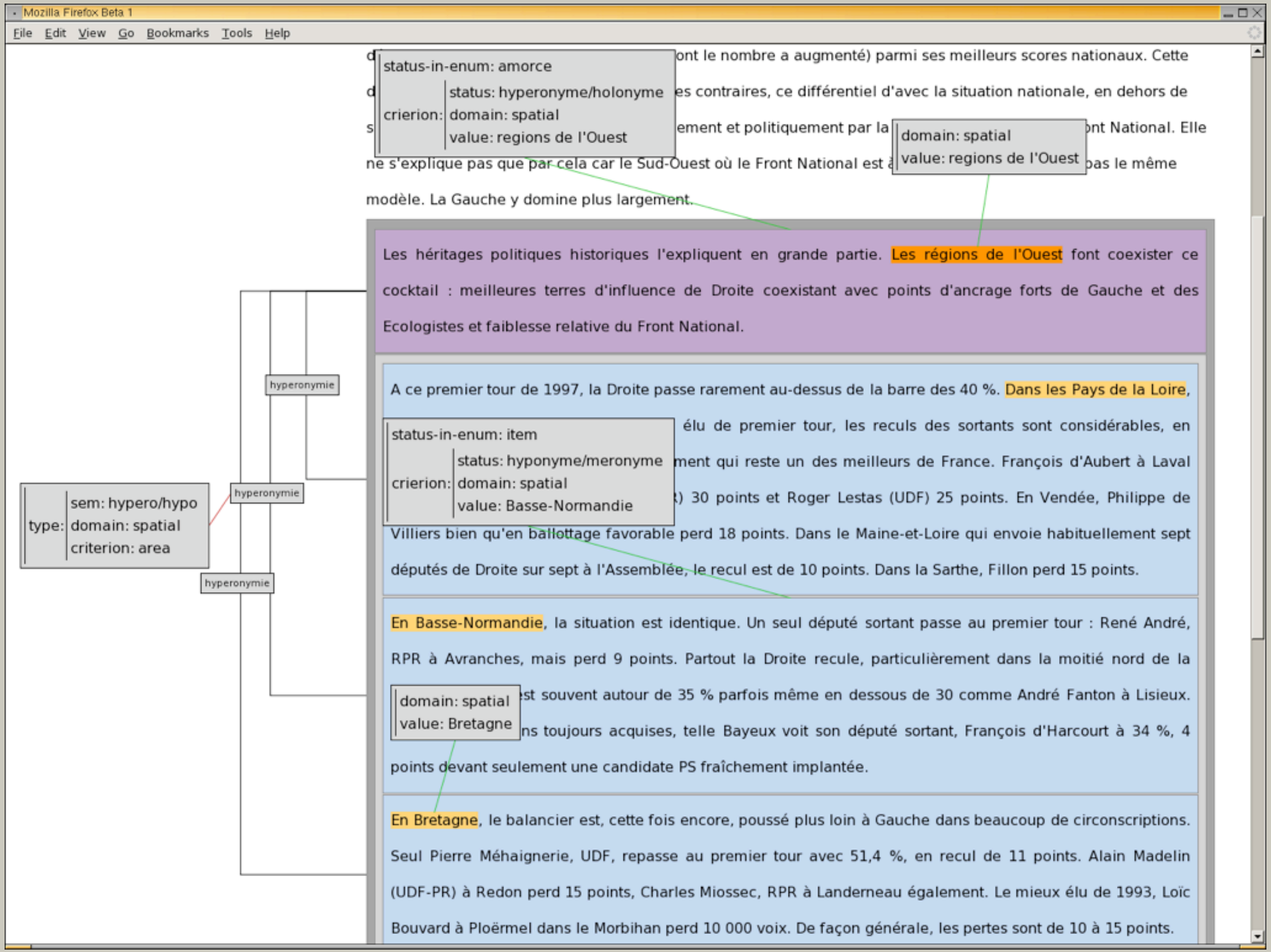
[Members]

Fundación CTIC (Centro Tecnológico para el Desarrollo en Asturias de las Tecnologías de la Información y la Comunicación)

FUNDACIÓN CTIC

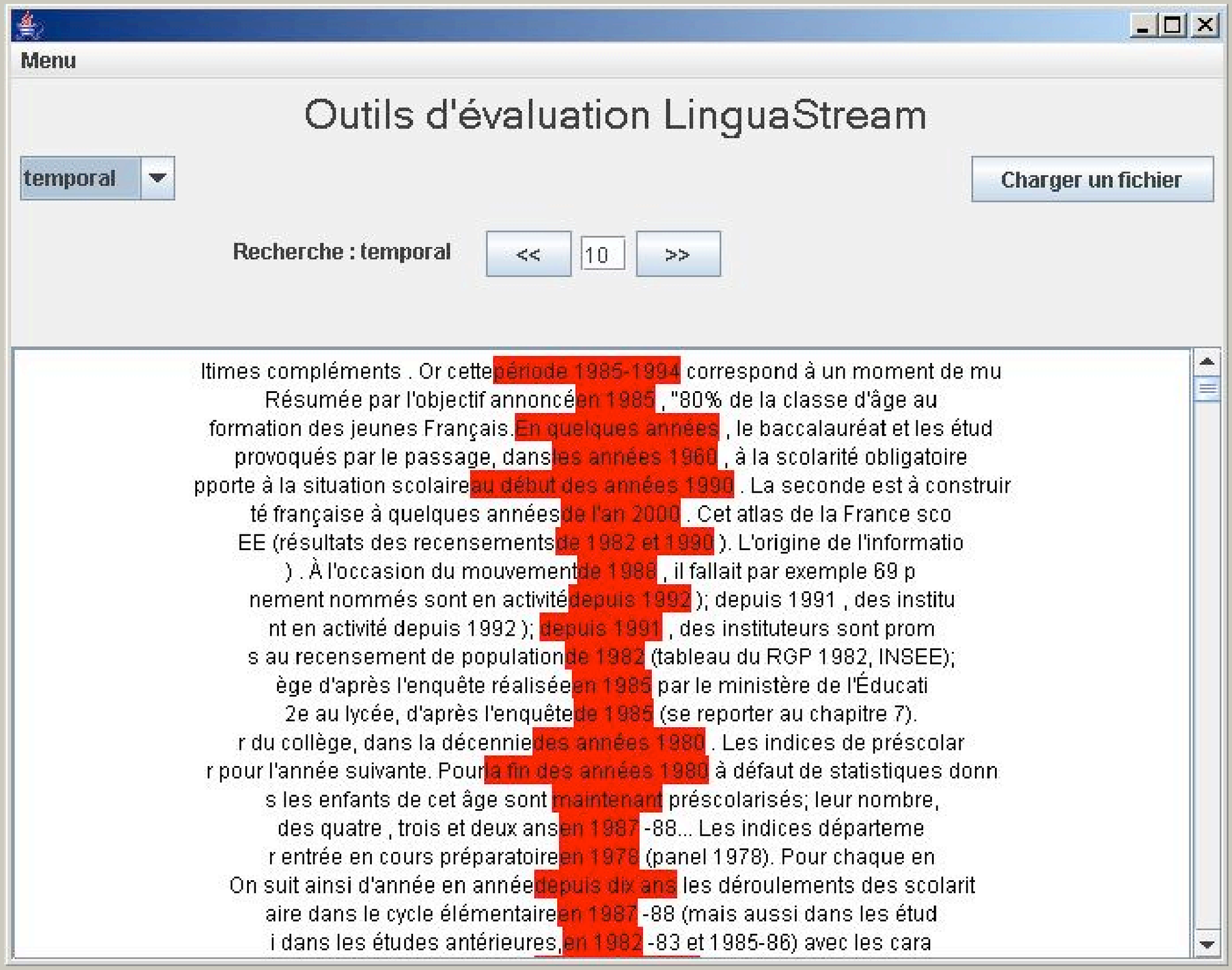
[The mission of the Information and Communication Technology Centre is the promotion and development of the Information Society within the social, institutional, and business communities.] [The Centre is commited to use W3C

Done


Mozilla Firefox Beta 1
File Edit View Go Bookmarks Tools Help
status-in-enum: amorce
crierion: status: hyperonyme/holonyme
domain: spatial
value: regions de l'Ouest
domain: spatial
value: regions de l'Ouest
Les héritages politiques historiques l'expliquent en grande partie. Les régions de l'Ouest font coexister ce cocktail : meilleures terres d'influence de Droite coexistant avec points d'ancrage forts de Gauche et des Ecologistes et faiblesse relative du Front National.
hyperonymie
A ce premier tour de 1997, la Droite passe rarement au-dessus de la barre des 40 %. Dans les Pays de la Loire,
status-in-enum: item
crierion: status: hyponyme/meronyme
domain: spatial
value: Basse-Normandie
type: sem: hypero/hypo
domain: spatial
criterion: area
hyperonymie
hyperonymie
Villiers bien qu'en ballottage favorable perd 18 points. Dans le Maine-et-Loire qui envoie habituellement sept députés de Droite sur sept à l'Assemblée, le recul est de 10 points. Dans la Sarthe, Fillon perd 15 points.
En Basse-Normandie, la situation est identique. Un seul député sortant passe au premier tour : René André, RPR à Avranches, mais perd 9 points. Partout la Droite recule, particulièrement dans la moitié nord de la
domain: spatial
value: Bretagne
points devant seulement une candidate PS fraîchement implantée.
En Bretagne, le balancier est, cette fois encore, poussé plus loin à Gauche dans beaucoup de circonscriptions. Seul Pierre Méhaignerie, UDF, repasse au premier tour avec 51,4 %, en recul de 11 points. Alain Madelin (UDF-PR) à Redon perd 15 points, Charles Miossec, RPR à Landerneau également. Le mieux élu de 1993, Loïc Bouvard à Ploërmel dans le Morbihan perd 10 000 voix. De façon générale, les pertes sont de 10 à 15 points.

Visualisation sélective

- Différents composants permettent de produire des "vues" spécifiques d'un document annoté
  - Vue type concordancier classique
  - Vue "macro-concordancier" plus adapté au grain discours
  - ...
- Les propriétés d'affichage sont spécifiées une fois pr ttes
- Facilité de développement de nouvelles "vues"


Menu
Outils d'évaluation LinguaStream
temporal
Charger un fichier
Recherche : temporal
<<
10
>>
ltimes compléments . Or cettepériode 1985-1994 correspond à un moment de mu
Résumée par l'objectif annoncéen 1985 , "80% de la classe d'âge au
formation des jeunes Français.En quelques années , le baccalauréat et les étud
provoqués par le passage, dansles années 1960 , à la scolarité obligatoire
pporte à la situation scolaireau début des années 1990 . La seconde est à construir
té française à quelques annéesde l'an 2000 . Cet atlas de la France sco
EE (résultats des recensementsde 1982 et 1990). L'origine de l'informatio
) . À l'occasion du mouvementde 1988 , il fallait par exemple 69 p
nement nommés sont en activitédepuis 1992); depuis 1991 , des institu
nt en activité depuis 1992); depuis 1991 , des instituteurs sont prom
s au recensement de populationde 1982 (tableau du RGP 1982, INSEE);
ège d'après l'enquête réaliséeen 1985 par le ministère de l'Éducati
2e au lycée, d'après l'enquêtede 1985 (se reporter au chapitre 7).
r du collège, dans la décenniedes années 1980 . Les indices de préscolar
r pour l'année suivante. Pourla fin des années 1980 à défaut de statistiques donn
s les enfants de cet âge sont maintenant préscolarisés; leur nombre,
des quatre , trois et deux ansen 1987 -88... Les indices départeme
r entrée en cours préparatoireen 1978 (panel 1978). Pour chaque en
On suit ainsi d'année en annéedepuis dix ans les déroulements des scolarit
aire dans le cycle élémentaireen 1987 -88 (mais aussi dans les étud
i dans les études antérieures,en 1982 -83 et 1985-86) avec les cara

Mozilla Firefox

file://localhost/tmp/LinguaStream_64539.tmp

{cadreman:14»En 1985 ... préparaient ... en trois ans ... en 1991 ... peine ... «cadreman:14}

{frame:6»Entre 1985 et 1992 ... «frame:6}

{cadreman:15»En cinq ans ... la rentrée 1985 ... de 1992 ... en deux ansest ... vont ... sont ... rejoignent ... la rentrée de 1993 ... entreprennent ... en deux ans ... Entre 1985 et 1992 ... ont ... est ... depuis quelques années ... cycle ... accueillentmaintenant ... la rentrée 1992 ... au milieu des années 1980 ... a ... cycle ... «cadreman:15}

{cadreman:16»En 1985 ... «cadreman:16}

{frame:7»En 1985 ... en 1987 ... «frame:7}

{cadreman:17»En 1985 ... ont ... «cadreman:17}

{frame:8»En 1985 ... ont ... «frame:8}

{cadreman:18»En 1990 ... approche ... «cadreman:18}

{frame:9»En 1990 ... approche ... «frame:9}

{cadreman:19»en 1992 ... «cadreman:19}

{frame:10»en 1992 ... «frame:10}

{cadreman:20»En une quinzaine d'années seulement ... est ... puis ... depuis le milieu des années 1980a ... cependant ... est ... sont ... traduit ... tendent ... resserrent ... rattrapent ... dépassent ... font ... caractérisentdepuis des décennies ... «cadreman:20}

{frame:11»En une quinzaine d'années seulement ... est ... puis ... depuis le milieu des années 1980a ... cependant ... est ... sont ... traduit ... tendent ... resserrent ... rattrapent ... dépassent ... font ... caractérisentdepuis des décennies ... mettaient ... sont ... en quelques années ... reste ... estmaintenant ... jouent ... poursuivent ... offrentsont ... entreprennent ... ont ... continuent ... conduisent ... sont ... sont ... sont ... est ... orienté ... offrent ... «frame:11}

{cadreman:21»Depuis 1987 ... est ... a ... a ... est ... «cadreman:21}

{frame:12»Depuis 1987 ... est ... a ... a ... est ... «frame:12}

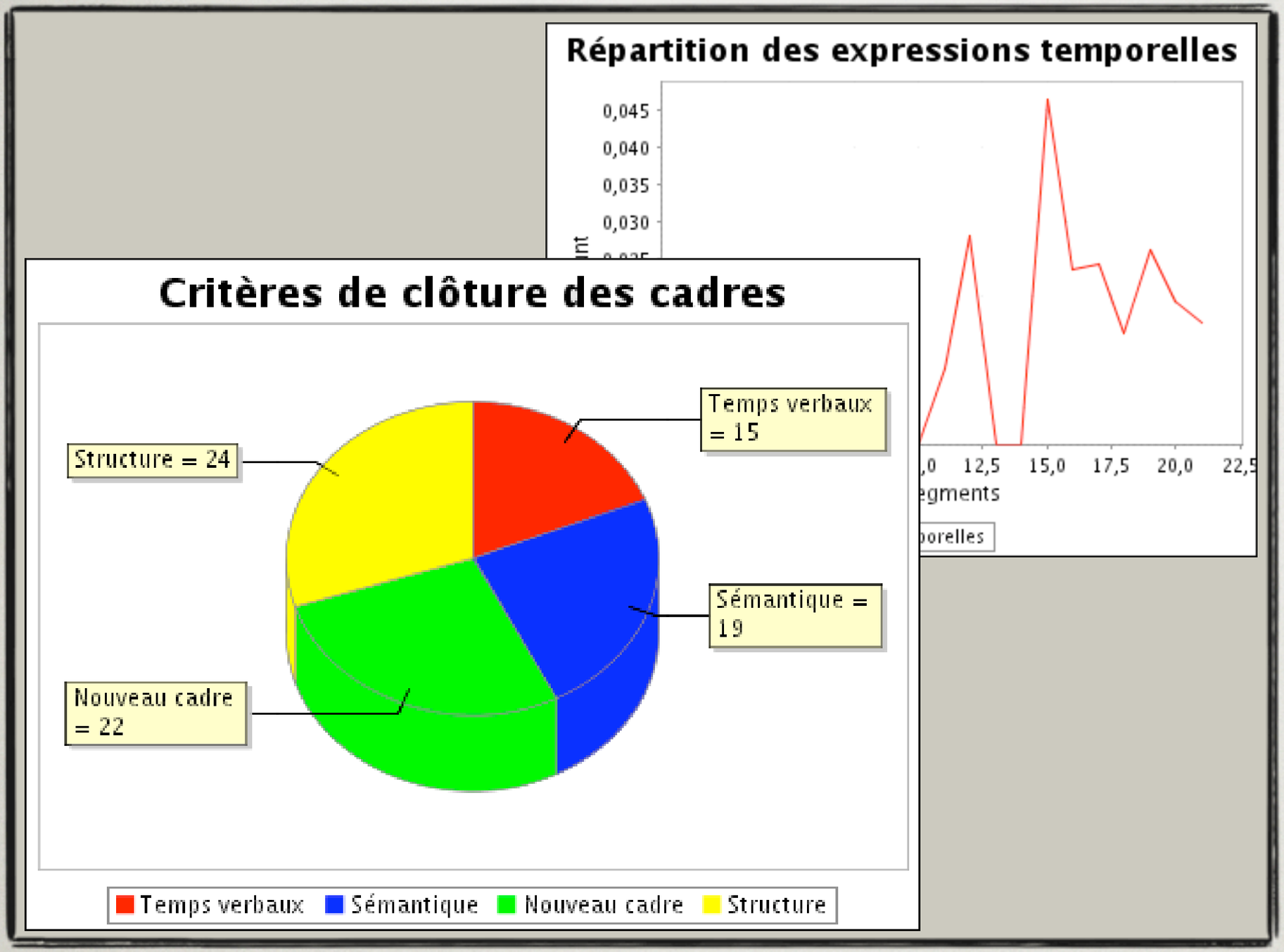
{cadreman:22»Entre 1988 et 2000 ... a ... en 1980 ... en 1984 ... en 1987 ... en 1990 ... est ... ont ... est ... est ... «cadreman:22}

{frame:13»Entre 1988 et 2000 ... a ... en 1980 ... en 1984 ... en 1987 ... en 1990 ... est ... ont ... est ... est ... «frame:13}

{cadreman:23»Cependant, depuis le début des années 1990 ... est ... rencontrent ... «cadreman:23}

{frame:14»Cependant, depuis le début des années 1990 ... est ... rencontrent ... «frame:14}

Terminé



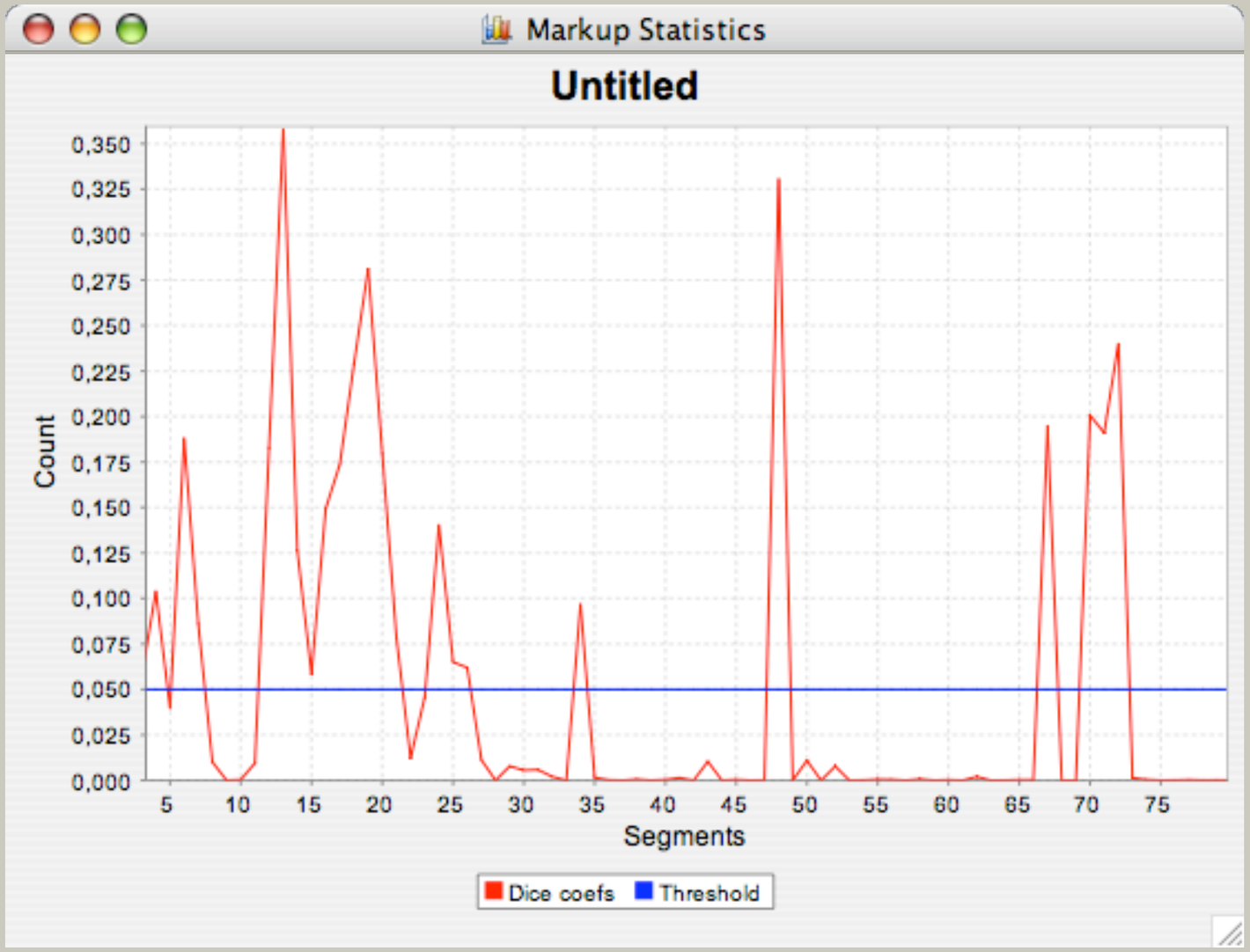
Visualisation sous forme de graphes

frameClosure.cd (Chart Descriptor Editor)

Standard | Expert

chartTitle	Critères de clôture des cadres
exclusiveItems	True
insertLegend	True
mode	Total Count
segmentTypes	

Name	Markup Type	Feature Set	Value Feature
Nouveau cadre	frame	{closureCriterion:newFrame}	
Sémantique	frame	{closureCriterion:semantics}	
Structure	frame	{closureCriterion:structure}	
Temps verbaux	frame	{closureCriterion:tenses}	
Autres	frame	{}	


Markup Statistics
Untitled
Count
0,350
0,325
0,300
0,275
0,250
0,225
0,200
0,175
0,150
0,125
0,100
0,075
0,050
0,025
0,000
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
Segments
Dice coefs
Threshold

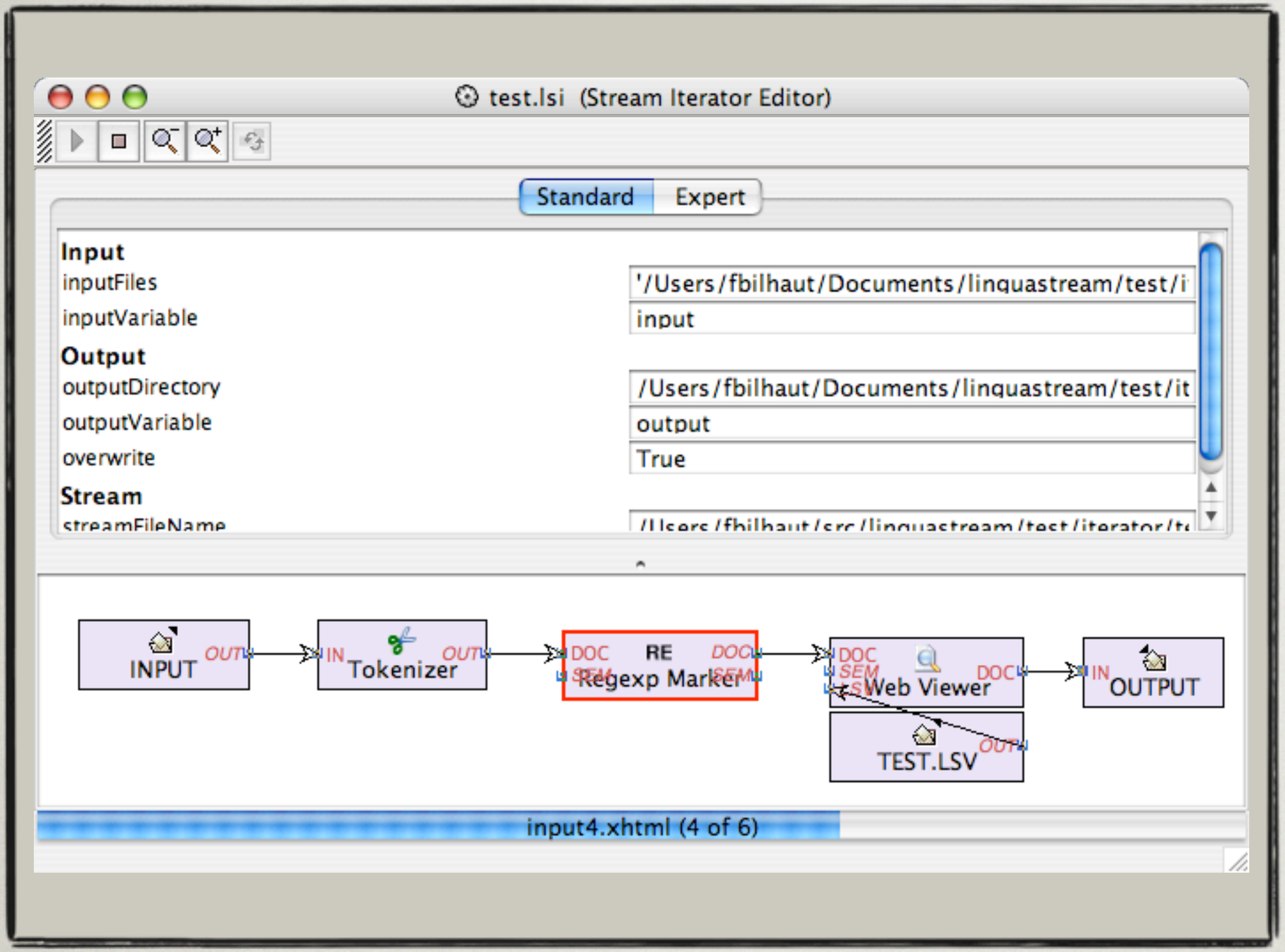
test.cd (Chart Descriptor Editor)

Standard | Expert

chartTitle	Untitled
exclusiveItems	False
insertLegend	True
mode	Feature Values
segmentTypes	

Name	Markup Type	Feature Set	Value Feature
Dice coefs	paragraph	{}	tfidf/dice
Threshold	paragraph	{}	tfidf/threshold

Gestion des chaînes des traitements



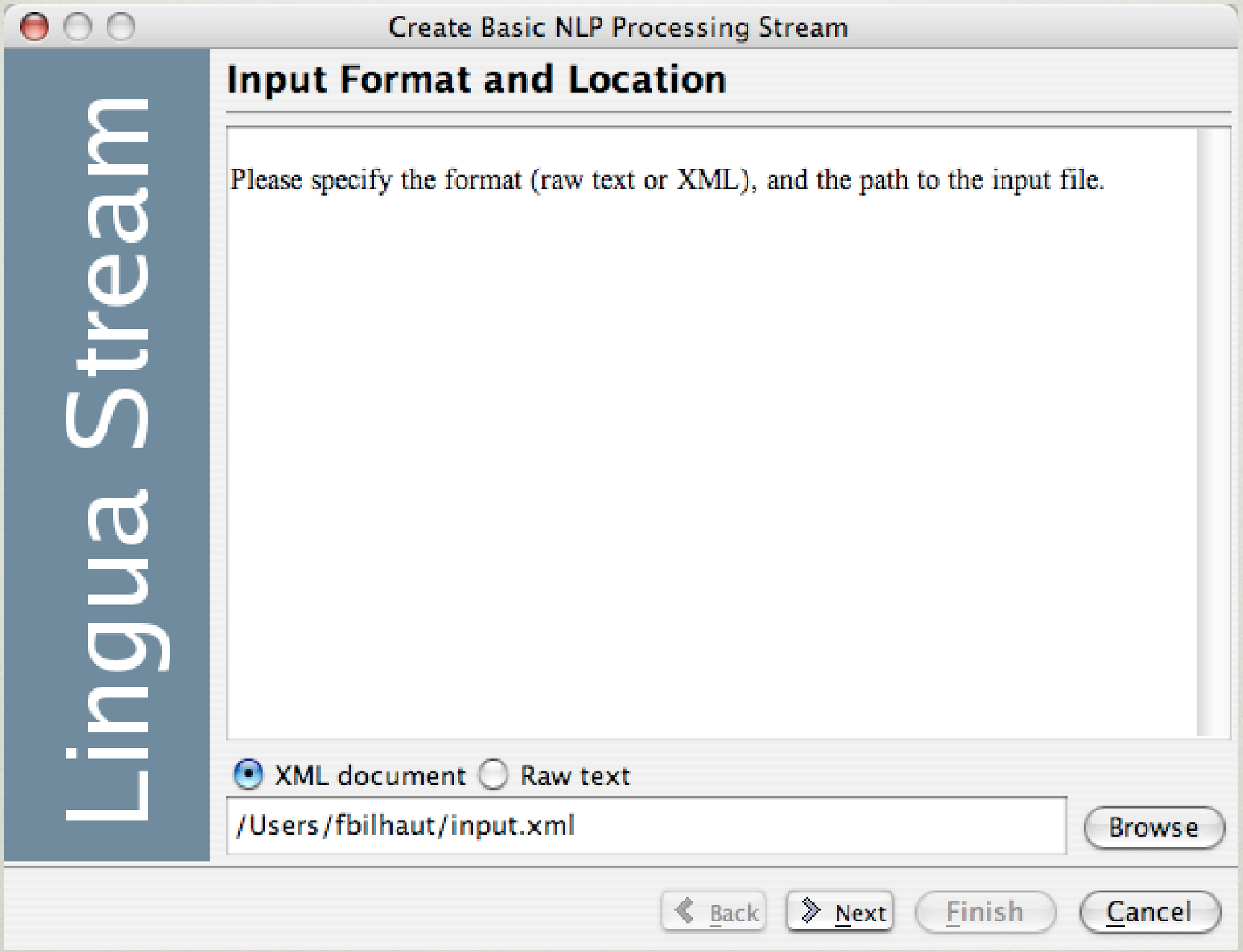
Application d'une chaîne sur un lot de documents

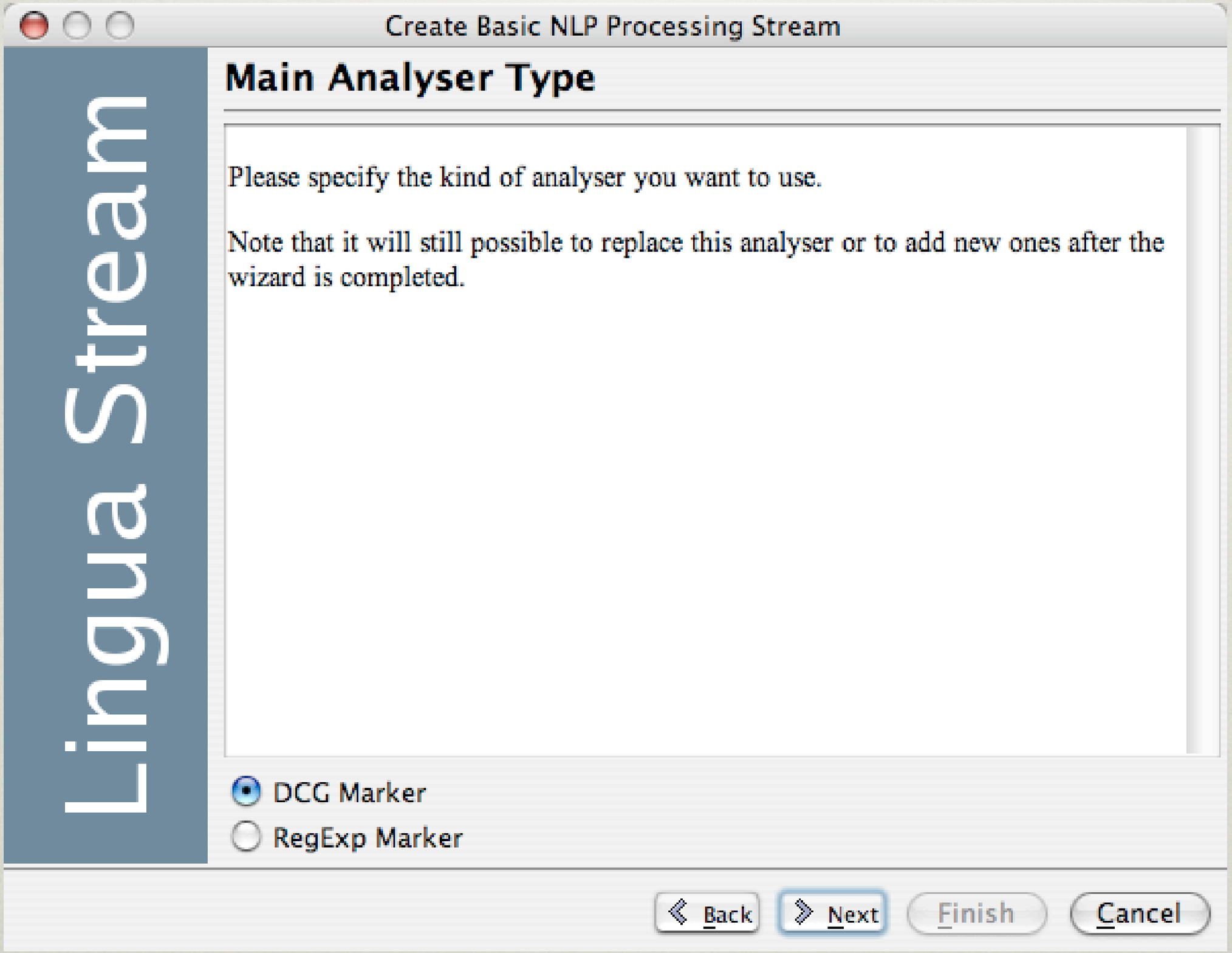
Scripts d'automatisation

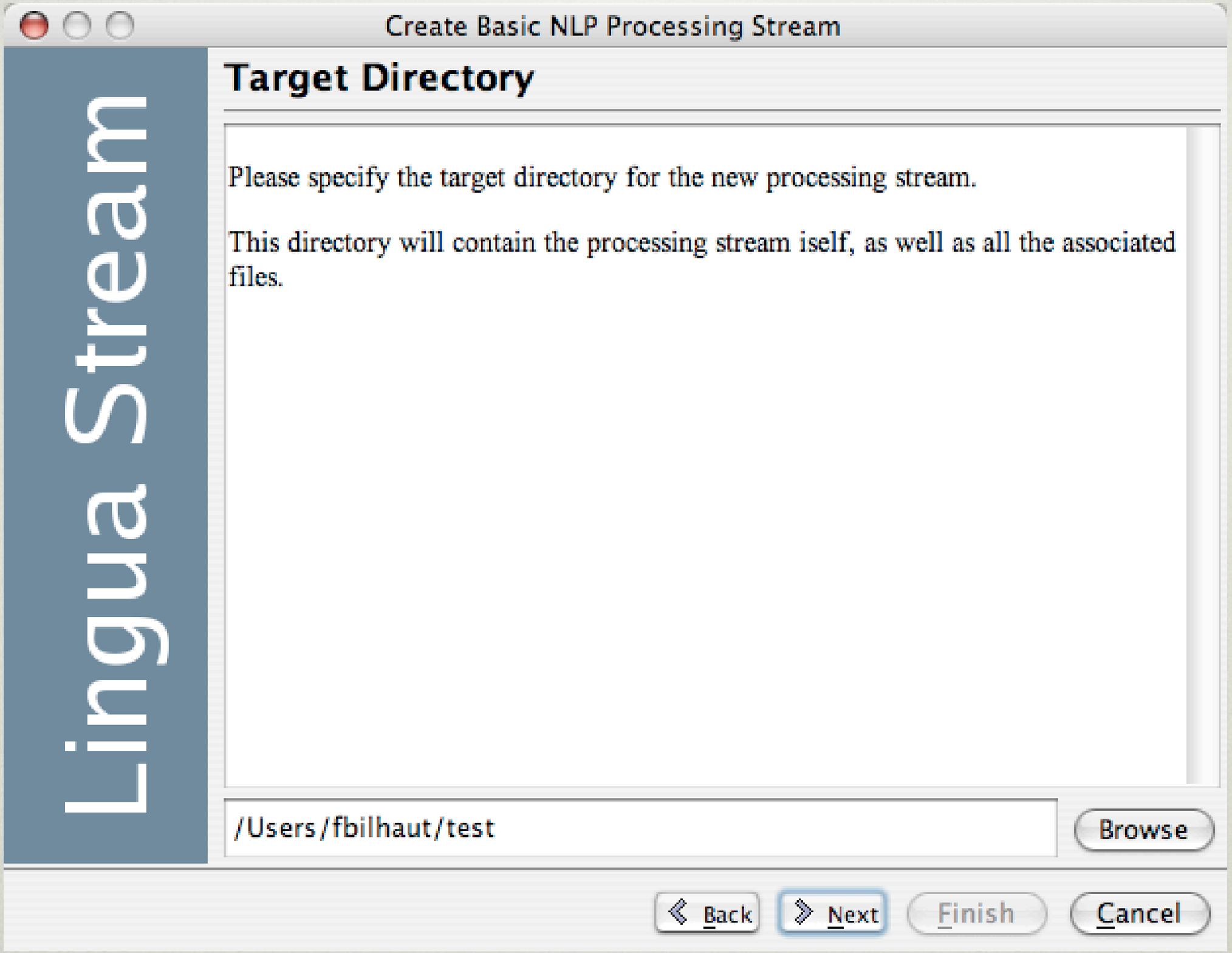
```
query = automation.getUserInput("Query:");

google = automation.getGoogle("xxxxxxxxxxx");
google.maxResults = 10;
google.languages = "lang_fr";

results = google.search(query);

stream = automation.openStream("analysis.ls");

for(res : results)
{
    stream.setParameter("input", res.URL);
    // ...
    stream.run();
}
```

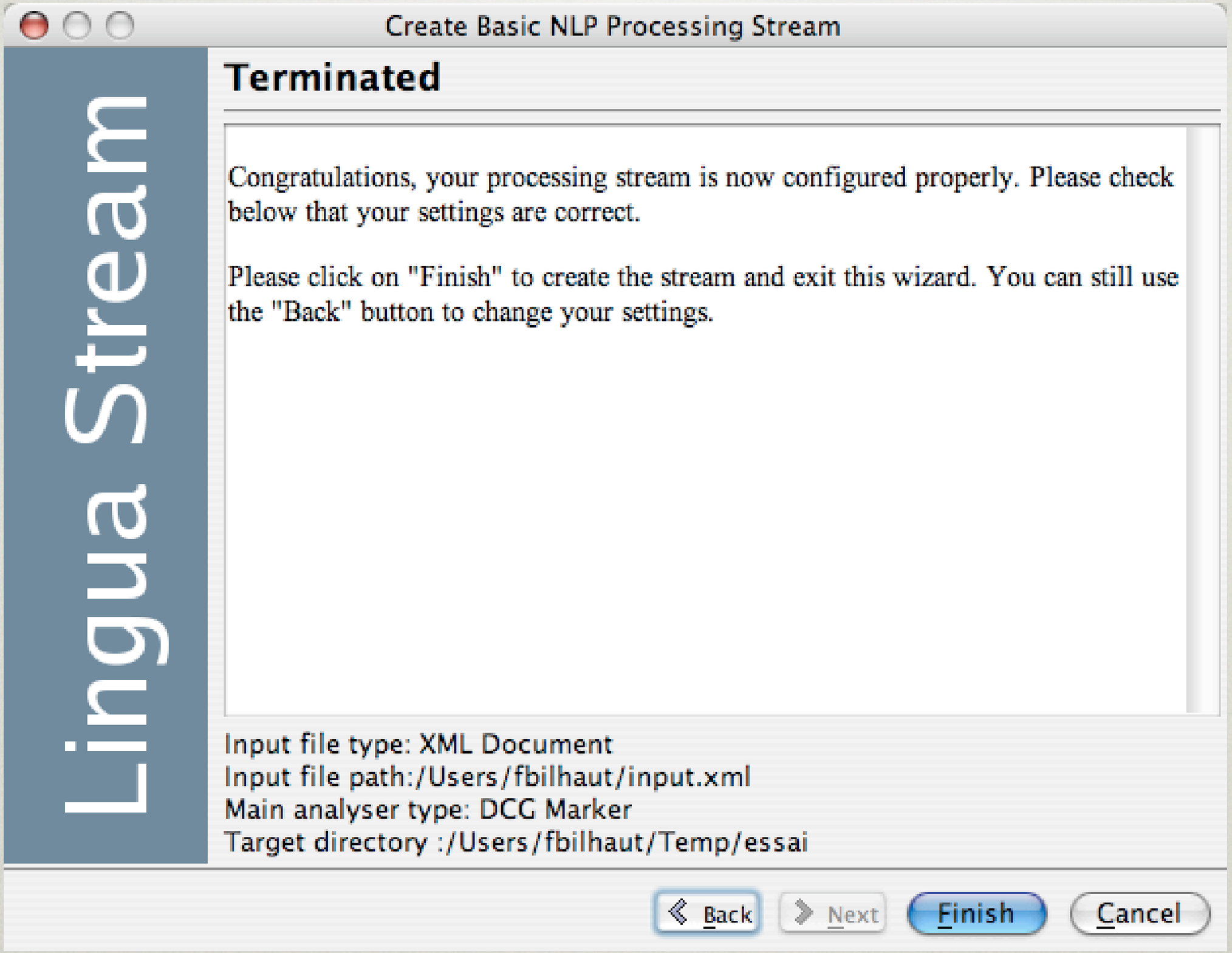
Assistants

- Interface type “wizard” permettant de créer facilement des chaînes de traitements
- Demande à l’utilisateur de renseigner les paramètres principaux
- Génère la chaîne et tous les documents associés (squelettes)


Create Basic NLP Processing Stream
Lingua Stream
Input Format and Location
Please specify the format (raw text or XML), and the path to the input file.
XML document
Raw text
/Users/fbilhaut/input.xml
Browse
Back
Next
Finish
Cancel


Create Basic NLP Processing Stream
Lingua Stream
Main Analyser Type
Please specify the kind of analyser you want to use.
Note that it will still possible to replace this analyser or to add new ones after the wizard is completed.
DCG Marker
RegExp Marker
Back
Next
Finish
Cancel


Create Basic NLP Processing Stream
Linguastream
Target Directory
Please specify the target directory for the new processing stream.
This directory will contain the processing stream iself, as well as all the associated files.
/Users/fbilhaut/test
Browse
Back
Next
Finish
Cancel


Create Basic NLP Processing Stream
Lingua Stream
Terminated
Congratulations, your processing stream is now configured properly. Please check below that your settings are correct.
Please click on "Finish" to create the stream and exit this wizard. You can still use the "Back" button to change your settings.
Input file type: XML Document
Input file path:/Users/fbilhaut/input.xml
Main analyser type: DCG Marker
Target directory :/Users/fbilhaut/Temp/essai
Back
Next
Finish
Cancel

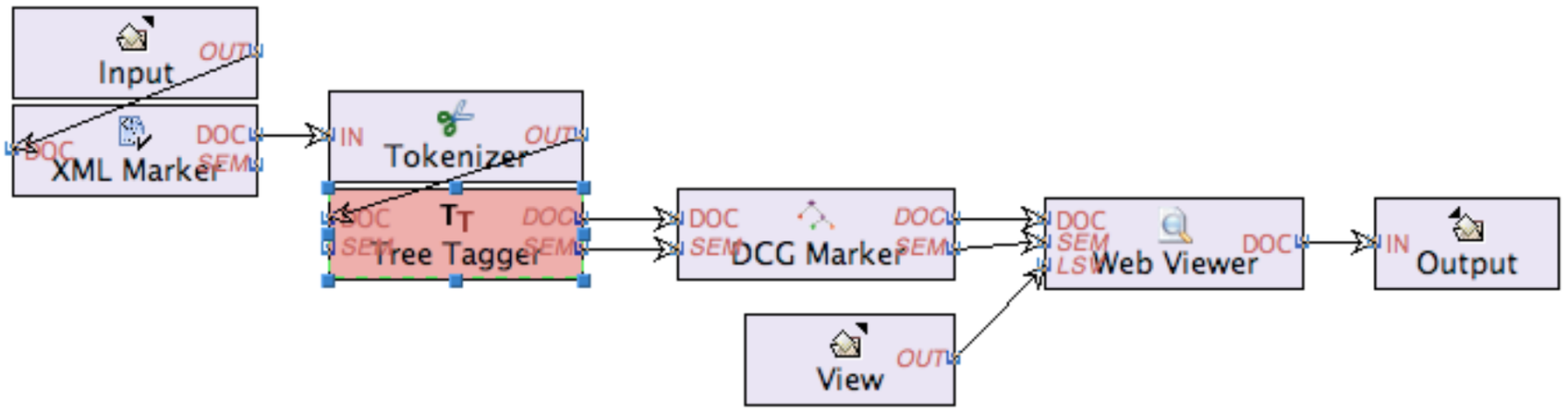
LinguaStream IDE 2.0.1

Explorer | Palette

- Explorer
  - LinguaStream/library
  - LinguaStream/demo
  - Home/src/geosem
  - Home/Documents/test/LS
  - Home/Desktop
  - Home/src/linguastream/test
  - Home/Temp/essai
    - grammar.pro
    - stream.ls
    - struct.xm
    - view.lsv

Working Area

stream.ls (Stream Editor)



Properties

Component

cached	True
class	Tree Tagger
comments	
minimized	False
name	Tree Tagger

Markup

tokenTypes	'word'

Tag Set

tagSet	French

Tree Tagger

parameterFilePath	${treetagger.par
taggerBinaryPath	${treetagger.bin

Standard | Expert

Console

```
Plugin 'Stream Format Converter' v. 2.0.1 successfully loaded
Plugin 'Database Tools' v. 2.0.1 successfully loaded
Plugin 'Text Editor' v. 2.0.1 successfully loaded
Plugin 'Lexicon Editor' v. 2.0.1 successfully loaded
Plugin 'Quick View' v. 2.0.1 successfully loaded
Plugin 'SWI Prolog' v. 2.0.1 successfully loaded
Plugin 'Tree Tagger' v. 2.0.1 successfully loaded
Plugin 'Regexp Tokenizer' v. 1.0.0 successfully loaded
Plugin 'Automation Script Editor' v. 2.0.1 successfully loaded
Plugin 'Markup Charter' v. 2.0.0 successfully loaded
Plugin 'Evaluation' v. 1.0.0 successfully loaded
```

Plugin 'GeoSem' v. 0.6.1 successfully loaded

--:--:--

Exemple d'application : Analyse des cadres de discours temporels

L'encadrement du discours (Charolles 97)

- Univers de discours
  - Ensemble de circonstances dans lesquelles une ou plusieurs propositions peuvent être dite "vraies" (critère véridictionnel) → segments textuels délimitables
- Introducteur de cadre
  - Groupe adverbial périphérique : expression détachée en initiale de phrase / en position pré-verbale
  - Cadre véridictionnel = ensemble de propositions soumises au critère d'interprétation spécifié par l'introducteur

- Différents types de cadres :
  - Temporels (“Depuis 1985, ...“),
  - Spatiaux (“En Normandie, ...”)
  - Thématiques (“Concernant X, ... “),
  - Médiatifs (“Selon X, ... “)
  - Praxéologiques (“En Chimie, ...“)
- Projet GeoSem : analyse automatique des cadres **spatiaux** et **temporels**
  - Indexation tri-partite phénomène / espace / temps
  - Indexation de passages
  - Moteur de recherche multi-critères

De 1965 à 1985, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif central, modérée en Bretagne et à Paris, l'augmentation a été considérable dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs ont souvent plus que doublé. Les variations de la population et les baisses plus ou moins fortes du nombre des naissances selon les régions ne suffisent pas à expliquer ces différences d'accroissement des effectifs du secondaire. Intervient aussi l'allongement des scolarités, qui a été plus marqué dans les départements où, au milieu des années 1960, la poursuite des études après l'école primaire était loin d'être la règle. [...]

?

Analyse automatique

- La borne gauche d'un cadre est assez facilement repérable car l'introducteur a une forme caractéristique
- La borne droite en revanche n'est pas signalée explicitement
- Le lecteur humain met vraisemblablement en oeuvre, plus ou moins consciemment, une grande diversité de mécanismes interprétatifs pour évaluer la borne finale
- L'automatisation de ce processus est une question très complexe

- Deux sous-problèmes :
  - Détection des introducteurs (borne gauche)
    - Formes reconnues par les analyseurs d'expr. spat. et temp.
    - Application de critères positionnels (en initiale détachée)
  - Délimitation de la borne droite
    - Nécessité d'établir des critères objectifs et opérationalisables
    - Plusieurs critères : sémantiques, énonciatifs, structurels, ...
    - Aucun de ces critères n'est vraiment fiable à lui seul

      On cherche à exploiter des faisceaux d'indices

Pré-traitements

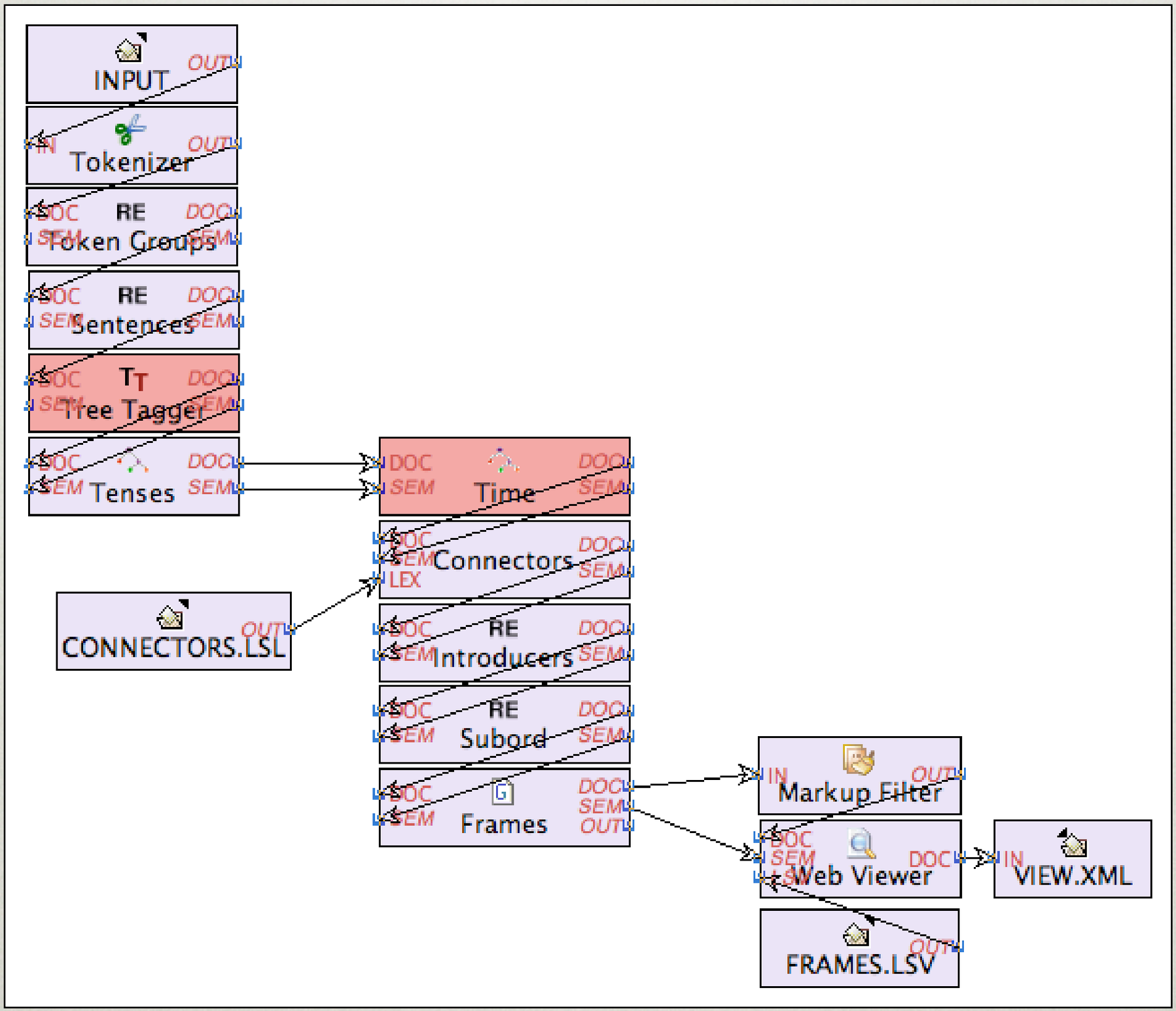
- Etiquetage morpho-syntaxique
- Délimitation des phrases
- Projection d'un lexique de connecteurs de discours
- Analyse syntaxico-sémantique des expressions temporelles

culturellement les plus défavorisés . Ces enjeux , et la demande des parents qu' ils ont contribué à susciter , ont été largement pris en charge par les pouvoirs publics . Les enfants de trois , quatre et cinq ans sont maintenant presque tous scolarisés , alors qu' au début des années 1960 les taux étaient , respectivement , de 36% , 63% et 91% . Et pour les enfants de deux ans , le taux est passé de 10% à 33% entre 1960 et 1982 ; il est maintenant stabilisé , depuis le milieu des années 1980 , à 35% . Cependant des inégalités subsistent , selon les milieux socio-économiques d' origine des enfants , comme selon les régions et départements . En ce qui concerne la préscolarisation à deux ans , le Grand Ouest , le Midi toulousain , le sud et l' est du

$$\begin{bmatrix} type : période \\ début : \begin{bmatrix} type : décénie \\ année : 1980 \\ delta : milieu \end{bmatrix} \end{bmatrix}$$

Critères d'analyse de la portée

- Critères sémantiques
  - Cohérence temporelle
  - Collaborations espace/temps
- Critères énonciatifs : temps verbaux
- Critères structurels : architecture textuelle
- Perspectives
  - Gestion de zones de "digression"
    - Critères a priori ?
- Chaîne de traitement LinguaStream


INPUT
OUT
Tokenizer
IN
OUT
RE
Token Groups
Sentences
TT
Tree Tagger
Tenses
DOC
SEM
Time
Connectors
LEX
CONNECTORS.LSL
Introducers
Subord
Frames
Markup Filter
Web Viewer
VIEW.XML
FRAMES.LSV

DSDL

```
Rule {type:"cadre"}
{
    start({type: "introducteur"})
    end({type : "phrase"})
    homogeneity(comparator:portée)
    size(mode:"longest")
    not presence(pattern : {type : "introducteur"}, amount : 2)
    size(mode : #LONGEST)
}

Comparator portée ({type: "verbe"} as $v1, {type: "verbe"} as $v2)
{
    $v1/temps = $v2/temps
}

Comparator portée ({type: "intro"} as $i, {type: "tempo"} as $t)
{
    (($i/debut >= $t/debut) and ($i/debut <= $t/fin))
  or
    (($i/fin >= $t/debut) and ($i/fin <= $t/fin))
}
```

De 1965 à 1985, le nombre de collégiens et de lycéens **a augmenté** de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif central, modérée en Bretagne et à Paris, l'augmentation **a été** considérable dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs **ont** souvent plus que **doublé**. Les variations de la population et les baisses plus ou moins fortes du nombre des naissances selon les régions **ne suffisent pas** à expliquer ces différences d'accroissement des effectifs du secondaire. Intervient aussi l'allongement des scolarités, qui a été plus marqué dans les départements où, **au milieu des années 1960**, la poursuite des études après l'école primaire était loin d'être la règle. [...]

d'auxiliaires.] [Ces phases successives du recrutement [expliquent] la surreprésentation actuelle des 40-50 ans.]

[[+][intro:36>Depuis le milieu des années 1980<intro:36], l'offre de recrutement (bien qu'en forte augmentation), ainsi [que le nombre de candidats se présentant aux concours, ne [sont] plus à la hauteur des besoins et des objectifs].] [De ce fait, la proportion de maîtres auxiliaires, personnels non titulaires, [augmente] à nouveau.] [C'[est] dans les lycées professionnels, [qui [attirent] le moins], [que leur proportion [est] la plus importante];] [[puis] dans les lycées, alors [que la stagnation des effectifs scolaires [explique] leur plus faible présence dans les collèges].] [Ce [sont] actuellement les lycées [qui [absorbent] l'essentiel de l'augmentation des effectifs enseignants].]

[[+][intro:37>En trente ans<intro:37], le corps enseignant du second degré s'[est] féminisé, particulièrement dans les collèges et dans les grades les moins élevés.] [Le taux de féminité s'[est] stabilisé autour de 55% depuis une dizaine d'années: 42% en lycée professionnel, 50% en lycée, 61% en collège.]

[[+][intro:38>Depuis le début des années 1960<intro:38], la composition du corps enseignant [a] été diversifiée: les disciplines, ca tipliés, avec l'apparition de CAPES et de CAPET artistiques et techniques et la cr C et plus récemment, PLP1 et PLP2).] [Même si la tendance actuelle [est] à la si (près d'une quinzaine) de même que les statuts (titulaire, titulaire académique, tit [Le corps professoral demeure hétérogène.]

intro/38 [X]

- *type: d*
  - *sem:*
  - *periode:*
    - *debut:*
      - *annees:*
        - *type: debut*
        - *annee: 1960*
  - *axis: time*

</ ans le public]</§>

<§ rés nationalement: les mutations [ont] donc pour cadre le territoire français dans son ensemble.] [Les certifiés et agrégés [sont] recrutés sur concours - avec une licence ou une maîtrise - alors [que les adjoints

Evaluation

- Evaluation de la qualité de la mesure de portée
- Comparaison avec une annotation manuelle
  - *Atlas de la France scolaire* ... (R. Hérin, 56 000 mots)
  - *Atlas politique* ... (P. Buléon, ? mots)
  - Journal *Le Monde* (volume à déterminer)
- Avancement :
  - Quelques numéros du monde en cours d'annotation
  - Premièrs résultats à partir de l'annotation de (Hérin)
    - Annotation de Marion Laignelet

- Alignement automatique des cadres
  - Exploitation de l'introducteur repéré automatiquement
  - Seuil de tolérance sur la borne gauche de l'intervalle
  - Première évaluation en termes de précision/rappel
- Mesure de l'écart entre les bornes droites uniquement
  - Dénombrement d'unités linguistiques quelconques
  - Modularité de la mesure, établie par exemple en nombre de mots ou de phrases
  - Mesure relative ou absolue

Méthode d'annotation et de comparaison

- En cas d'ambiguïté l'expert peu annoter une "plage de fin" pour un cadre
- Plusieurs méthodes de comparaison sont envisageables :
  - Distance relativement à la borne la plus proche
  - Pseudo-distance nulle si fin comprise dans l'intervalle
  - ...
- Unité de mesure :
  - En nombre de mots ? de phrases ?
  - Les outils de comparaison exploitent le grain variable de LS

[§168][La diversification des possibilités][/§]

[§169][Par orientations successives de la 2e à la Terminale s'offrent ainsi aux lycéens près de trente séries et options du baccalauréat.] [Certaines préservent pratiquement toutes les possibilités ultérieures: c'est le cas de la série C et, à un degré moindre, des séries D et E.] [D'autres sont déjà très spécialisées, telles les séries F à petits effectifs;] [il en est [+][qui sont peu prisées]: les séries G, voire les séries A.] [Nul ne conteste qu'il existe ainsi une hiérarchie des séries du baccalauréat et des cursus [+][qui y conduisent].] [Il est bien connu [+][que les lycéens sont sensiblement différents d'une série à l'autre].] [En G, série à recrutement très majoritairement féminin, les élèves sont en moyenne plus âgés, ont redoublé une ou deux classes, font fréquemment le parcours de la 2e à la Terminale en quatre ans, et même plus;] [leurs origines sociales sont assez sem-blables à celles des élèves de classes de BEP.] [À l'opposé, les Terminales C sont jeunes, comptent le plus souvent une forte majorité de garçons, et les familles de cadres supérieurs, d'enseignants, de médecins, avocats et autres professions libérales sont de beaucoup les mieux représentées.]

{cadreman:118»[[+]**{intro:74»**Au milieu des années 1960**«intro:74}**, les Terminales des lycées d'enseignement classique et moderne, ainsi [+][que les écoles normales [+][qui recrutaient encore les futurs instituteurs au niveau de la 2e]], ne préparaient qu'à trois baccalauréats: philosophie, sciences expérimentales et mathématiques élémentaires.] [Elles avaient au total une centaine de milliers d'élèves.]**«cadreman:118}** [L'enseignement technique et professionnel conduisait une vingtaine de milliers de lycéens soit aux baccalauréats mathématiques et technique ou technique et économie, ancêtres des actuelles séries F et G, soit aux brevets de

frame/65 [X]

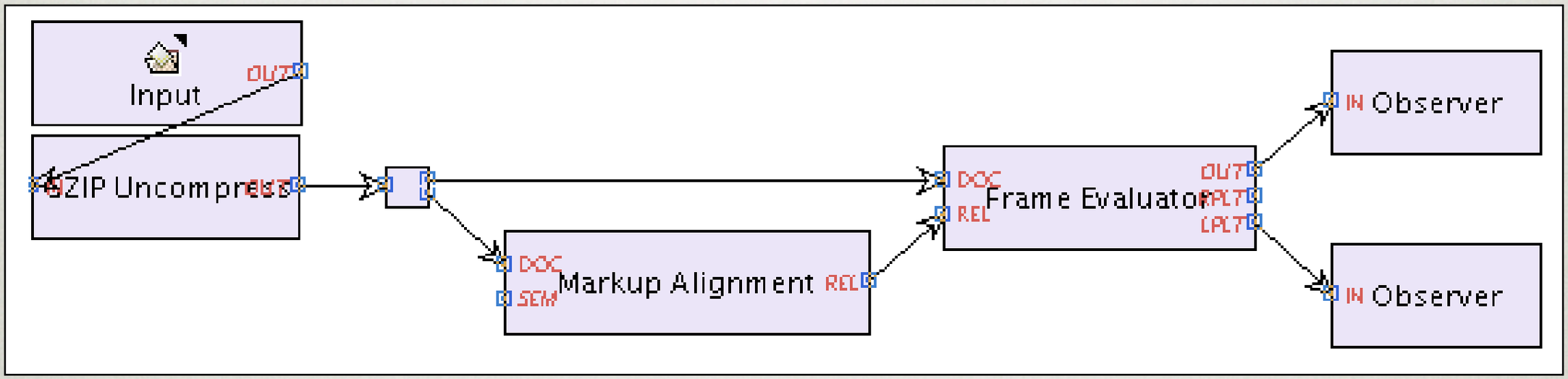
sem: | *annees:* | *type: milieu* | *annee: 1960*

axis: time

closureCriterion: tenses

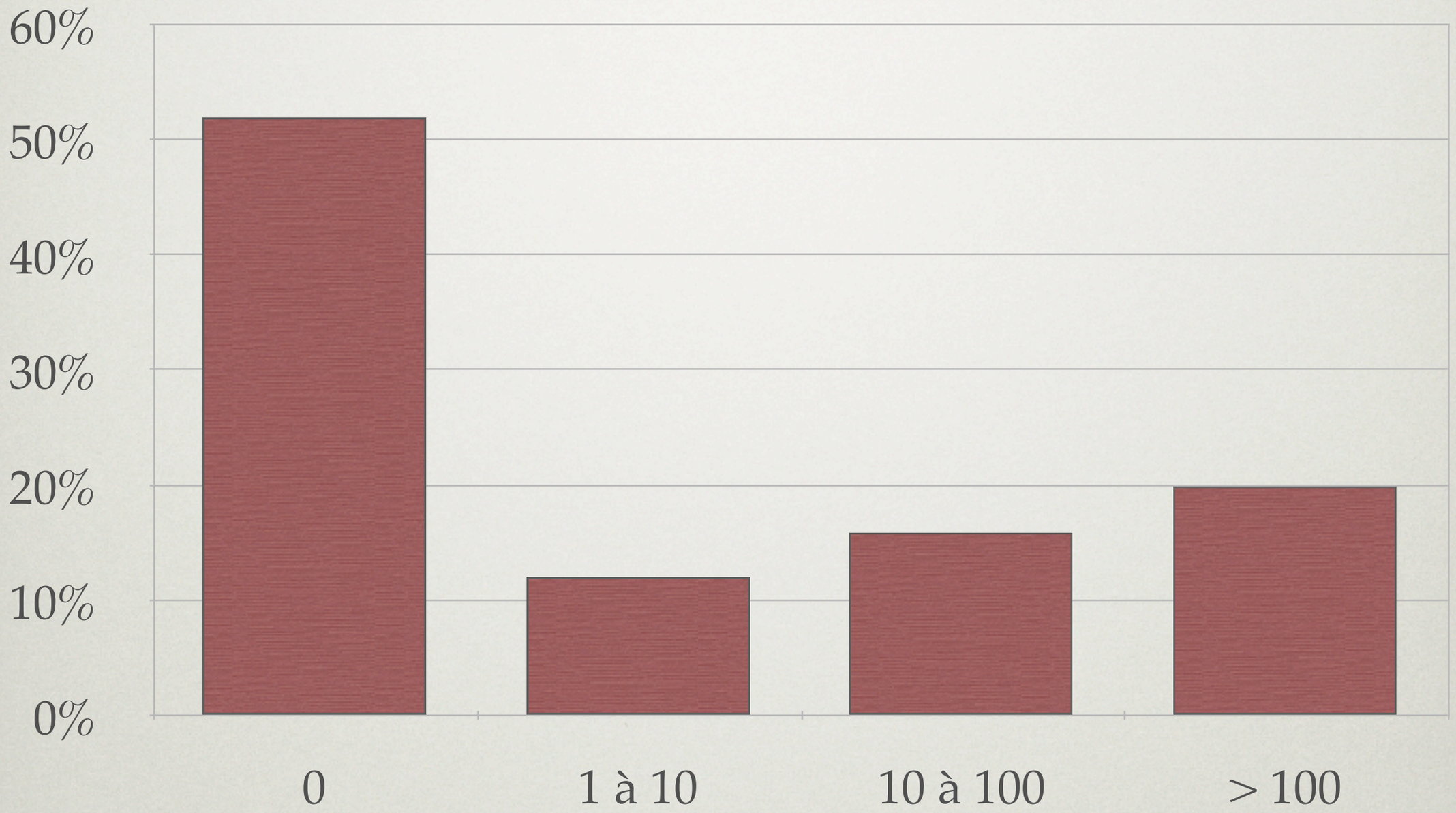
[**{cadreman:119»** Depuis, les cursus scolaires dans le deuxième et technologique ont été considérablement diversifiés;] [dans l'enseignement technologiqu nseignement général, les formations ont été spécialisées], dans un effort d'adaptation des alifications et des profils des emplois.] [De plus, des passerelles ont été créées.] [Des élèv 1re de lycée par les 1re dites d'adaptation;] [inversement, des élèves de 2e peuvent être les élèves des classes de CAP et BEP ont la perspective de continuer leurs études jusqu'a [Près de 30% des élèves du cycle professionnel poursuivent ainsi, généralement après le CAP, des études dans le cycle long.]**«cadreman:119}**[ ][/§]

«cadreAW:2} {cadreSF:3» {cadreFB:3» {cadreAW:3»A la naissance de son fils Dino, en 1931, Enzo Ferrari raccroche ses gants de pilote pour devenir directeur des courses de l'Alfa-Romeo.«cadreSF:3} «cadreFB:3} Entouré d'une quarantaine d'ouvriers, il est chargé de créer chez lui à Modène la " scuderia Ferrari ". Il frappe aussitôt ses voitures à ses armes : le " cheval rampant ", emblème d'un as de la première guerre disparu, au nom faussement prédestiné, le comte Baracca.«cadreAW:3} Ce cavallino noir sur champ d'or qui ornera pendant plus d'un demi-siècle les cockpits écarlates va connaitre des galops pas ordinaires. Le nouveau style des ces " Alfettes " à la ligne racée tranche avec les Bugatti trapues que nous allions admirer, parmi des foules énormes, roulant à tombeau ouvert sur l'anneau de Montlhéry. Encore fallait-il les tenir en main, ces joujoux magiques ! Or un pilote de grand prix ne se trouve pas sous le pied d'un cheval, même cabré. Antonio Ascari et Campari étant tombés au champ d'honneur des autodromes ? le premier à Montlhéry, le second à Monza, ? le choix prémonitoire d'Enzo Ferrari se porta sur deux futurs campionissimi : Tazio Nuvolari, le petit homme au masque de doge émacié, dévoré par la passion d'être toujours devant, forant l'espace, et son vivant contraste, Achille Varzi, un garibaldino méticuleux et peu communicatif, bouclé à double tour sur ses romanesques aventures sentimentales (1).«cadreSF:3} «cadreFB:3}


Input
OUT
GZIP Uncompress
DOC
SEM
Markup Alignment
REL
DOC
REL
Frame Evaluator
OUT
APL7
LPL7
IN Observer
IN Observer

Premiers résultats

Reference:	cadreman
Evaluated:	frame
Alignment relation:	alignement
Aligned frames:	72
Reference frames:	156
Measured frames:	80
Mean reference frame length:	119.88
Mean evaluated frame length:	174.88
Mean absolute left delta:	0.0
Mean relative left delta:	0.0
Mean positive left delta:	NaN (0.0%)
Mean negative left delta:	NaN (0.0%)
Min left delta:	0 (ref 8 vs. eval 3)
Max left delta:	0 (ref 8 vs. eval 3)
Mean absolute right delta:	47.59
Mean right delta:	47.59
Mean positive right delta:	100.79 (47.22%)
Mean negative right delta:	NaN (0.0%)
Min right delta:	0 (ref 13 vs. eval 5)
Max right delta:	518 (ref 98 vs. eval 51)


60%
50%
40%
30%
20%
10%
0%
0
1 à 10
10 à 100
> 100

Comparaison entre annotations manuelles

- Confrontation des annotations de trois “experts”
- Une journée du journal *Le Monde*
- Nombres de cadres annotés :
  - E_1 : 81 ; E_2 : 117 ; E_3 : 147
- Nombres d’alignements deux à deux :
  - E_1/E_2 : 76 ; E_1/E_3 : 76 ; E_2/E_3 : 108
- Gradation dans ce que les experts ont considéré comme “cadre temporel” : plutôt que des différences deux à deux, on trouve l’ensemble des annotations de l’un incluses dans celles de l’autre.

WWW.LINGUASTREAM.ORG