



---

# Les nouveautés d'Unitex 3.0beta

Sébastien Paumier

LIGM, Université Paris-Est

Séminaire du CENTAL, 4 novembre 2011



# Texte: offsets+formats

---

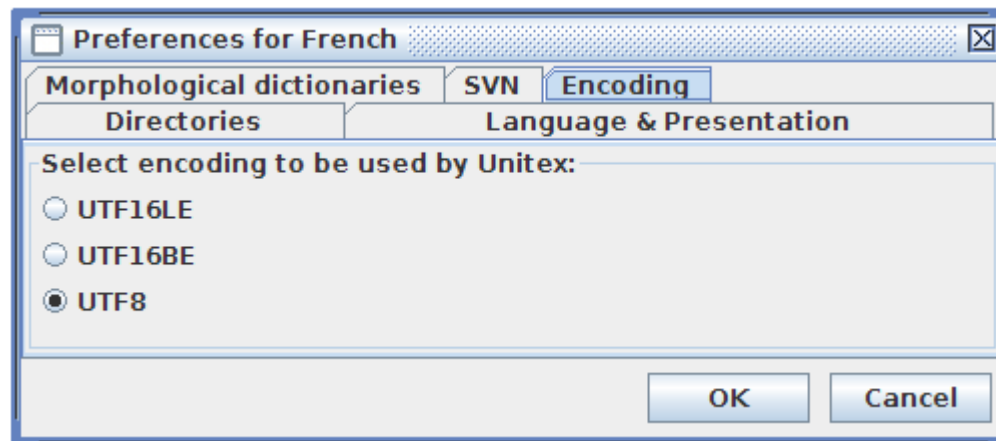
- on peut garder un historique de toutes les modifications faites par Unitex sur le texte d'entrée
- possibilité de synchroniser une concordance avec le texte d'origine
- possibilité de lire du html et du xml



# Encodage

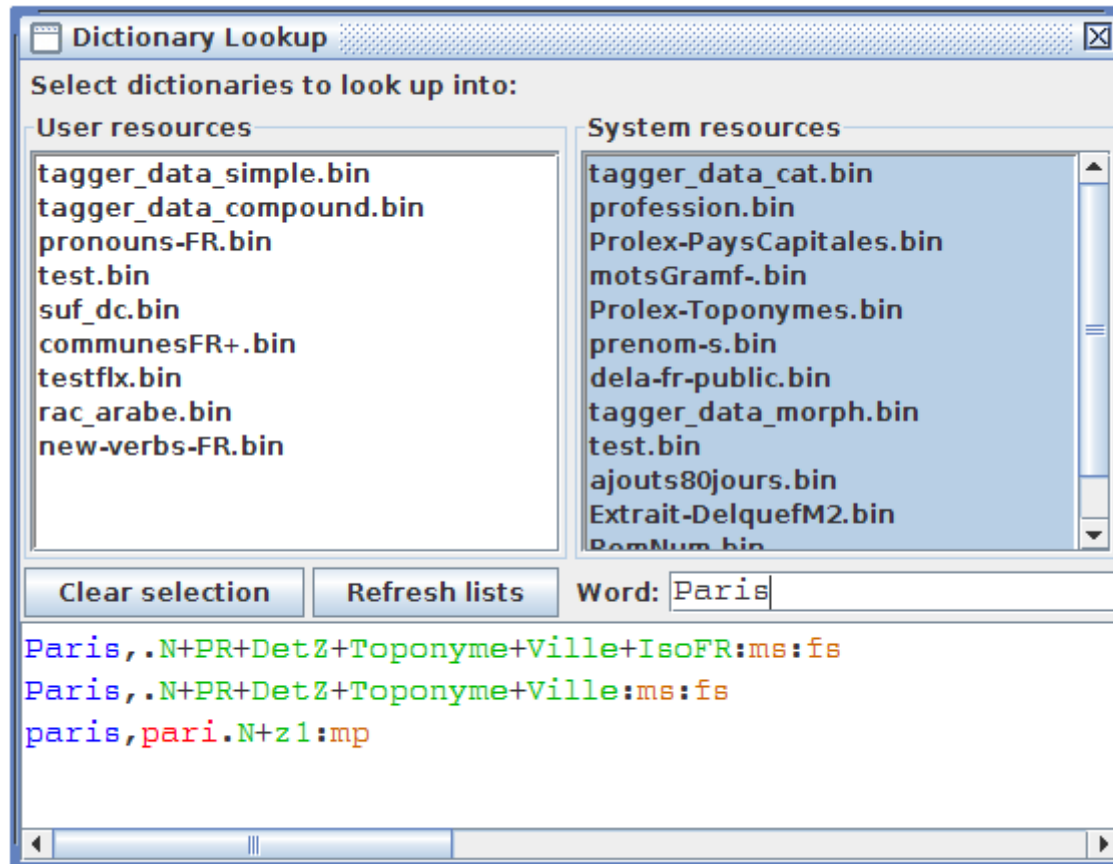
---

- utf8, et tous les utf16 sont supportés en entrée par tous les programmes
- l'encodage préféré pour les fichiers créés est paramétrable:





# Dico: lookup





# Dico: recherche

```
/home/paumier/unitex/French/Dela/dela-fr-public.dic
phtirius
phtalonitriles, phtalonitrile.N:mp
phtalyl, .PFX
phtaléine, .N:fs
phtaléines, phtaléine.N:fp
phtanite, .N:fs
phtanites, phtanite.N:fp
phtiriase, .N:fs
phtiriases, phtiriase.N:fp
phtiriasique, .A:ms:fs
phtiriasiques, phtiriasique.A:mp:fp
phtiriasis, .N:ms:mp
phtirius, .N:ms:mp
phtisie, .N+z2:fs
phtisie dorsale, .N+NA:fs
phtisie tuberculeuse, .N+NA:fs
phtisies, phtisie.N+z2:fp
phtisies galopantes, phtisie galopante.N+NA+z2:fp
phtisies ulcéreuses, phtisie ulcéreuse.N+NA:fp
phtisioqène, .A:ms:fs
phtisioqènes, phtisioqène.A:mp:fp
phtisiologie, .N:fs
phtisiologies, phtisiologie.N:fp
phtisiologique, .A:ms:fs
phtisiologiques, phtisiologique.A:mp:fp
```



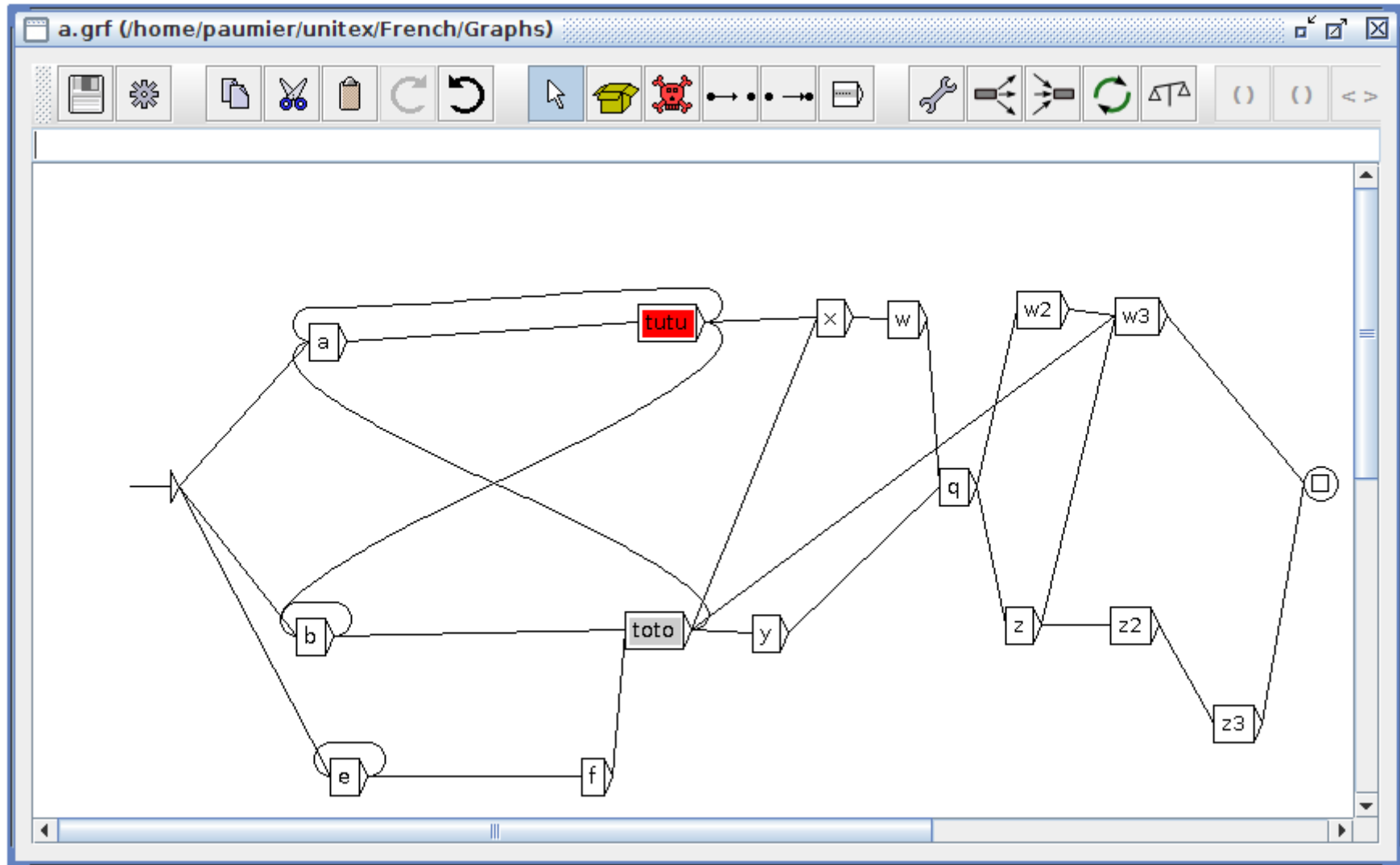
# Dico: formats

---

- nouveau format `.bin`:
  - plus compact
  - sans limite de taille
- format caché `.bin2`:
  - tout dans un seul fichier
  - plus économique quand on a beaucoup d'entrées avec beaucoup de codes uniques (ex: identifiants pour une ontologie)



# Graphes non trouvés





# Called graphs

Sentence.grf (/home/paumier/unitex/French/Graphs/Preprocessing/Sentence)

Called graphs:

- > AbrPoint
- > AbrPointMilFin
- > Abr\_nbAmb
- > LettreMaj
- > LettreMin
- > Millions
- > MotsComposesAvecMaj
- > MotsSuivisDeLettreMaj
- > NN
- > NenN
- > Nombres
- > PhTh
- > Prenoms
- > Symboles1Maj
- > abr\_nb
- cas2
- cas3
- cas4
- > crochets
- crochets
- > motifAnthro
- > motifSymboles
- > nb\_abr
- > parTel
- > parentheses
- parentheses
- > rois
- > sigles

Placement des marques de séparation de phrases {S}

Cas général Ponctuation

Ponctuation suivie de cas particuliers, noms, symboles...



# Caller graphs

Nombres.grf (/home/paumier/unitex/French/Graphs/Preprocessing/Sentence)

Caller graphs: ✕

- abr\_nb
- cas3
- cas4
- motifSymboles

Décrit les nombres contenant un point ou pas et les nombre

ex : "01-23-43-45-44"  
"1.000.000"

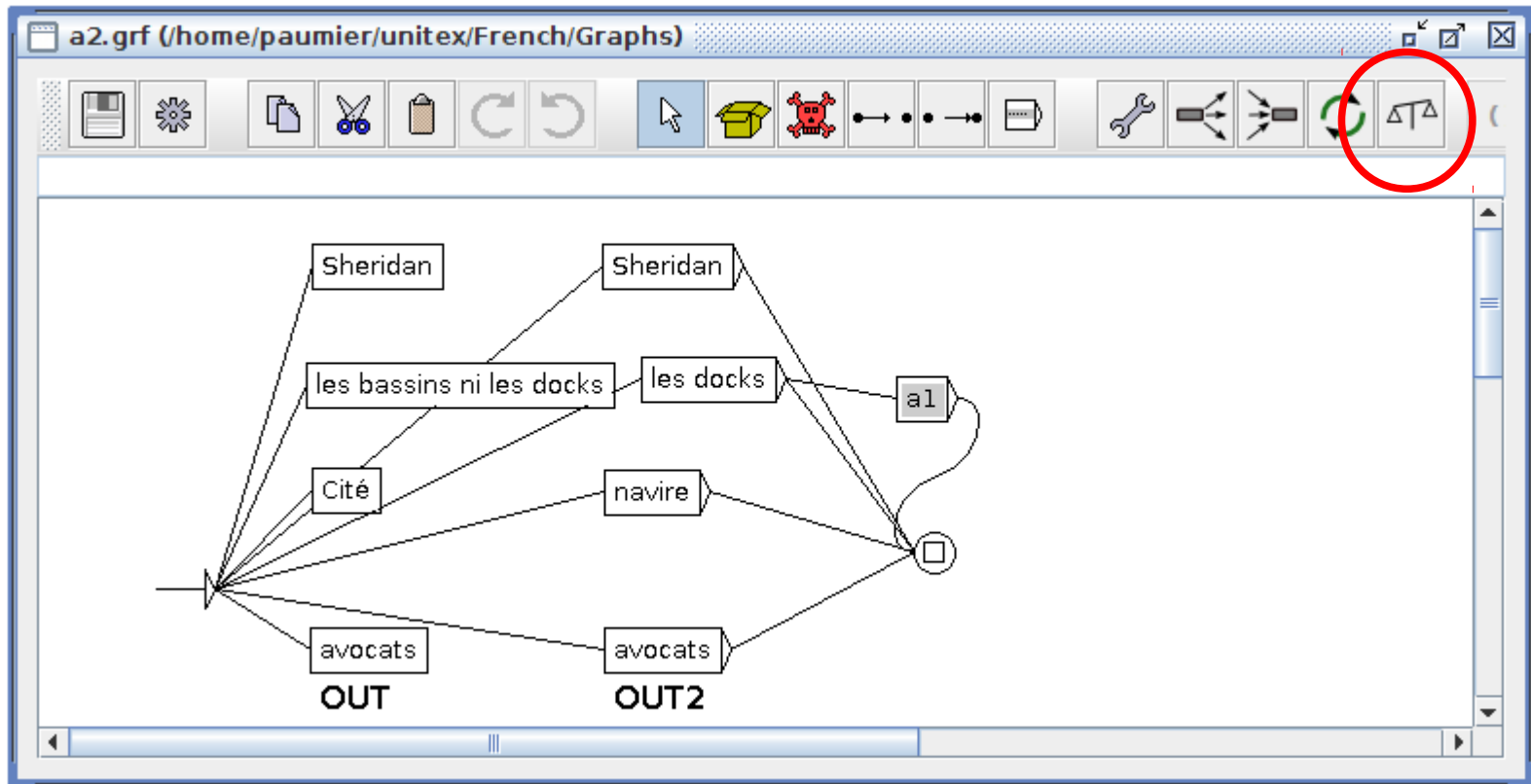


# (auto+ <E>) refresh

The screenshot shows a window titled "Nombres.grf (/home/paumier/unitex/French/Graphs/Preprocessing/Sentence)". The toolbar contains a refresh icon (a circular arrow) which is circled in red. Below the toolbar, there is red text: "Décrit les nombres contenant un point ou pas et les nombres contenant un t". A dialog box is open in the center with the message "Graph has changed on disk. Do you want to reload it ?" and "Yes" and "No" buttons. Below the dialog box, a graph diagram is shown with two nodes labeled "<NB>" connected by a line. A central box contains two patterns: "#.#" and "#-#".

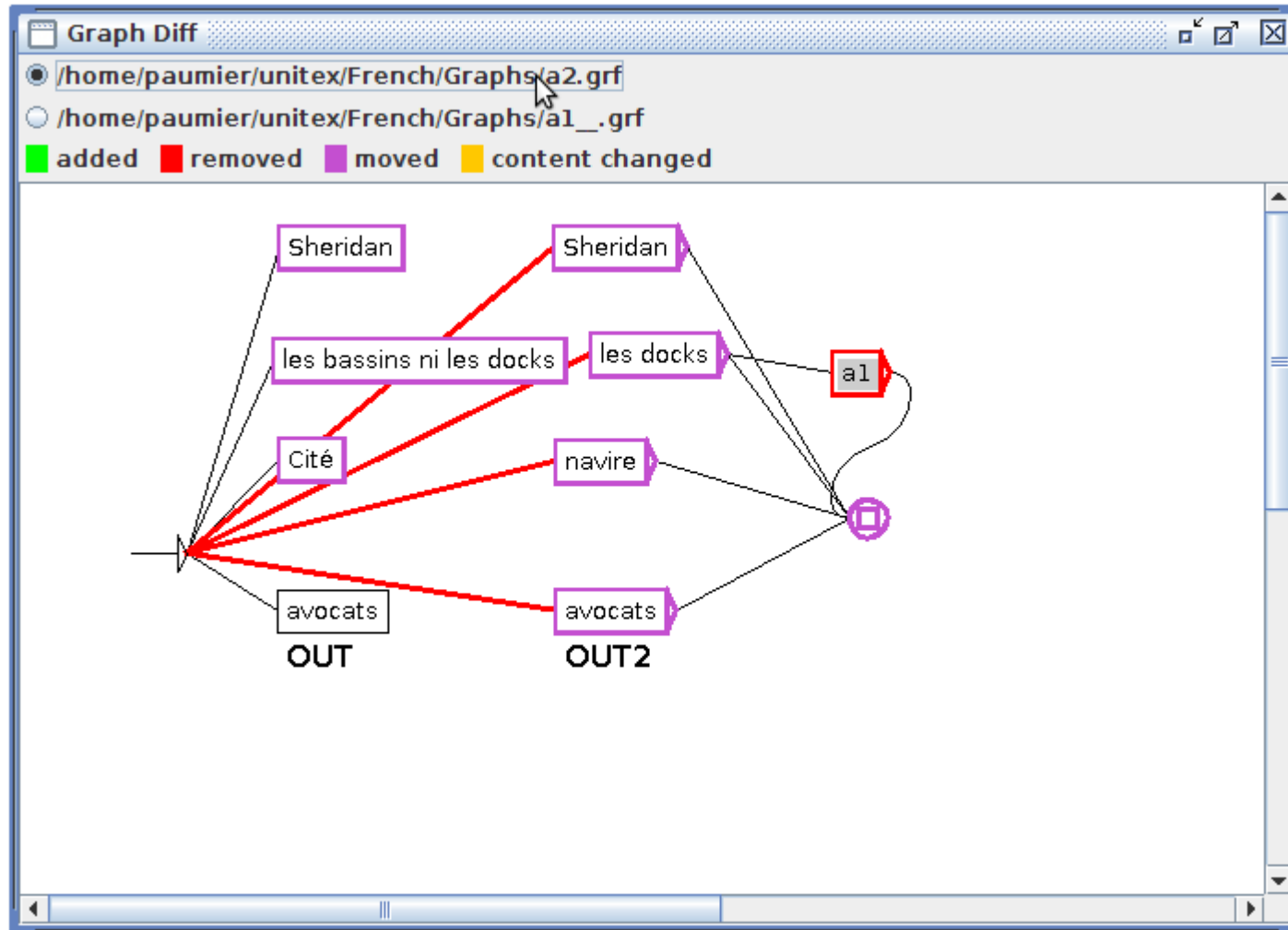


# SVN: GrfDiff



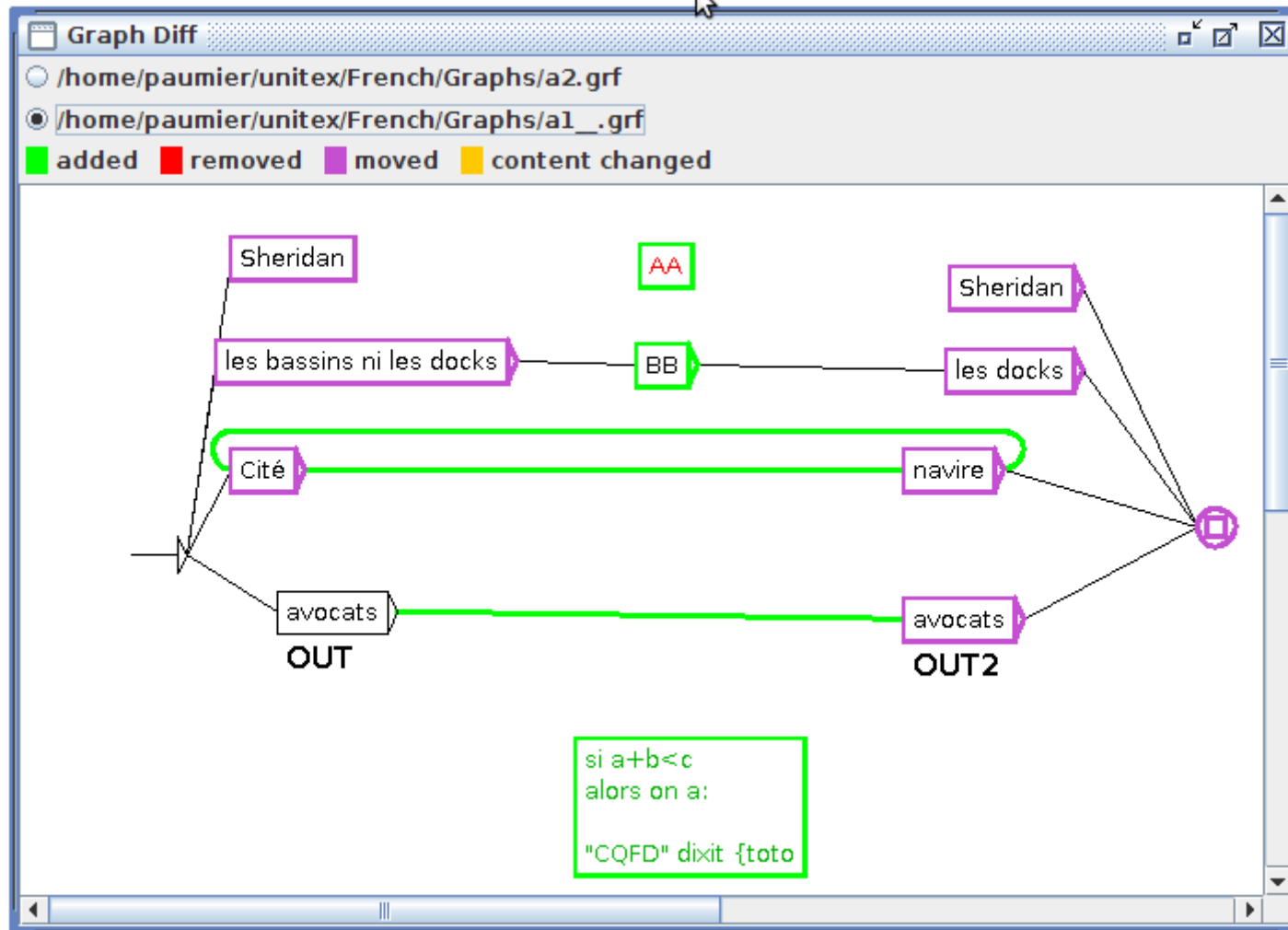


# SVN: GrfDiff





# SVN: GrfDiff







# Poids dans les graphes

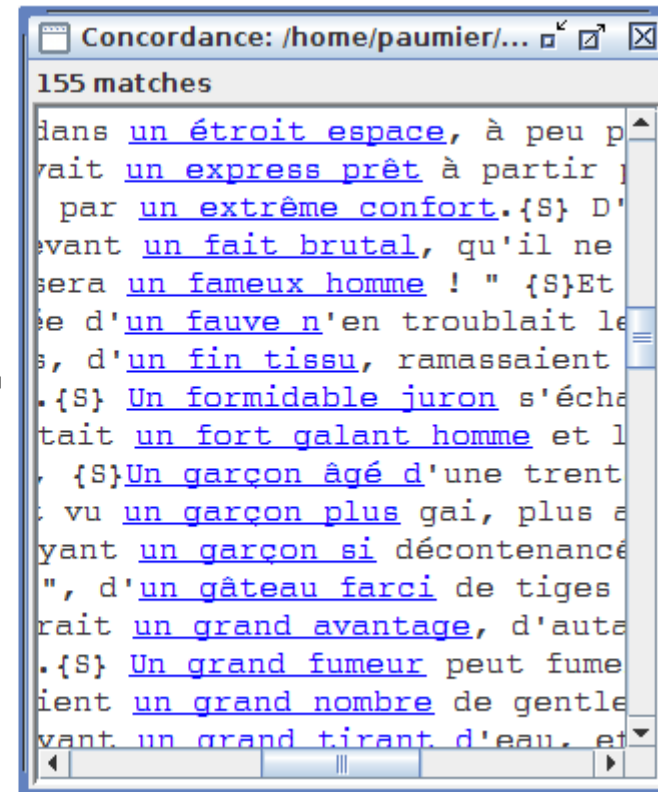
The screenshot shows a graph editor window titled "poids.grf (/home/paumier/unitex/French/Graphs)". The graph has a root node on the left, which branches into two nodes: "le" and "<DET>". The "le" node is associated with the label "\${1} \${cas particulier}" and the "<DET>" node is associated with "\${0} \${cas général}". Both nodes point to a final node on the right, which is a square with a circle inside.

Overlaid on the graph editor is a concordance window titled "Concordance: /home/paumier/unitex/French/Corpus/80j...". It displays "200 matches" and a list of text excerpts with blue underlines indicating matches for the graph nodes. The excerpts include:

- ... Fogg, esq., l'[cas général]un des membres les p...
- ... ention. {S}A l'[cas général]un des plus grands c...
- ... lant homme et l'[cas général]un des plus beaux g...
- ... CIPROQUEMENT L'[cas général]UN COMME MAÎTRE, L'AU...
- ... erait de ce qu'[cas général]un gentleman aussi m...
- ... lui un combat, [cas général]une lutte contre une...
- ... vert.{S} De là [cas général]une certaine « surfa...
- ... ne lutte contre [cas général]une difficulté, mais...
- ... e inspirées par [cas général]une seconde vue, tar...
- ... ne parût avoir [cas général]une connaissance spé...
- ... figuraient pour [cas général]une somme importante...
- ... an appoint pour [cas général]une chose noble, ut...
- ... ifficulté, mais [cas général]une lutte sans mouve...
- ... cercle.{S} Sur [cas général]vingt-quatre heures,
- ... pour gagner.{S} [cas particulier]Le jeu était po...
- ... ue suffisait à [cas particulier]le servir.{S} Déj...
- ... t l'honneur de [cas particulier]le connaître un...
- ... confortables que [cas particulier]le Reform-Club t...
- ... payés à vue par [cas particulier]le débit de son...
- ... rculaient dans [cas particulier]le club au sujet...
- ... ipalement dans [cas particulier]le but de détrui...



# Intervalles



[m,n]

[,n] (entre 0 et n)

[m,] (m ou plus)



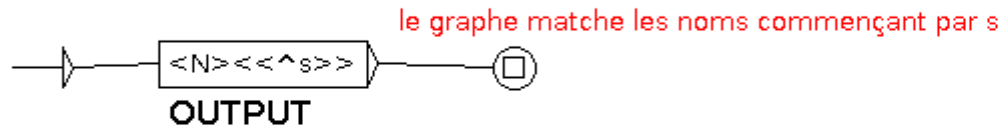
# GrfTest

```
@TEST:R:L@  
le<silence>est d'or  
OUTPUT
```

ce test passe

```
@TEST:R:L@  
le<silence>est d'or  
ERR
```

ce test échoue, car l'output n'est pas celui attendu par le test



```
@TEST:I:L@  
la<sortie>
```

ce test passe

```
@TEST:N:L@  
<salut> toi!
```

ce test échoue, car l'option de test N demande que la séquence entre angles ne soit pas reconnue, or "salut" est matché par le graphe

```
@TEST:M:L@  
la<sortie>
```

test invalide, car le mode M demande une troisième ligne contenant l'output attendu



# Mode debug

Concordance: /home/paumier/unitex/French/Corpus/80jours\_snt/concord.html

Tag	Output	Matched
<E>		
<ADV>		à minuit
<A>		précis
<E>		

202 matches

.{S} Cependant sa vie était à jour, mais ce qu'il faisait  
amais vu ni à la Bourse, ni à la Banque, ni dans aucun de  
chés comme ceux d'un soldat à la parade, les mains appuyé  
{S}En ce moment, on frappa à la porte du petit salon dan  
ez lui que pour se coucher, à minuit précis, sans jamais  
rcelaine spéciale et sur un admirable linge en toile de s  
ue ». dit-il. {S}Un garçon agé d'une trentaine d'années

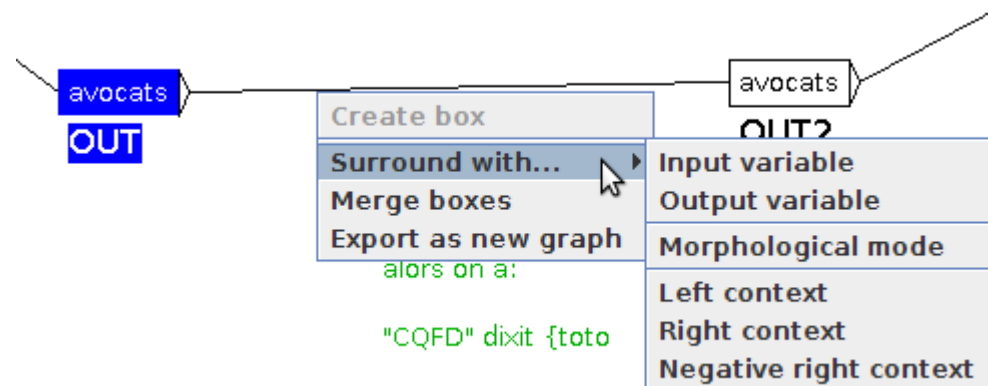
Double-click to open the graph:

The graph illustrates the syntactic dependencies between words in the text. Nodes represent words or phrases, and edges represent dependencies with associated weights. The 'si <ADV>' node is highlighted in green, indicating it is the current focus of the debug mode.



# Editeur de graphes

- sélection boîte par boîte avec Ctrl+Shift+clic
- support de la touche pomme sous Mac, en remplacement de Ctrl
- menu contextuel:





# Inclure des .fst2

---

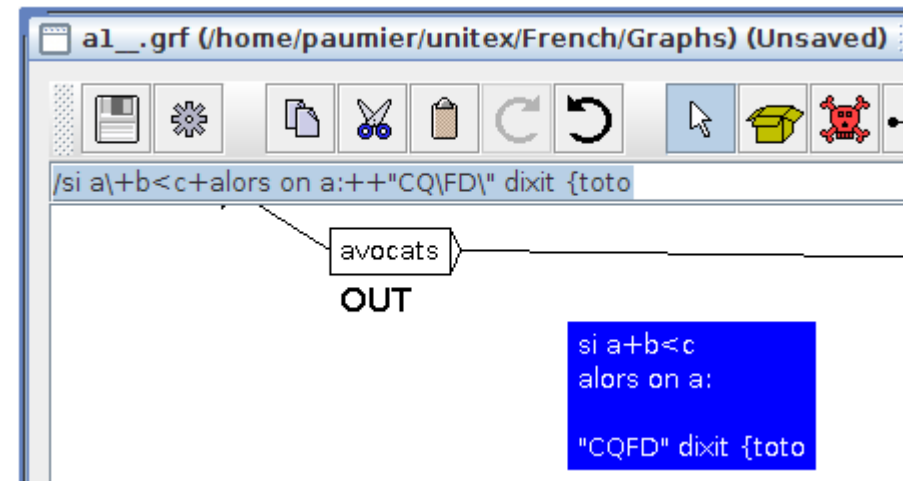
- si toto.grf n'est pas trouvé lors d'un appel à :toto, le compilateur cherche aussi toto.fst2
- possibilité de compiler une grammaire à partir de sous-grammaires précompilées





# Commentaires

- une boîte commençant par / est une "vraie" boîte de commentaire, affichée en vert:
  - possibilité de lignes vides
  - seul le + doit être protégé
  - pas de possibilité de transitions entrantes ou sortantes





# Divers

---

- dans Fst2Txt, Locate et LocateTfst, les tags {^} et {\$} permettent de matcher le début et la fin du texte
- option -v d'injection de variables de sortie pour Locate et LocateTfst
- option --only\_ambiguous de Concord
- option --no\_separator\_normalization de Normalize
- option -f de SortTxt



# Diff

- nouveau format cliquable pour les diffs de concordance:

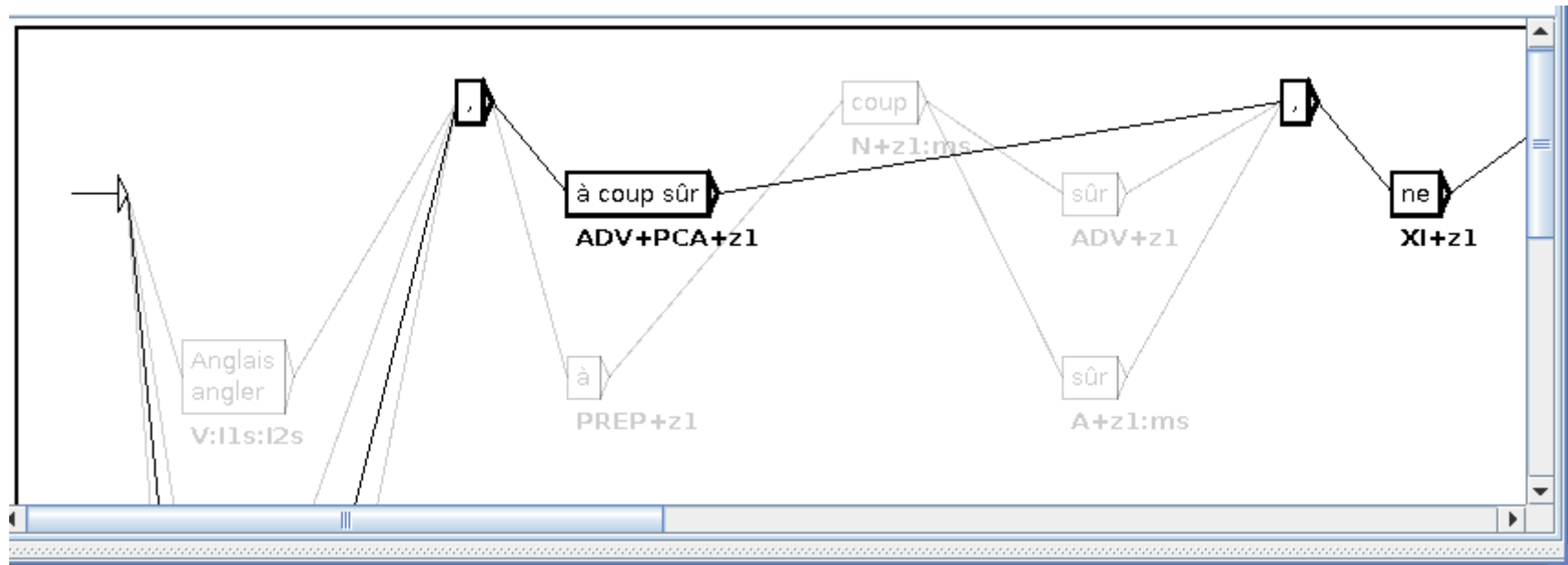
```
Concordance Diff
Violet: identical sequences with different outputs
Red: similar but different sequences
Green: sequences that occur in only one of the two concordances
Grey background=previous matches White background=new matches

{S}En l'année 1872, la[val=] maison portant le numéro 7 de S
{S}En l'année 1872, la[val=coucou] maison portant le numéro
beaux gentlemen de la[val=] haute société anglaise. {S}On d
beaux gentlemen de la[val=coucou] haute société anglaise. {
blait à Byron par la[val=] tête, car il était irréprochabl
blait à Byron par la[val=coucou] tête, car il était irrépr
vait jamais vu ni à la[val=] Bourse, ni à la Banque, ni dans
vait jamais vu ni à la[val=coucou] Bourse, ni à la Banque, n
i à la Bourse, ni à la[val=] Banque, ni dans aucun des compt
i à la Bourse, ni à la[val=coucou] Banque, ni dans aucun des
un des comptoirs de la[val=] Cité.{S} Ni les bassins ni les
un des comptoirs de la[val=coucou] Cité.{S} Ni les bassins n
s il ne plaïda ni à la[val=] Cour du chancelier, ni au Banc
```



# Tagging manuel

- élagage manuel facilité avec le clic droit dans l'automate du texte:





# JNI

---

- Unitex peut se construire sous forme de JNI
- possibilité conserver les .bin, les .fst2 et les fichiers alphabet en mémoire de façon persistante
- économie de 30% de temps sur des petits corpus traités à la chaîne
- cf. `TestUnitexJniPersistence.java`



# Pour finir...

---

- de nouveaux développements orientés vers l'industrialisation
- le projet GramLab:
  - un IDE pour les grammaires locales
  - utilisera les standards du développement: maven, svn, tests unitaires, tests de non regression, etc