جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

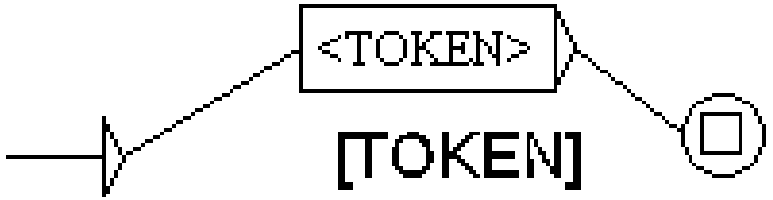
Extraction d'information

Cours 6

Nassim ZELLAL

Les sorties

- Il est possible d'associer une sortie à une boîte, en utilisant le caractère spécial ***/***



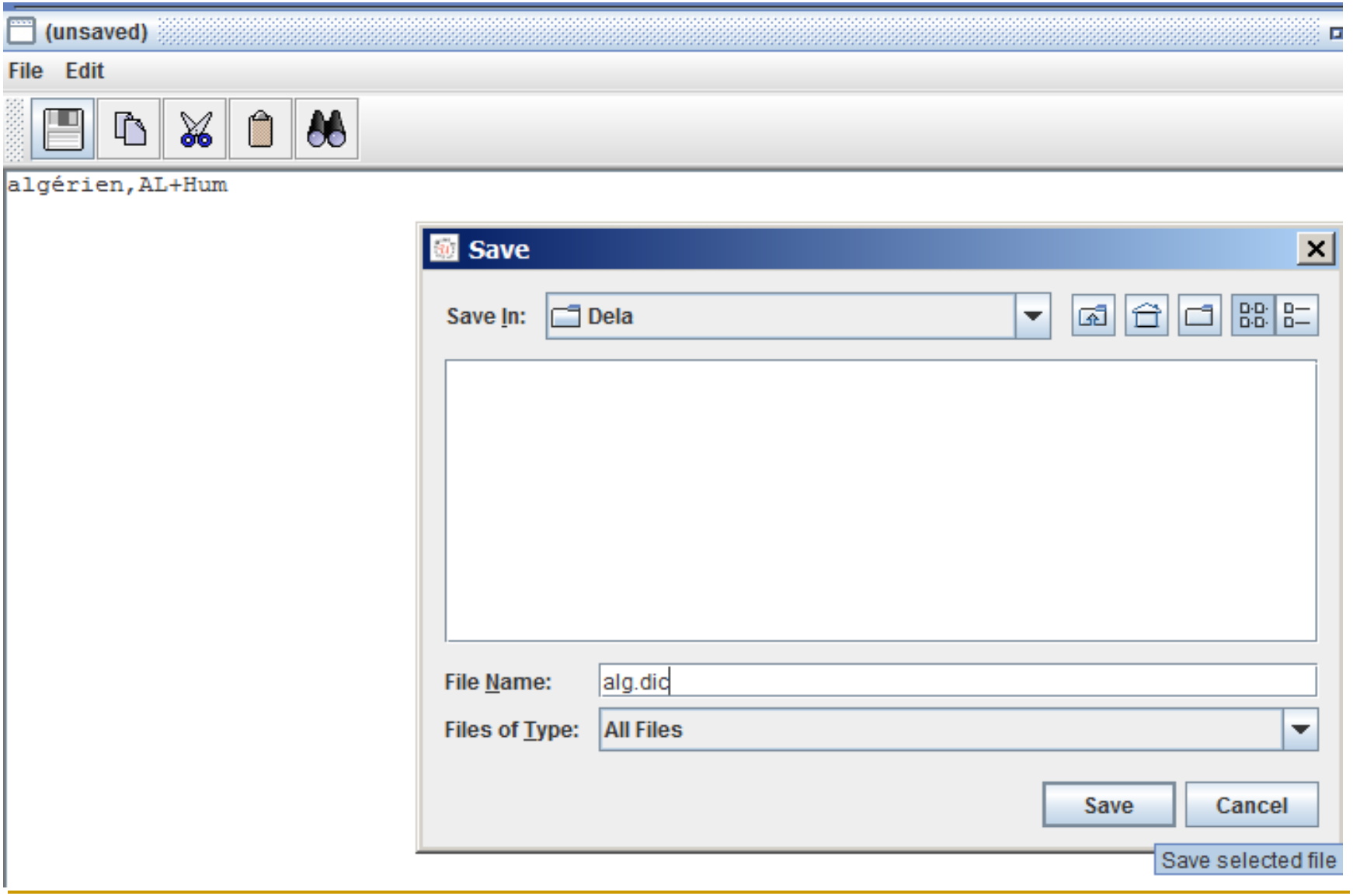
Pour créer cette sortie :

<TOKEN>***/***[TOKEN]

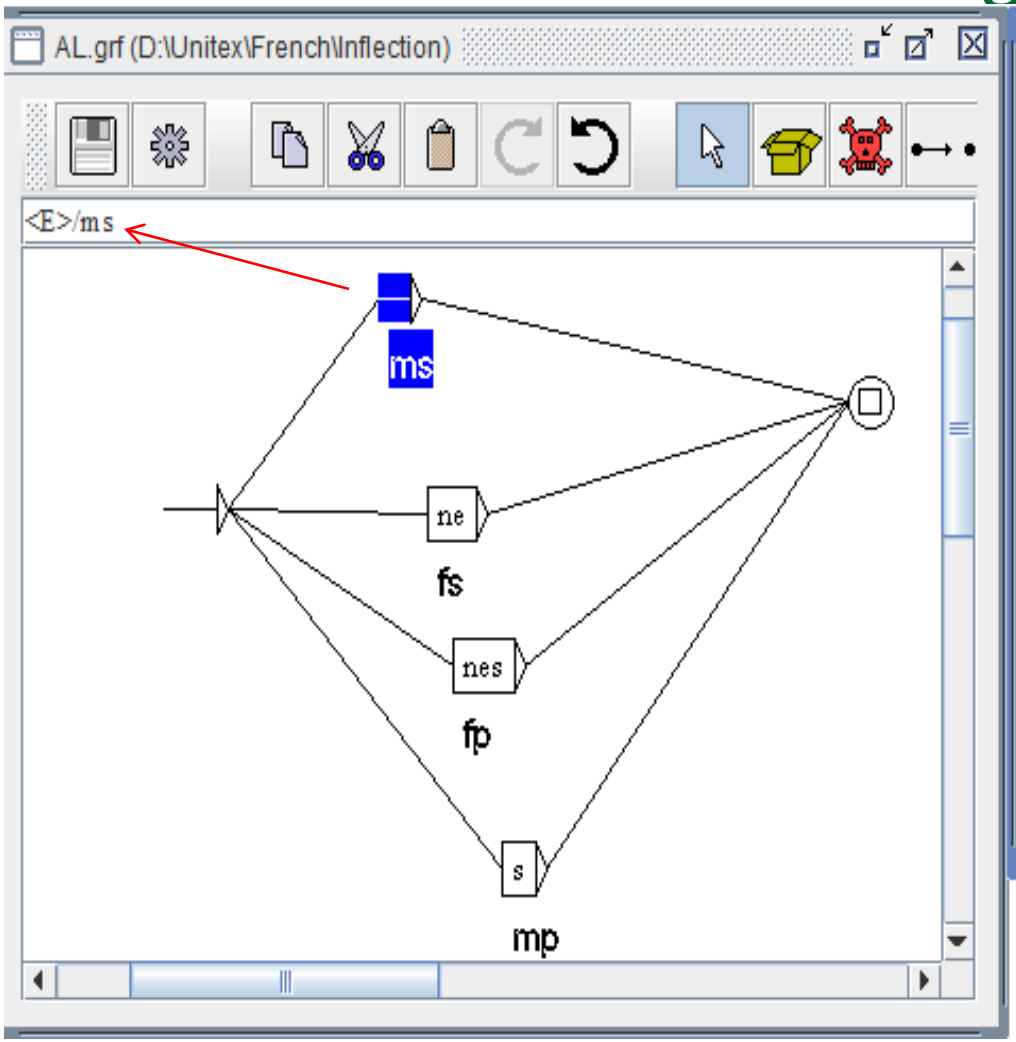
Format DELAS - flexion automatique

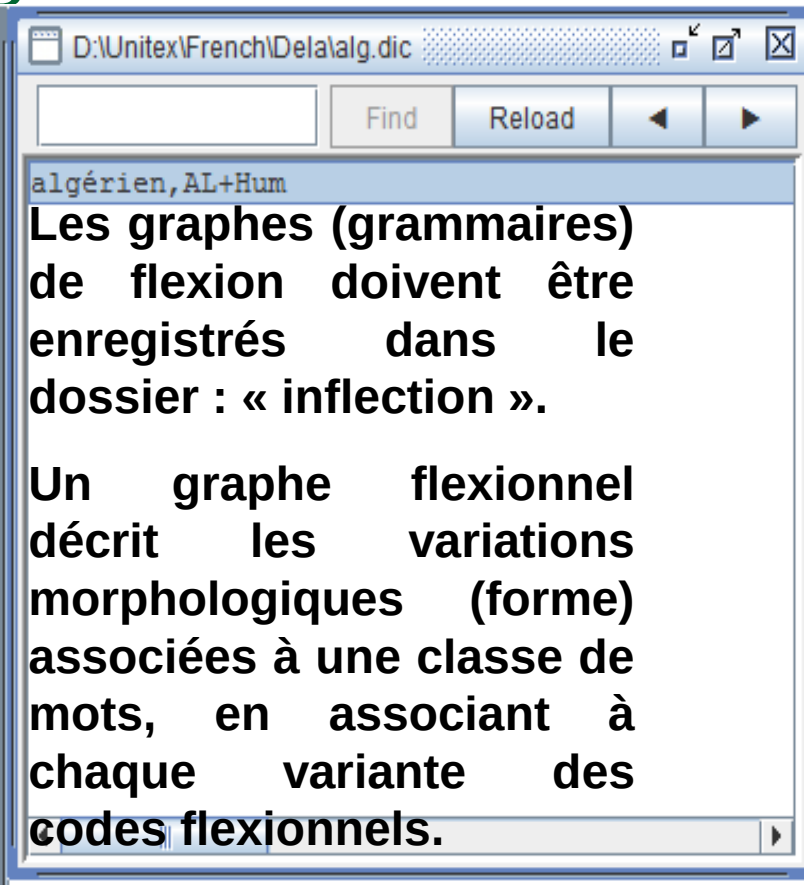
- **La flexion est un procédé morphologique fondamental consistant à adjoindre à la base d'un mot des désinences (terminaisons) : le cas, le genre, le nombre, la personne, le temps, le mode, l'aspect et la voix.**
- **Dans l'exemple qui suit, nous utilisons un « DELAS » pour la flexion du lemme « algérien ».**
- **algérien,AL+Hum**
- **Lemme,graphe de flexion+information sémantique (information facultative)**

Création d'un DELAS



Création du graphe de flexion « AL.grf » et du DELAS « alg.dic »

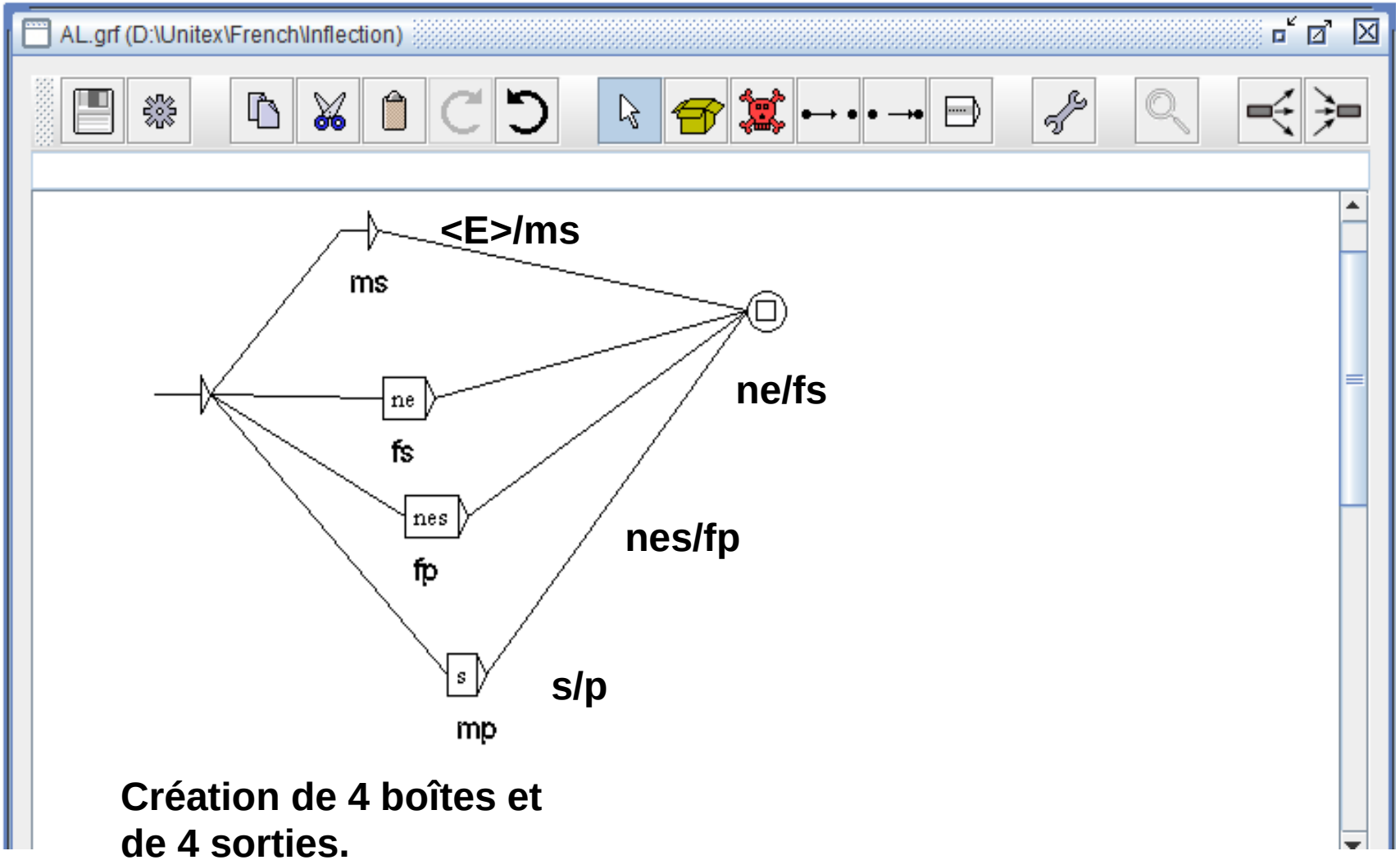




Les graphes (grammaires) de flexion doivent être enregistrés dans le dossier : « inflection ».

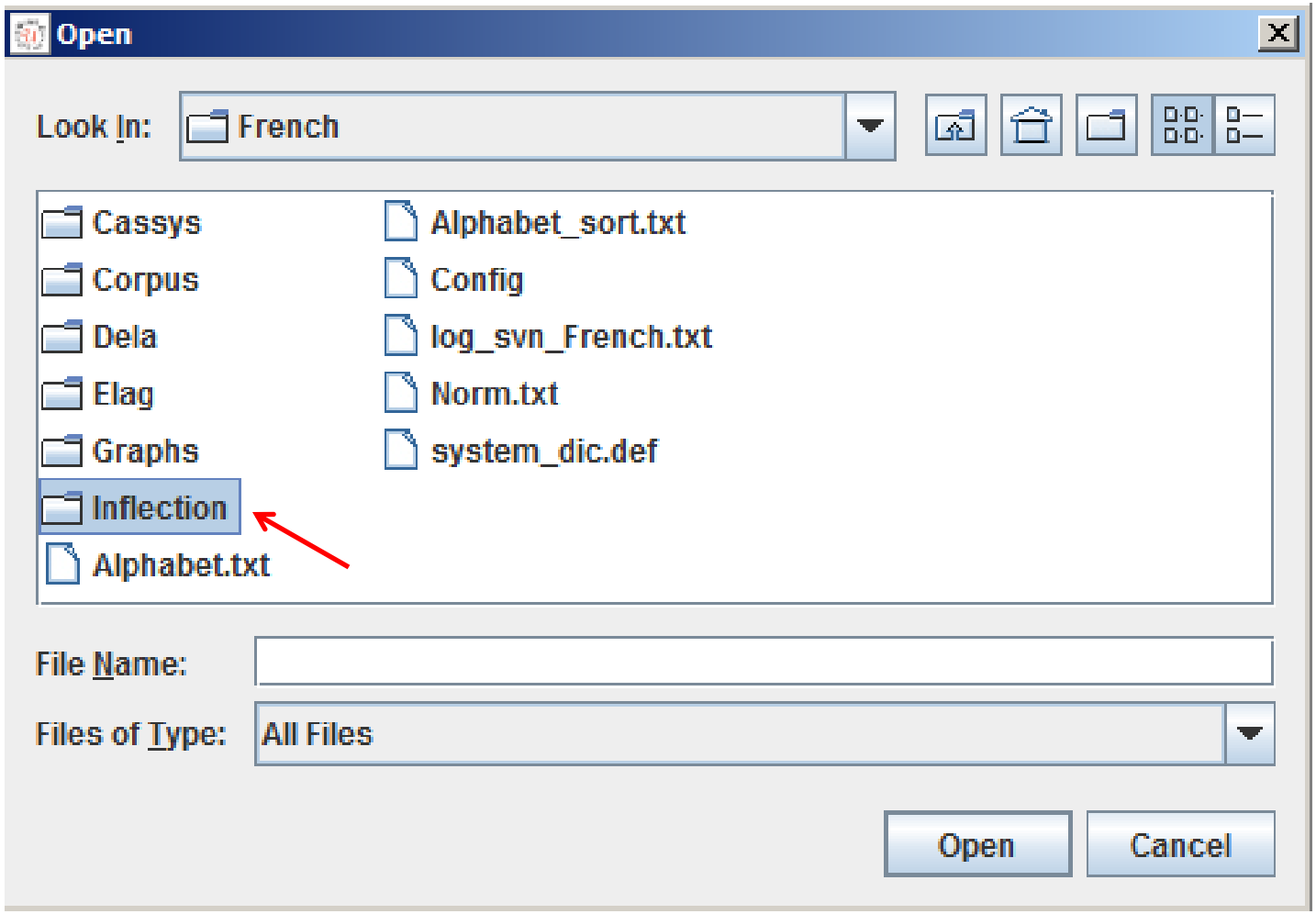
Un graphe flexionnel décrit les variations morphologiques (forme) associées à une classe de mots, en associant à chaque variante des codes flexionnels.

Création du graphe de flexion « AL.grf »

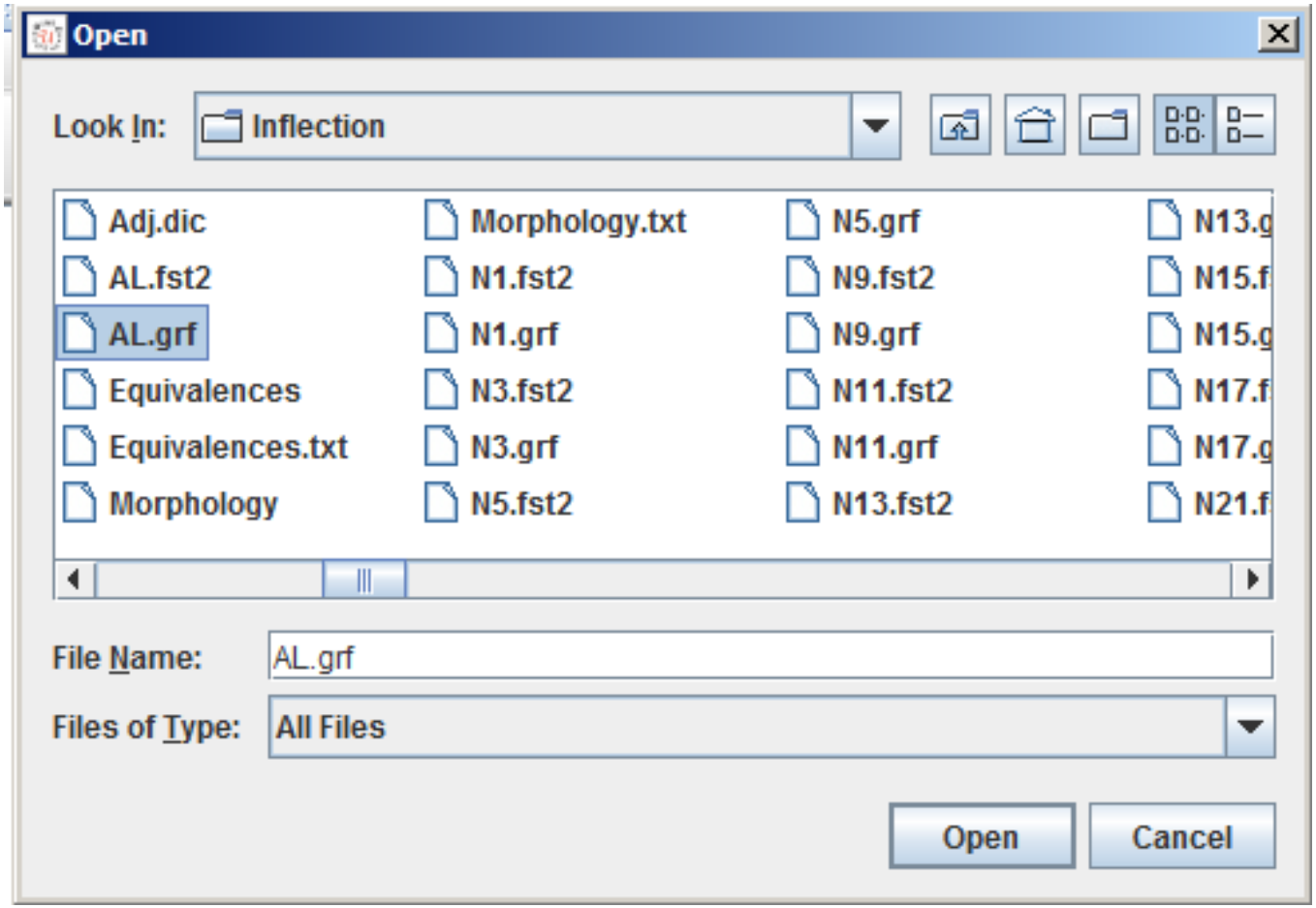


Création de 4 boîtes et de 4 sorties.

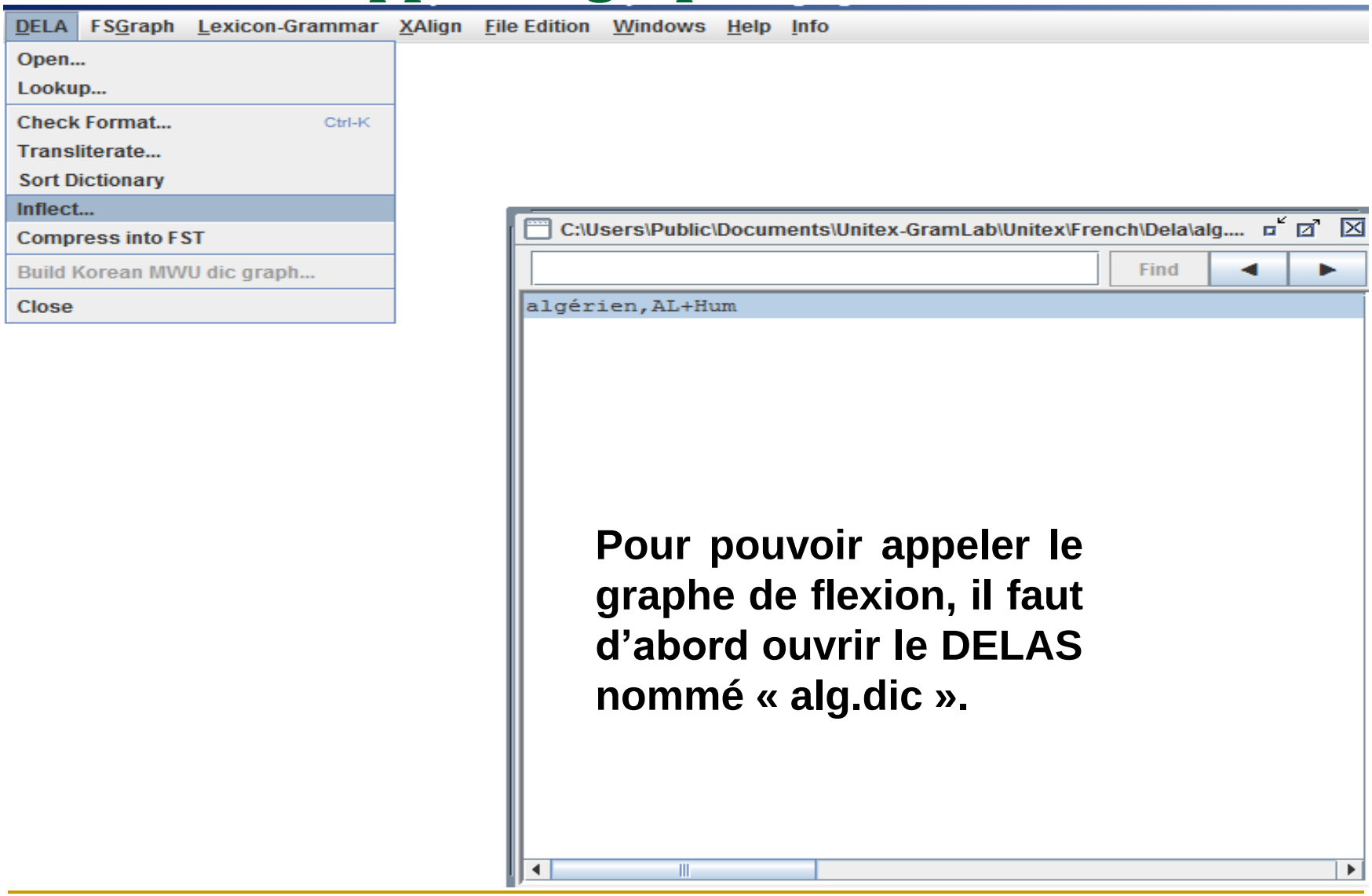
Enregistrement du graphe de flexion



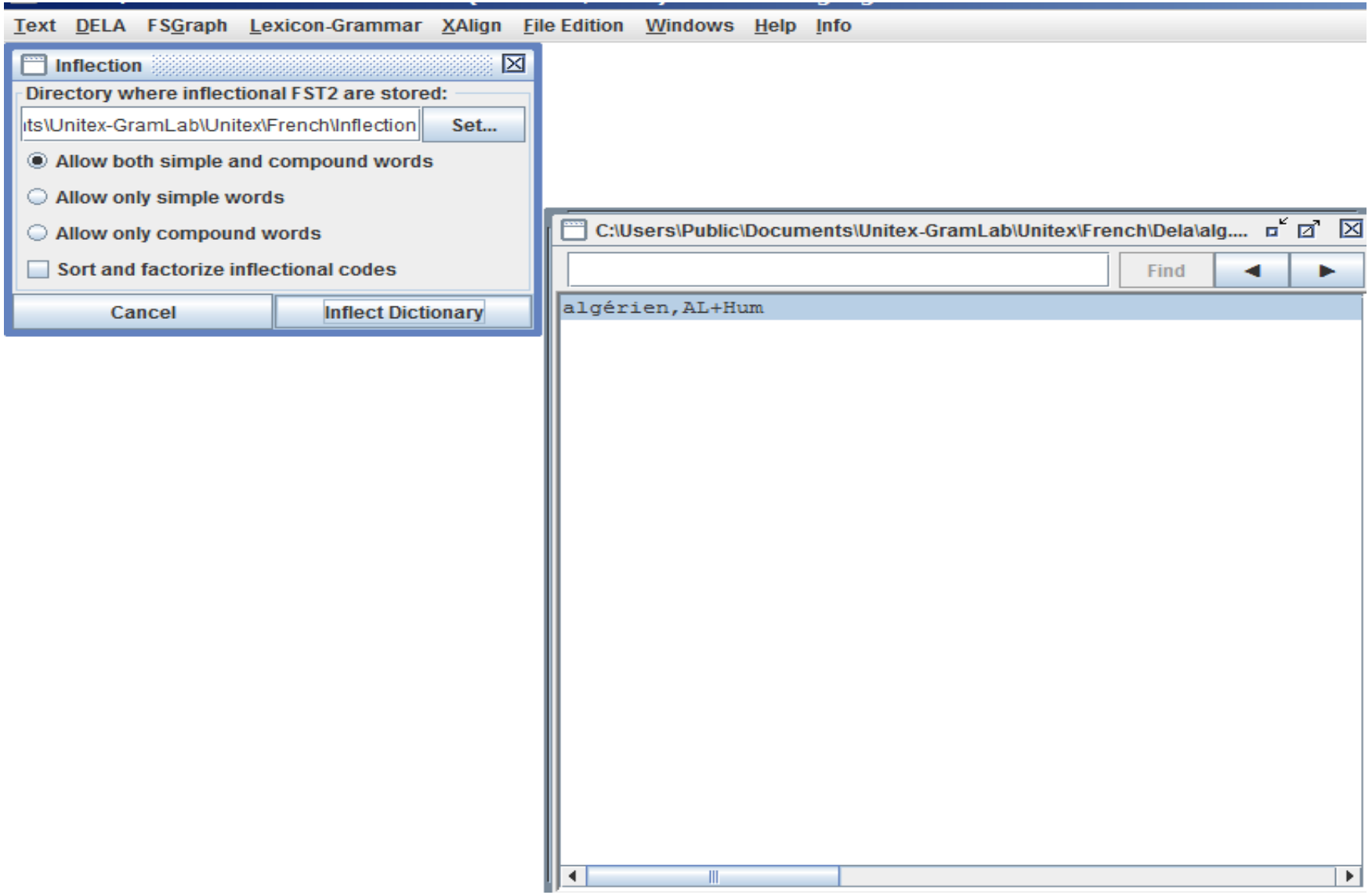
Enregistrement du graphe de flexion



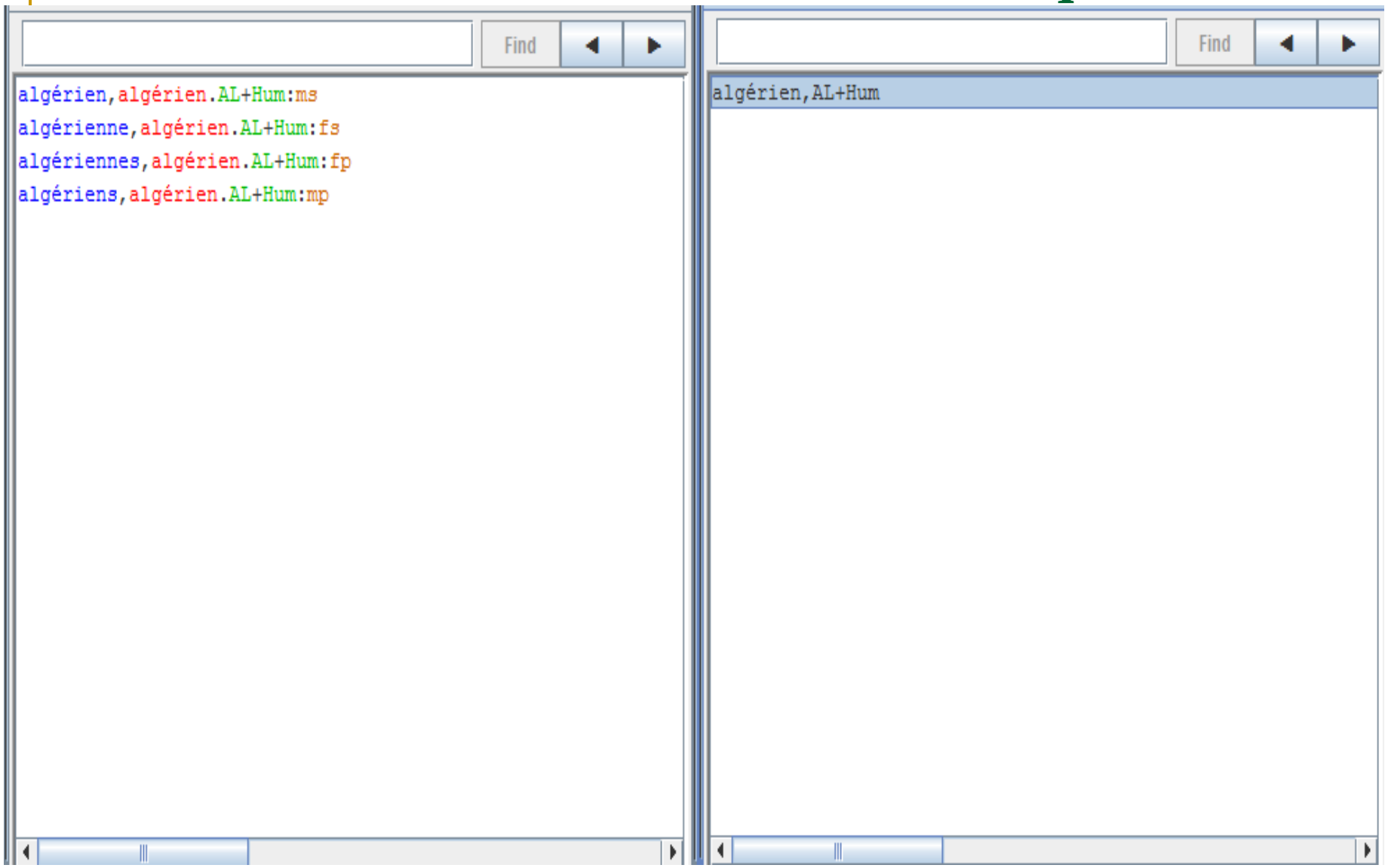
Appel du graphe de flexion



Appel du graphe de flexion



Résultat de la flexion automatique

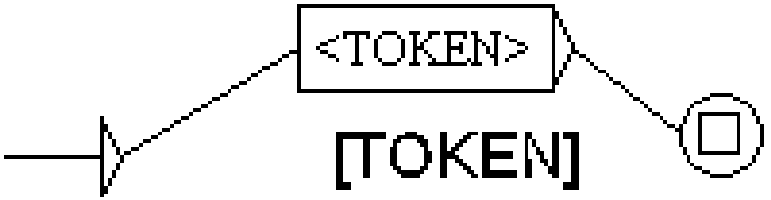


Les sorties - « Merge » et « Replace »

- **Mode Replace (substitution/remplacement) :** permet de remplacer les séquences reconnues dans le texte par les **séquences produites par les « sorties »** du **transducteur**.
- En d'autres termes, les séquences reconnues par le **transducteur** seront remplacées par les **séquences produites** par celui-ci.

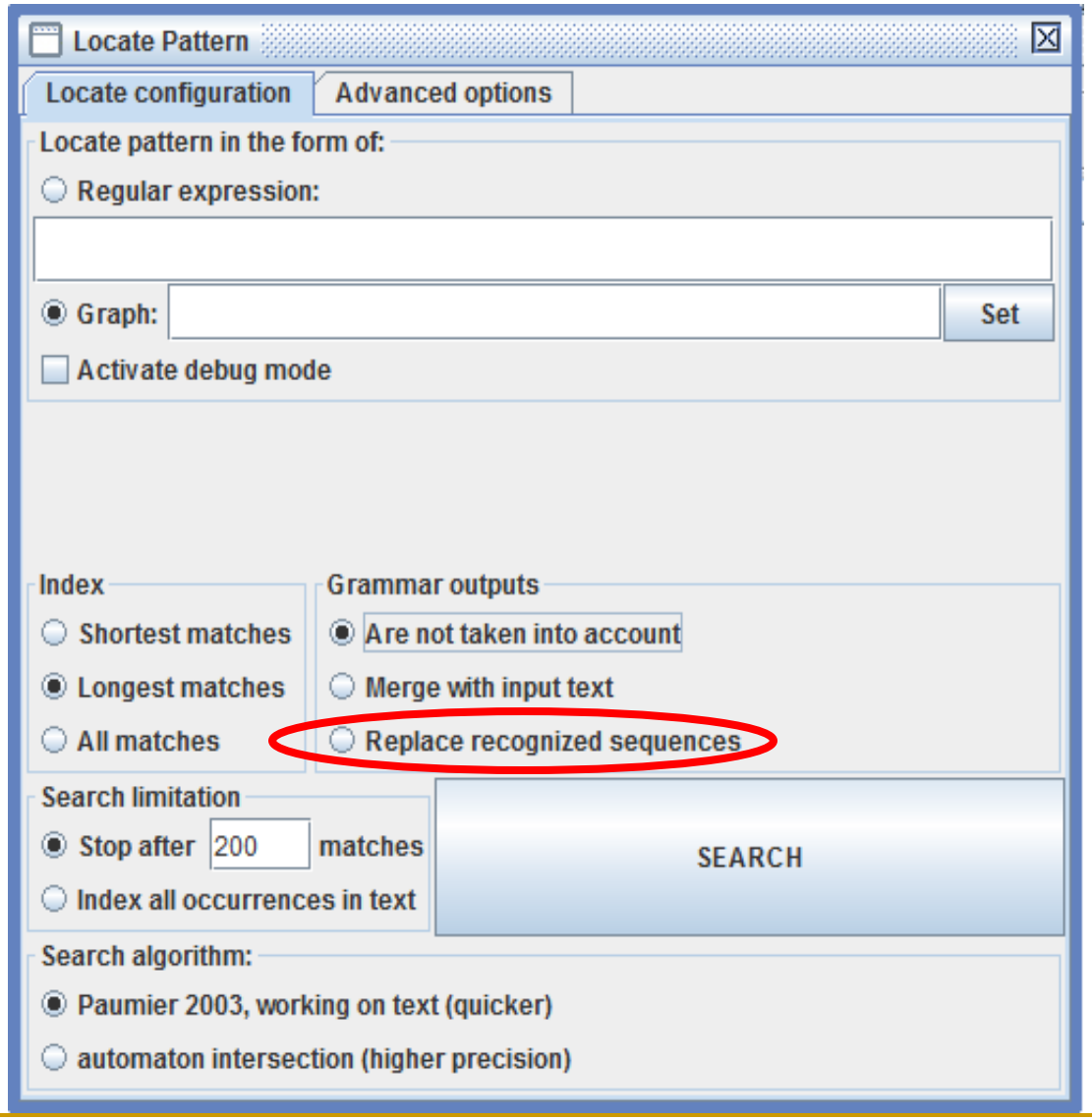
- **Mode Merge (fusion/insertion) :** permet d'insérer avant ou après les séquences reconnues dans le texte, les **séquences produites par les « sorties »** du **transducteur**.
- **Remarque : Lorsqu'une grammaire/graphe produit une « sortie », on utilise le terme de « transducteur ».**

Les sorties - « Merge » et « Replace »



Pour créer ce graphe avec cette sortie, écrire dans la zone de saisie :
<TOKEN>/ [TOKEN]

Mode « Replace »



Sortie avec le mode « Replace »

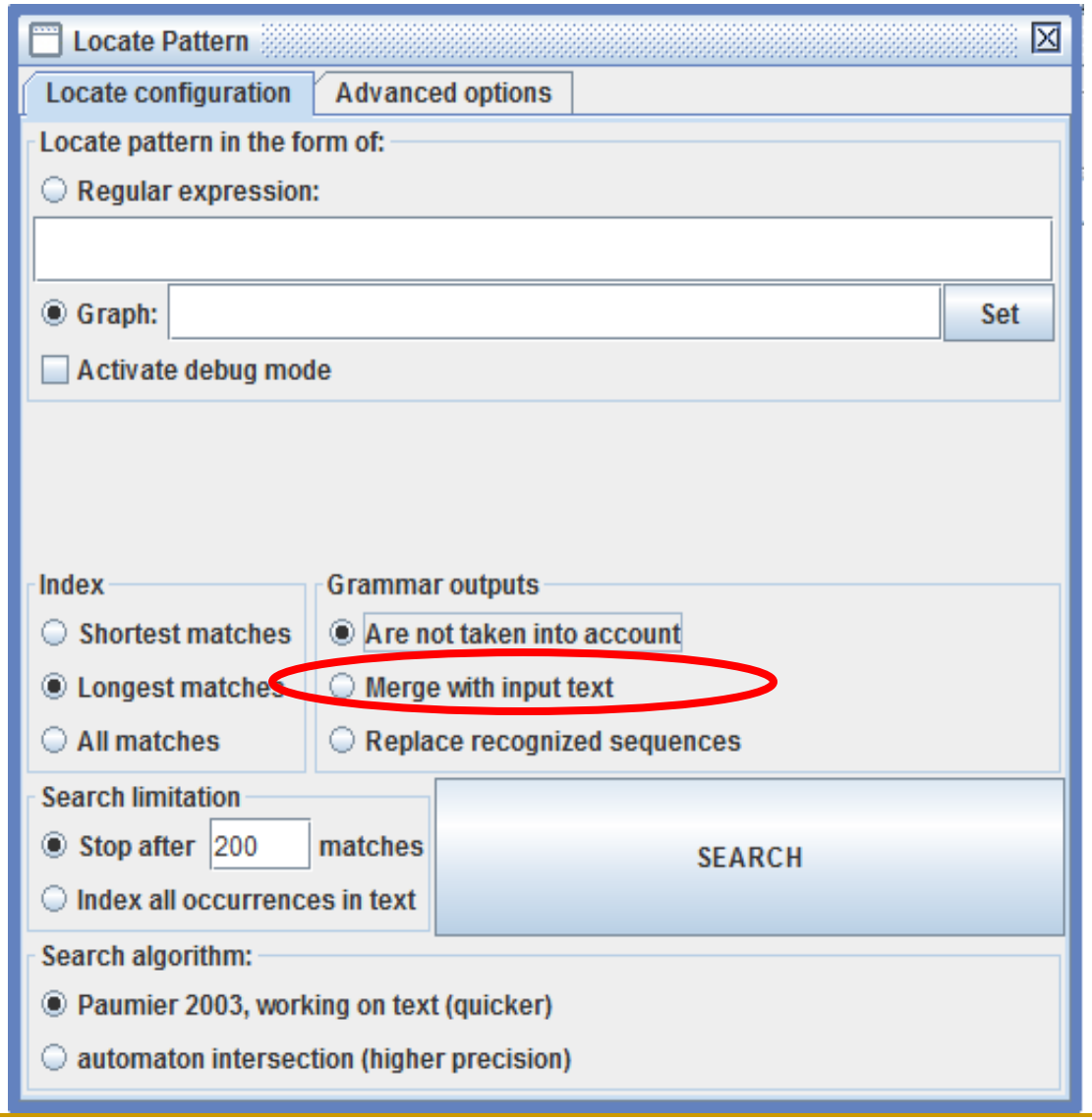
```
 Burlington Gardens _ maison dans  [TOKEN] Sheridan mourut en 1814
                Chapitre I {S}DANS  [TOKEN] PHILEAS FOGG ET PASSEPA
e Saville-row, Burlington Gardens  [TOKEN] maison dans laquelle Sh
n des membres les plus singuliers  [TOKEN] les plus remarqués du R
S}A l'un des plus grands orateurs  [TOKEN] honorent l'Angleterre,
r l'attention. {S}A l'un des plus  [TOKEN] orateurs qui honorent l
ort galant homme et l'un des plus  [TOKEN] gentlemen de la haute s
, esq., l'un des membres les plus  [TOKEN] et les plus remarqués d
s les plus singuliers et les plus  [TOKEN] du Reform-Club de Londr
ub de Londres, bien qu'il semblât  [TOKEN] à tâche de ne rien fair
```

80jours.snt (C:\Users\user\Documents\Unitex-GramLab\Unitex\French\Corpus)

3652 sentence delimiters, 165239 (9452 diff) tokens, 71859 (9422) simple forms, 438 (10) digits
67702 occurrences (12355 DLF entries) simple words, 0 occurrence (0 DLC entries) compound word, 4157 occurren

```
{S}A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Ph
personnage énigmatique, dont on ne savait rien, sinon que c'était un fort galant h
l'un des plus beaux gentlemen de la haute société anglaise.
{S}On disait qu'il ressemblait à Byron _ par la tête, car il était irréprochable q
pieds _ , mais un Byron à moustaches et à favoris, un Byron impassible, qui aurait
```

Mode « Merge »


Locate Pattern
Locate configuration
Advanced options
Locate pattern in the form of:
Regular expression:
Graph:
Set
Activate debug mode
Index
Shortest matches
Longest matches
All matches
Grammar outputs
Are not taken into account
Merge with input text
Replace recognized sequences
Search limitation
Stop after 200 matches
Index all occurrences in text
SEARCH
Search algorithm:
Paumier 2003, working on text (quicker)
automaton intersection (higher precision)

Sortie avec le mode « Merge »

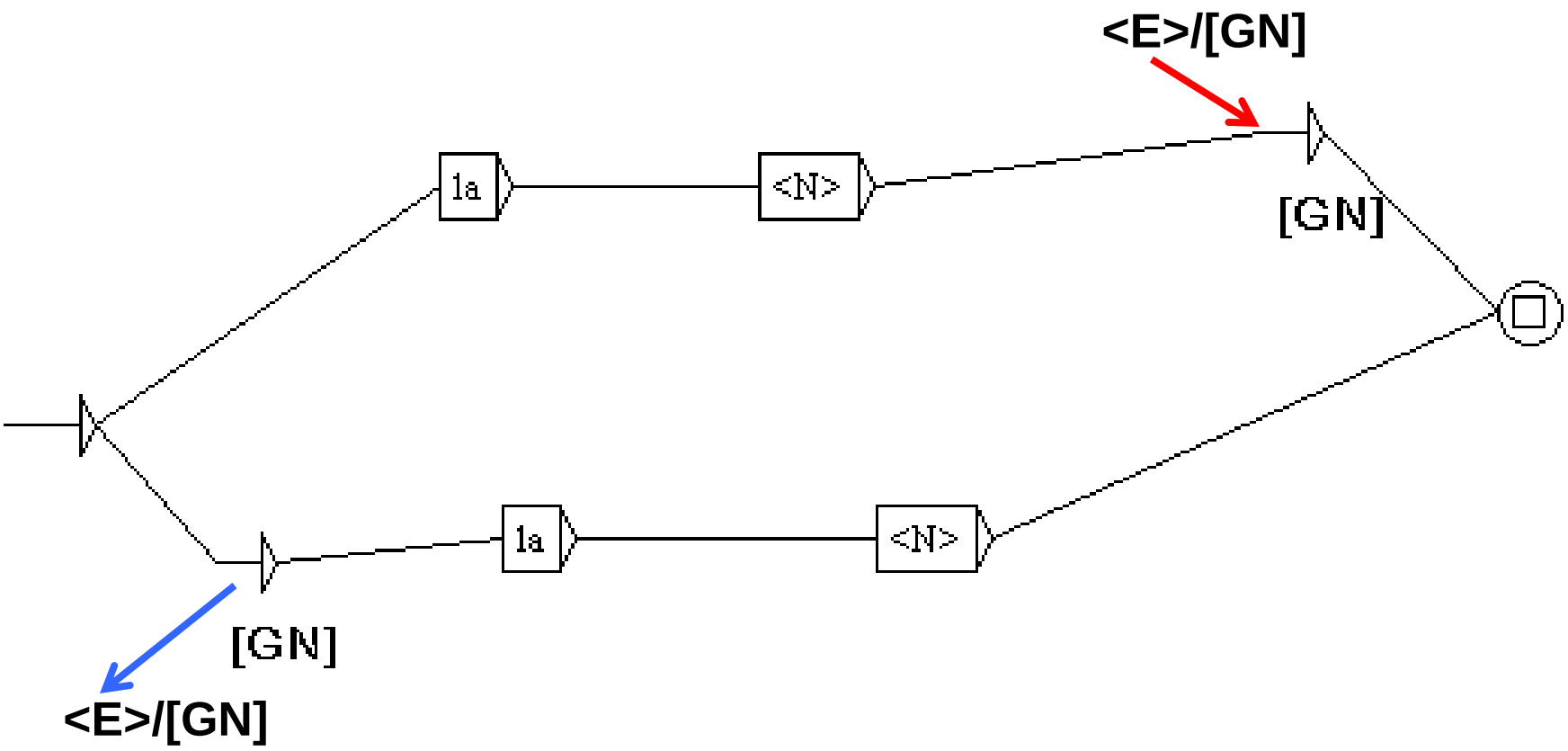
aville-row, Burlington Gardens _ maison [TOKEN]dans laquelle Sheridan mourut en 18
Chapitre I {S} [TOKEN]DANS LEQUEL PHILEAS FOGG ET PASSEPA
née 1872, la maison portant le numéro 7 [TOKEN]de Saville-row, Burlington Gardens
rs et les plus remarqués du Reform-Club [TOKEN]de Londres, bien qu'il semblât pren
res, bien qu'il semblât prendre à tâche [TOKEN]de ne rien faire qui pût attirer l'
homme et l'un des plus beaux gentlemen [TOKEN]de la haute société anglaise. {S}On
it habitée par Phileas Fogg, esq., l'un [TOKEN]des membres les plus singuliers et
qui pût attirer l'attention. {S}A l'un [TOKEN]des plus grands orateurs qui honore
ue c'était un fort galant homme et l'un [TOKEN]des plus beaux gentlemen de la haut
men de la haute société anglaise. {S}On [TOKEN]disait qu'il ressemblait à Byron

80jours.snt (C:\Users\user\Documents\Unitex-GramLab\Unitex\French\Corpus)

3652 sentence delimiters, 165239 (9452 diff) tokens, 71859 (9422) simple forms, 438 (10) digits
67702 occurrences (12355 DLF entries) simple words, 0 occurrence (0 DLC entries) compound word, 4157 occurrences (457 ERR lines...

{S}A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Phileas Fogg,
personnage énigmatique, dont on ne savait rien, sinon que c'était un fort galant homme et
l'un des plus beaux gentlemen de la haute société anglaise.
{S}On disait qu'il ressemblait à Byron _ par la tête, car il était irréprochable quant aux
pieds _ , mais un Byron à moustaches et à favoris, un Byron impassible, qui aurait vécu mille
ans sans vieillir.
{S}Anglais, à coup sûr, Phileas Fogg n'était peut-être pas Londonner.{S} On ne l'avait jamais

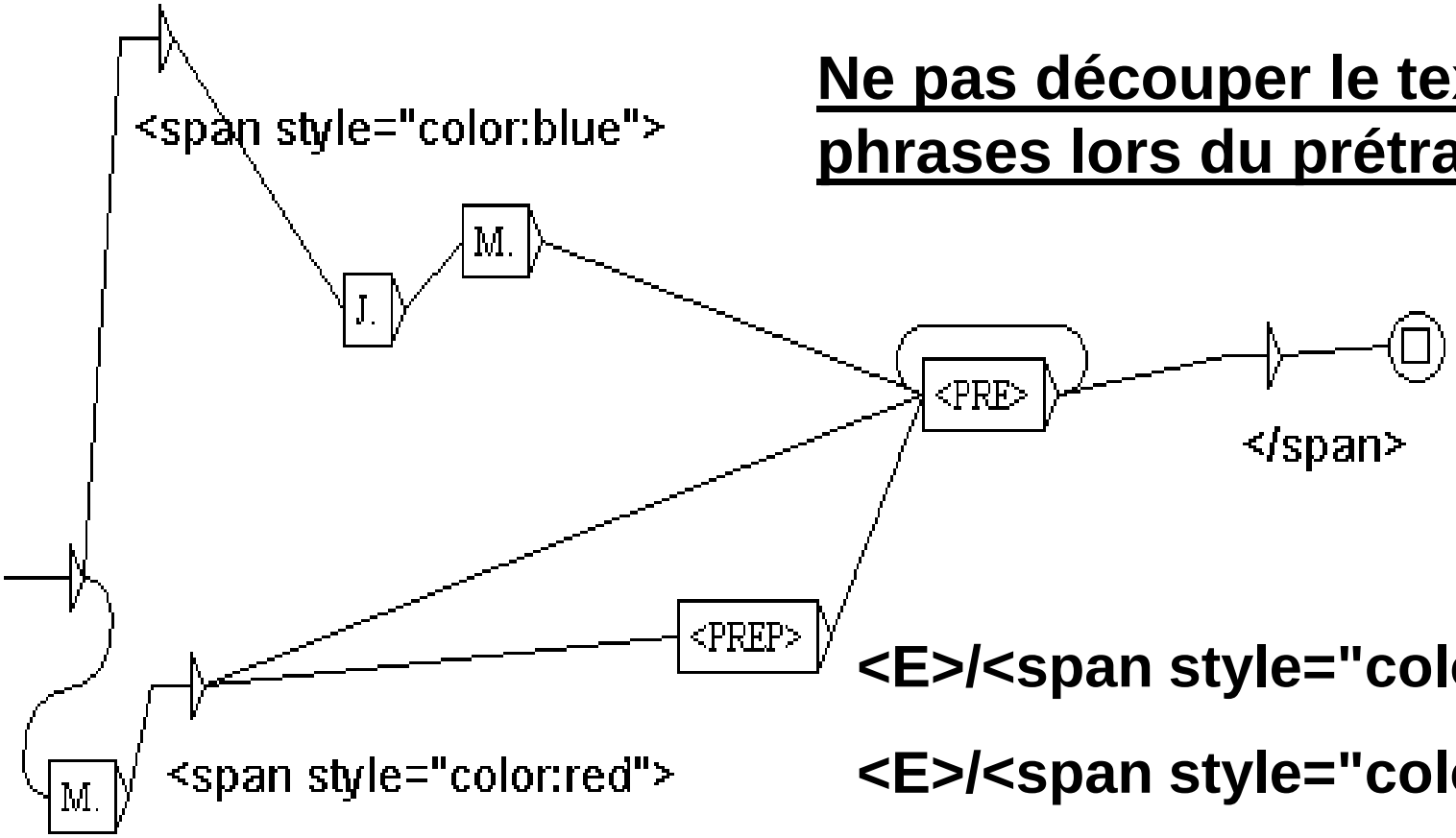
Sorties avec le mode « Merge »



Résultat de l'insertion (Merge)

:e et que! voulant gouter de [GN]la vie de famille! je suis v
: le panorama pittoresque de [GN]la ville ; mais la plupart d
se dessinaient au-dessus de [GN]la ville sous les pâles rayo
)tion passa aux journaux par [GN]la voie des reporters! et de
reprit Mr. Fogg sans élever [GN]la voix davantage. Passepart
: dans un cirque! faisant de [GN]la voltige comme Léotard! et
ıl auquel concourent l'ouïe! [GN]la vue et l'odorat. J'ai arr
.en! répondit Andrew Stuart! la Banque[GN] en sera pour son a
;s la fermeture des bureaux! la Banque[GN] d'Angleterre n'ava
ıais vu ni à la Bourse! ni à la Banque[GN]! ni dans aucun des
)orts! après le vol commis à la Banque[GN] d'Angleterre. Ce F
:te du caissier principal de la Banque[GN] d'Angleterre. A qu
;ci : Dans une des salles de la Banque[GN] où il se trouvait
.'un des sous-gouverneurs de la Banque[GN] se trouvait parmi
ı! un des administrateurs de la Banque[GN] d'Angleterre! - pe
ıs la salle des paiements de la Banque[GN]. Le détective! trè
ıres vingt-trois! l'eau pour la barbe[GN] de neuf heures tren

Balises HTML avec sorties en mode « Merge »



Ne pas découper le texte en phrases lors du prétraitement.

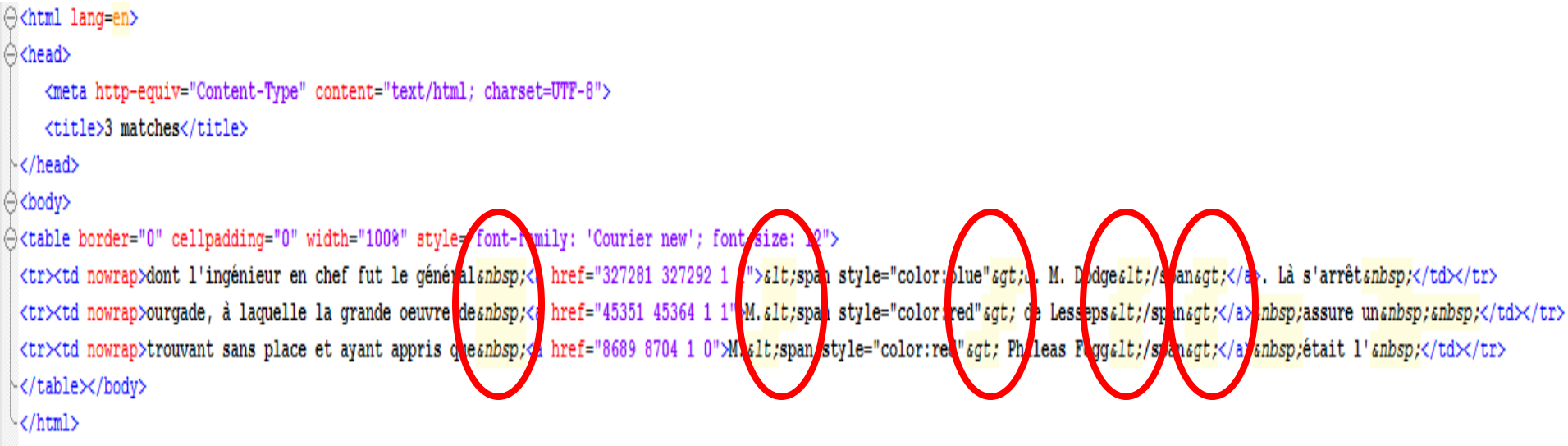
<E>/

<E>/

<E>/

Dans Index, sélectionner « Longest matches »

Code source HTML du fichier « concord.html » généré par UNITEX



Les caractères spéciaux : espace, chevron ouvrant et chevron fermant sont respectivement convertis en   < et >

Résultat de l'exécution du fichier « concord.html » avec un navigateur Web

fut le général J. M. Dodge. Là s'arrêt
rande oeuvre de M. de Lesseps assure un
yant appris que M. Phileas Fogg était l'

1- Copier le contenu de cette page Web « concord.html » générée par le navigateur Web et le mettre dans un autre fichier en .html

2- Exécuter ce nouveau fichier HTML avec votre navigateur Web.

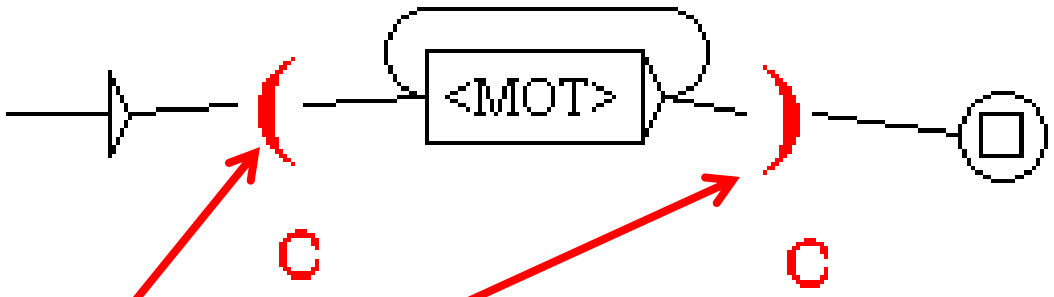
Résultat de l'exécution avec Firefox, Chrome ou IE

dont l'ingénieur en chef fut le général J. M. Dodge. Là s'arrêt ourgade! à laquelle la grande oeuvre de M. de Lesseps assure un trouvant sans place et ayant appris que M. Phileas Fogg était l'

Variables d’entrée

- **Définition : une variable d’entrée permet de sélectionner des parties du texte reconnu par une grammaire (graphe).**
- **Les variables d’entrée sont globales, cela signifie qu’il est possible de définir une variable dans un graphe et l’appeler dans un autre (voir graphe «appel-varent.grf » et « det_entree.grf »).**

Création d'une variable d'entrée



$C(➔ partie ouvrante de la variable d'entrée

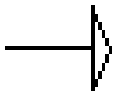
$C) ➔ partie fermante de la variable d'entrée

Création d'une variable d'entrée

$C(

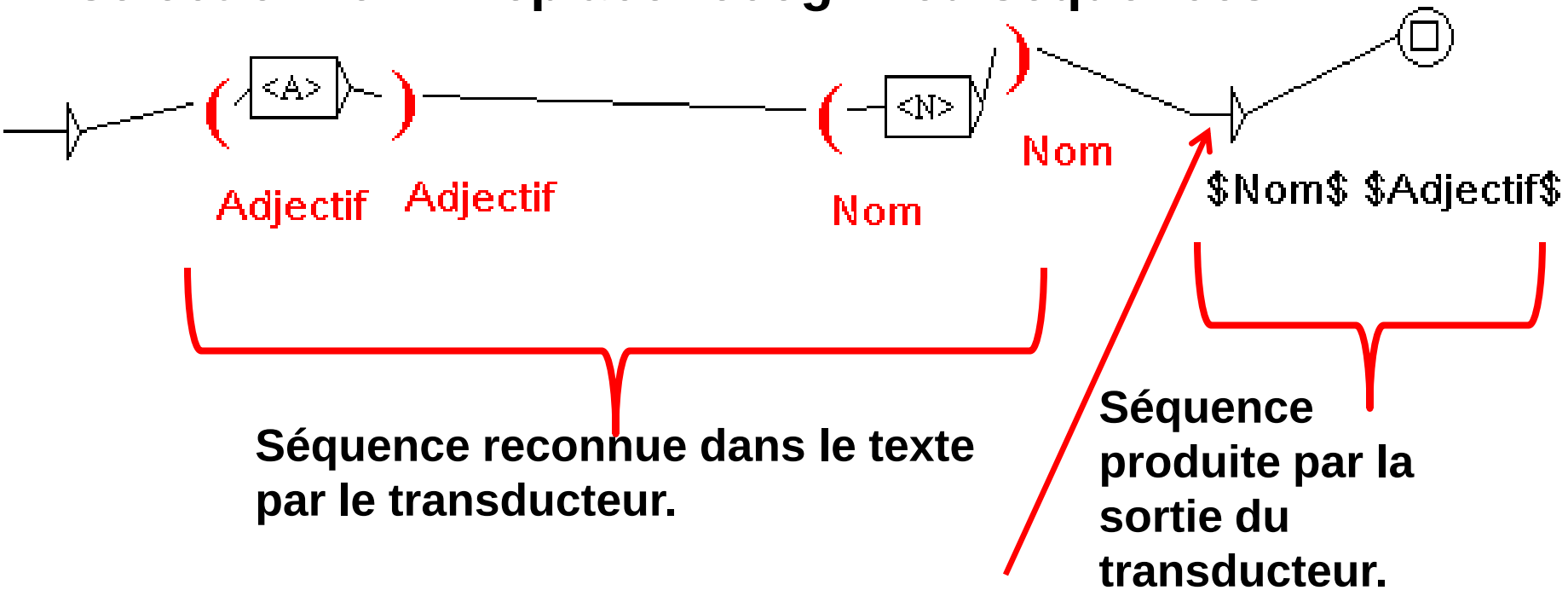
Écrire dans une boîte : $C((partie ouvrante de la variable)

Écrire dans une boîte : $C) (partie fermante de la variable)



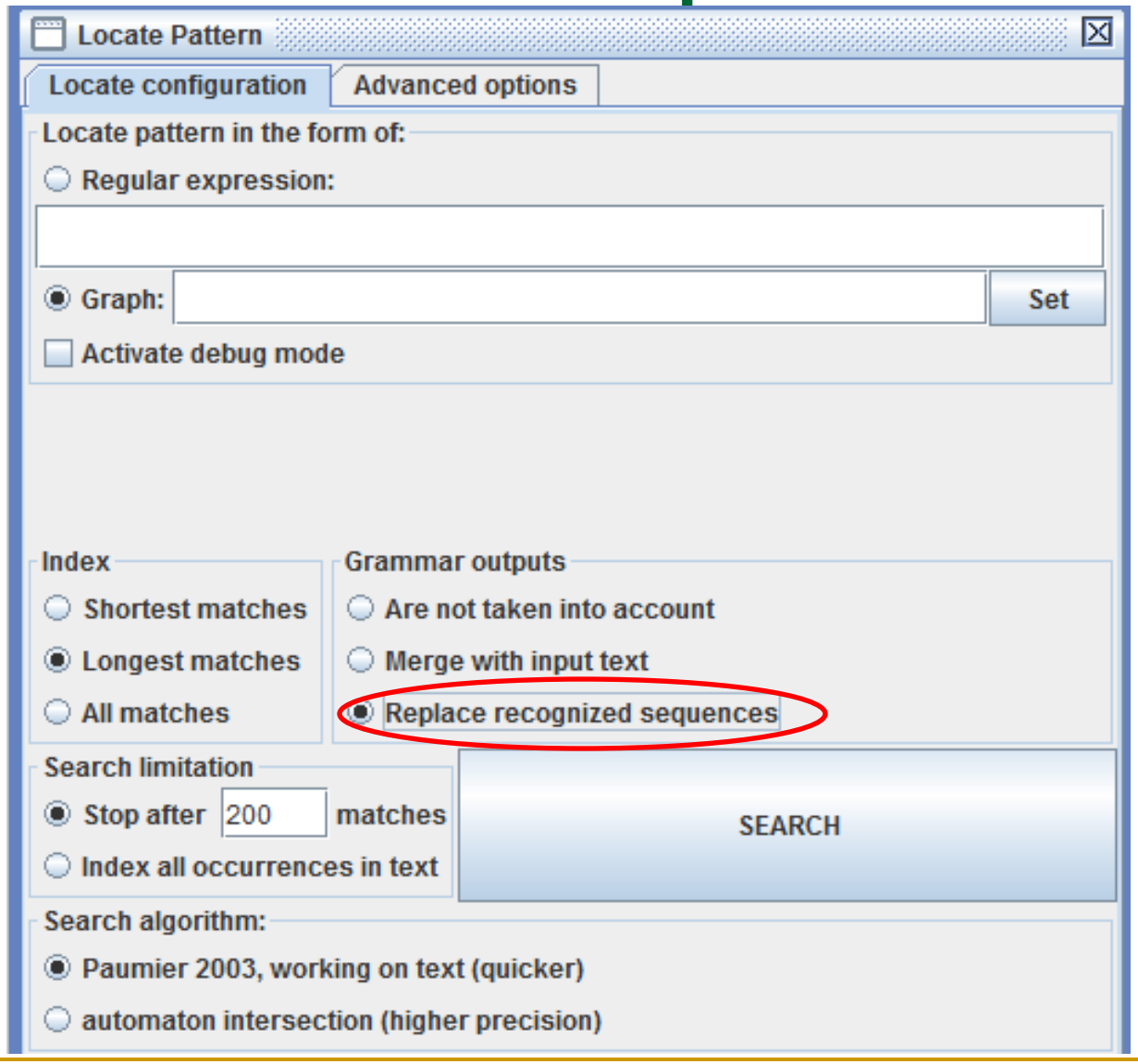
Variables d'entrée avec sortie en mode «Replace»

Pour utiliser le mode « Replace », aller dans Text>Locate Pattern>Grammar outputs, il faut sélectionner: «Replace recognized sequences».

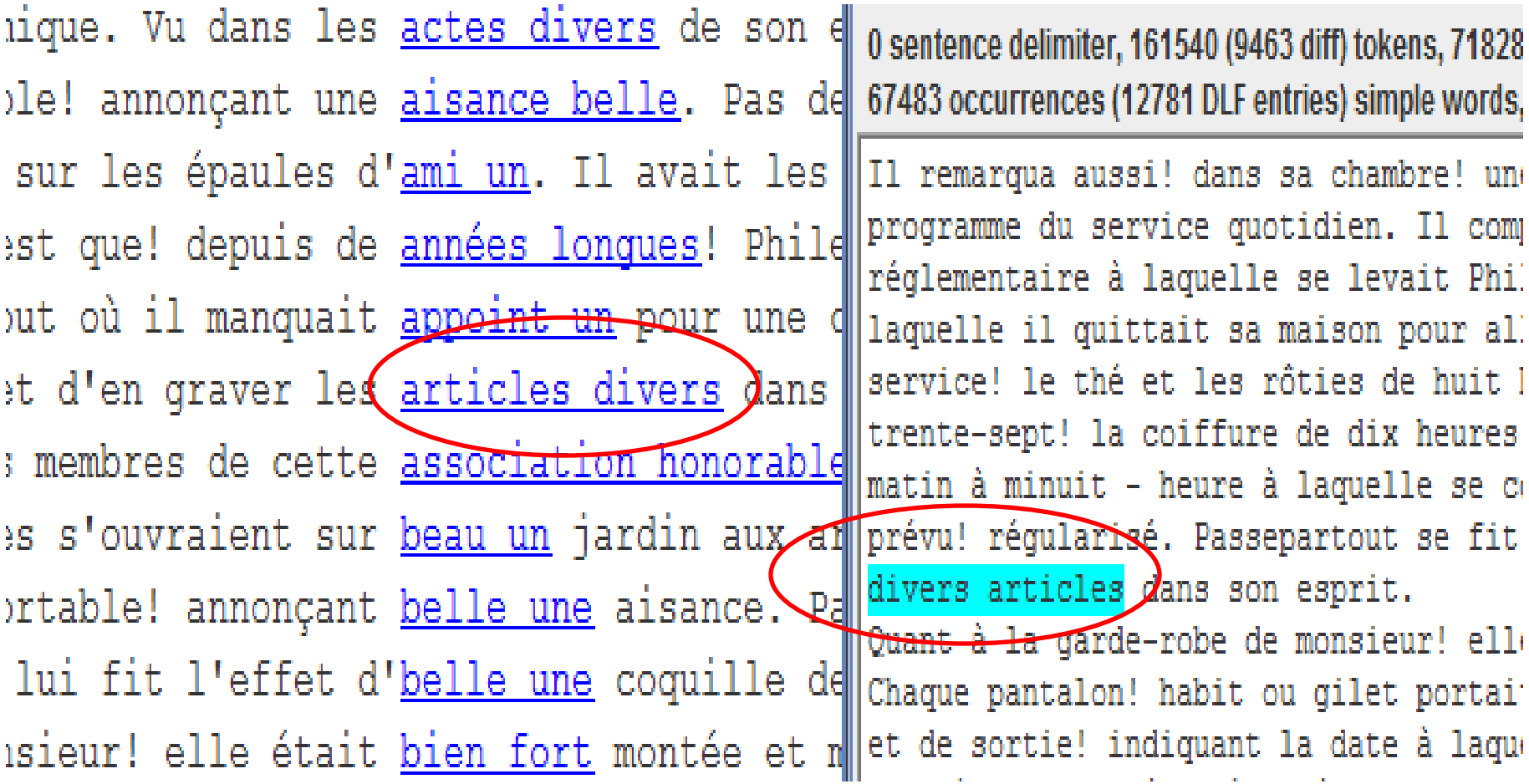


Pour créer cette sortie et appeler les deux variables Adjectif et Nom, il faut écrire dans une boîte : <E>/Nom $Adjectif$

Mode « Replace »

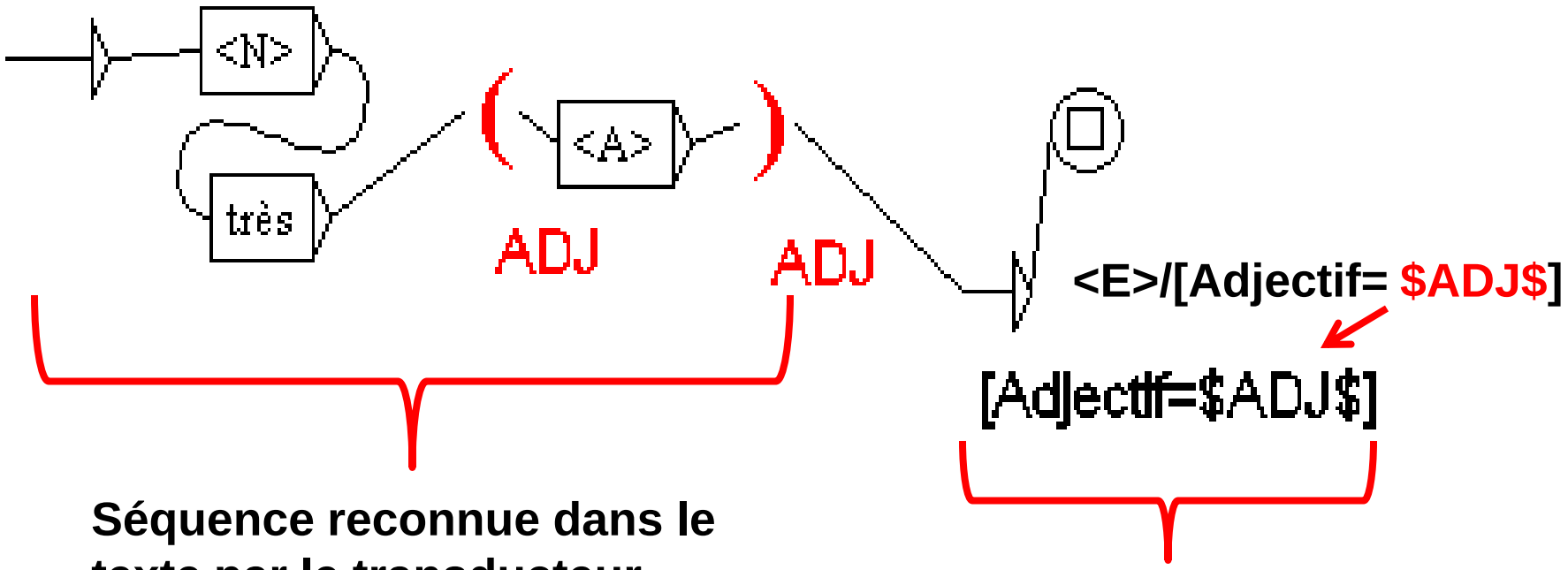


Résultat du remplacement (Replace)



Variable d'entrée avec une sortie en mode « Merge »

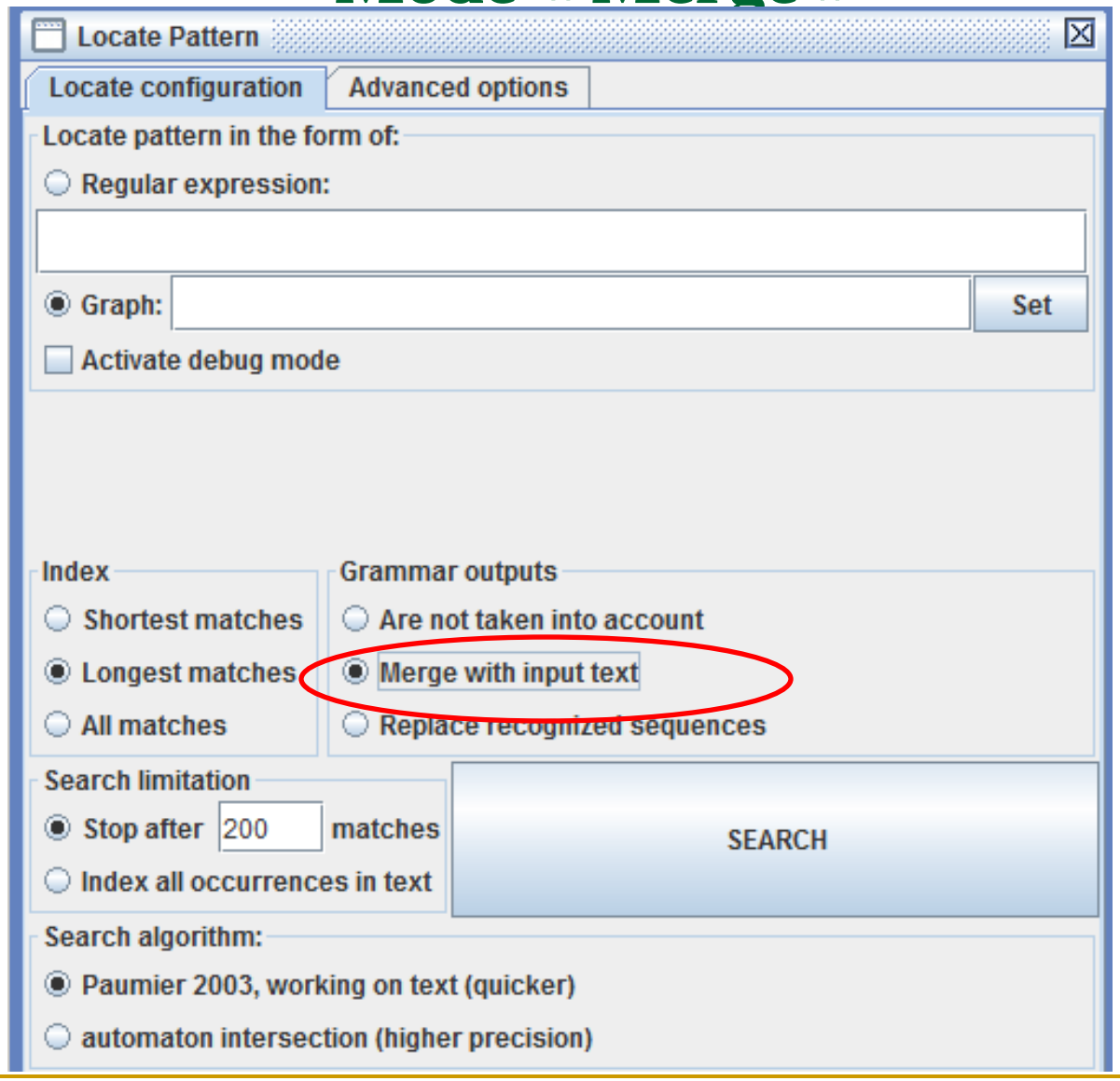
Pour utiliser le mode « Merge », aller dans Text>Locate Pattern>Grammar outputs et sélectionner : « Merge with input text ».



Séquence reconnue dans le texte par le transducteur.

Séquence produite par la sortie du transducteur.

Mode « Merge »


Locate Pattern
Locate configuration
Advanced options
Locate pattern in the form of:
Regular expression:
Graph:
Set
Activate debug mode
Index
Shortest matches
Longest matches
All matches
Grammar outputs
Are not taken into account
Merge with input text
Replace recognized sequences
Search limitation
Stop after 200 matches
Index all occurrences in text
SEARCH
Search algorithm:
Paumier 2003, working on text (quicker)
automaton intersection (higher precision)

Résultat de l'insertion (Merge)

: mauvais, le froid très vif[Adjectif=vif]. Fix, assis sur un banc da
:e ne soit un homme très fort[Adjectif=fort], répondit le consul. Vou
: dressait un mât très élevé[Adjectif=élevé], sur lequel s'enverguait
ıdre dans ces mers très fréquentées[Adjectif=fréquentées] aux approch
ıgea dans les montagnes très ramifiées[Adjectif=ramifiées] des Ghâtes
. éprouva des mouvements de tangage très violents[Adjectif=violents],

Séquence reconnue dans le texte par le transducteur.

Séquence produite par la sortie du transducteur.

Variables de sortie

- **Définition : une variable de sortie permet de mémoriser des sorties produites par une grammaire (graphe).**
- **Les variables de sortie sont globales, cela signifie qu'il est possible de définir une variable dans un graphe et l'appeler dans un autre (voir graphe «appel-varsrt.grf » et « det_sortie.grf ».**

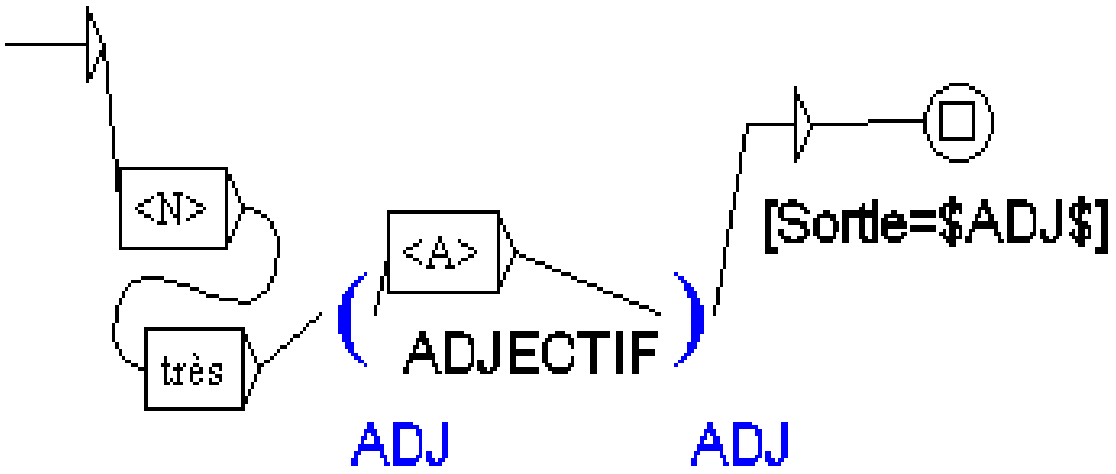
Création d'une variable d'entrée

$|ADJ(

Écrire dans une boîte : $|ADJ((partie ouvrante de la variable)

Écrire dans une boîte : $|ADJ) (partie fermante de la variable)

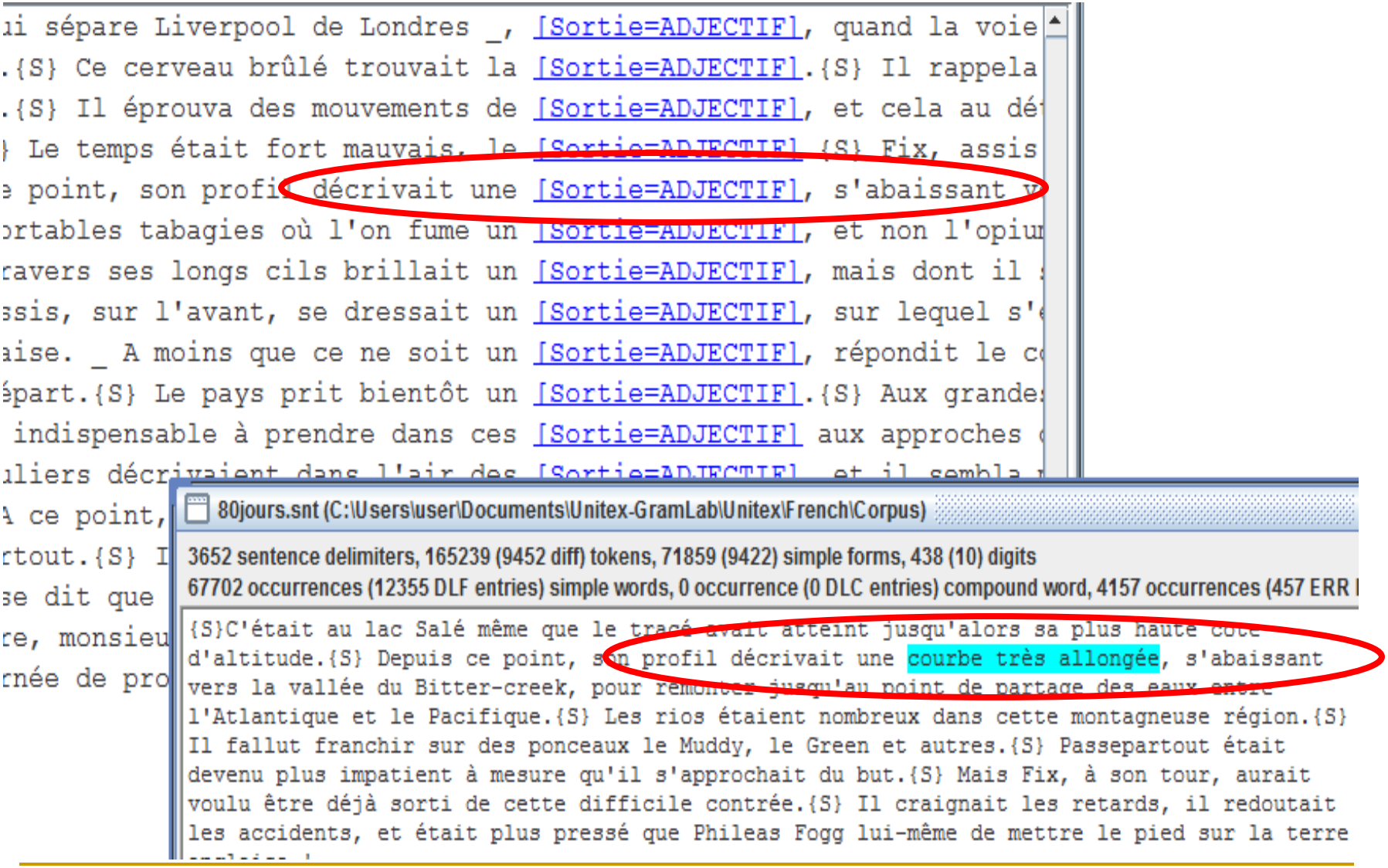
Variable de sortie « ADJ »



$|ADJ(➔ **partie ouvrante de la variable de sortie (**

$|ADJ) ➔ **partie fermante de la variable de sortie)**

Résultat du remplacement (Replace)



Résultat de l'insertion (Merge)

t.{S} Le pays prit bientôt un aspect très sauvage[Sortie=ADJECTIF].{S} Aux grandes fo
épare Liverpool de Londres _, chose très faisable[Sortie=ADJECTIF], quand la voie est
Ce cerveau brûlé trouvait la chose très faisable[Sortie=ADJECTIF].{S} Il rappela mêm
int, son profil décrivait une courbe très allongée[Sortie=ADJECTIF], s'abaissant vers
monsieur Fogg ? dit-elle. _ C'est très simple[Sortie=ADJECTIF], répondit le gentleman
de promenade, il se sentit l'estomac très creux[Sortie=ADJECTIF].{S} Il avait bien r
lit que l'infortuné Fix devait être très désappointé[Sortie=ADJECTIF], très humilié da
temps était fort mauvais, le froid très vif[Sortie=ADJECTIF].{S} Fix, assis sur un b
. _ A moins que ce ne soit un homme très fort[Sortie=ADJECTIF], répondit le consul.{S
, sur l'avant, se dressait un mât très élevé[Sortie=ADJECTIF], sur lequel s'enverguai
lispensable à prendre dans ces mers très fréquentées[Sortie=ADJECTIF] aux approches de
point, il s'engagea dans les montagnes très ramifiées[Sortie=ADJECTIF] des Ghâtes-Oc
rs ses longs
t.{S} Il sava
bles tabagies
Il éprouva d
rs décrivaien

80jours.snt (C:\Users\user\Documents\Unitex-GramLab\Unitex\French\Corpus)

3652 sentence delimiters, 165239 (9452 diff) tokens, 71859 (9422) simple forms, 438 (10) digits
67702 occurrences (12355 DLF entries) simple words, 0 occurrence (0 DLC entries) compound word, 4157 occurrences (457 ERR lin

{S}C'était au lac Salé même que le tracé avait atteint jusqu'alors sa plus haute cote d'altitude.{S} Depuis ce point, son profil décrivait une courbe très allongée, s'abaissant vers la vallée du Bitter-creek, pour remonter jusqu'au point de partage des eaux entre l'Atlantique et le Pacifique.{S} Les rios étaient nombreux dans cette montagneuse région.{S} Il fallut franchir sur des ponceaux le Muddy, le Green et autres.{S} Passepartout était devenu plus impatient à mesure qu'il s'approchait du but.{S} Mais Fix, à son tour, aurait voulu être déjà sorti de cette difficile contrée.{S} Il craignait les retards, il redoutait les accidents, et était plus pressé que Phileas Fogg lui-même de mettre le pied sur la terre

Exercice

Construire un graphe qui extrait les suites adjectif nom et qui les transforme en nom adjectif, mais en ne conservant que les adjectifs qui se terminent par « able » et les noms qui sont au « féminin singulier ». Le résultat contient les 23 occurrences (matches) suivantes: