Plate-formes logicielles pour le TAL 1 : graphes et grammaires locales dans Unitex

Karën Fort

karen.fort@sorbonne-universite.fr / http://karenfort.org

23 novembre 2018

Quelques sources d'inspiration

- Manuel d'Unitex : http://www-igm.univ-mlv.fr/~unitex/index.php?page=4
- M. Constant (Université de Marne-la-Vallée / IGM), qui a patiemment répondu à toutes mes questions

Recherches de motifs complexes

Rechercher dans le TDM

- toutes les occurrences des pronoms personnels *(je, tu, il, . . .)* $\rightarrow$ *je* + *tu* + *il* + *elle* + *nous* + *vous* + *ils* + *elles*
- toutes les occurrences des pronoms personnels qui sont suivis par un verbe $\rightarrow$ *je* + *tu* + *il* + *elle* + *nous* + *vous* + *ils* + *elles* $< V >$
- toutes les suites d'au moins 3 adjectifs (A) $\rightarrow < A > . < A > . < A >$
- toutes les suites de noms. Pourquoi le motif <*N*>* produit-il une erreur ? Que faire pour l'éviter ?$\rightarrow < N > . < N > *$

Recherches utilisant les informations grammaticales|flexionnelles|sémantiques

Rechercher dans le TDM

- tous les adjectifs au féminin pluriel $\rightarrow < A : fp >$
- tous les noms possédant le trait sémantique « humain collectif » $\rightarrow < N + HumColl >$
- tous les verbes à l'imparfait du langage courant $\rightarrow < V + z1 : I >$

Recherches complexes utilisant la concaténation et l'union

Rechercher dans le TDM

- tous les verbes, soit à l'imparfait, soit au présent ou à l'imparfait du subjonctif $\rightarrow < V : I : S : T >$

Recherches utilisant les négations

ou pas

Rechercher dans le TDM

- tous les mots qui ne sont pas dans le dictionnaire $\rightarrow <!DIC>$
- tous les mots qui ne sont pas écrits tout en minuscules $\rightarrow <!MIN>$
- tous les noms non humains $\rightarrow <N \sim Hum \sim HumColl>$

Recherches à l'aide de méta motifs

ou pas

Rechercher

- tout les mots commençant par une majuscule $\rightarrow$ $< PRE >$
- tous les mots qui possèdent le trait sémantique « concret » $\rightarrow$ $< +Conc >$

Recherches utilisant les filtres

Rechercher dans le TDM

- tous les mots qui commencent par *anti* ou *pro*, suivis par un tiret facultatif → <<^(*anti*|*pro*)−?>>
- tous les mots composés contenant un tiret → <CDIC><<->>
- tous les mots qui ne sont pas dans le dictionnaire et qui se terminent par *es* → <!DIC><<es$>>

Expressions régulières avancées

Rechercher dans le TDM

- tous les adjectifs qui ne sont pas très spécialisés
- tous les verbes un peu ou très spécialisés, soit au participe passé, soit à l'infinitif
- toutes les séquences
  - commençant par le verbe avoir (et)
  - se terminant par un verbe au participe passé (et)
  - dans lesquelles peuvent s'insérer des séquences quelconques de mots entre virgules (*eût, au contraire, perdu*)
- tous les verbes au subjonctif passé ou présent, contenant *uiss*

Expressions régulières avancées

Rechercher dans le TDM

- tous les adjectifs qui ne sont pas très spécialisés $< A \sim z3 >$
- tous les verbes un peu ou très spécialisés, soit au participe passé, soit à l'infinitif
- toutes les séquences
  - commençant par le verbe avoir (et)
  - se terminant par un verbe au participe passé (et)
  - dans lesquelles peuvent s'insérer des séquences quelconques de mots entre virgules (*eût, au contraire, perdu*)
- tous les verbes au subjonctif passé ou présent, contenant *uiss*

Expressions régulières avancées

Rechercher dans le TDM

- tous les adjectifs qui ne sont pas très spécialisés
- tous les verbes un peu ou très spécialisés, soit au participe passé, soit à l'infinitif $< V + z3 : W : K > + < V + z2 : W : K >$
- toutes les séquences
  - commençant par le verbe avoir (et)
  - se terminant par un verbe au participe passé (et)
  - dans lesquelles peuvent s'insérer des séquences quelconques de mots entre virgules (*eût, au contraire, perdu*)
- tous les verbes au subjonctif passé ou présent, contenant *uiss*

Expressions régulières avancées

Rechercher dans le TDM

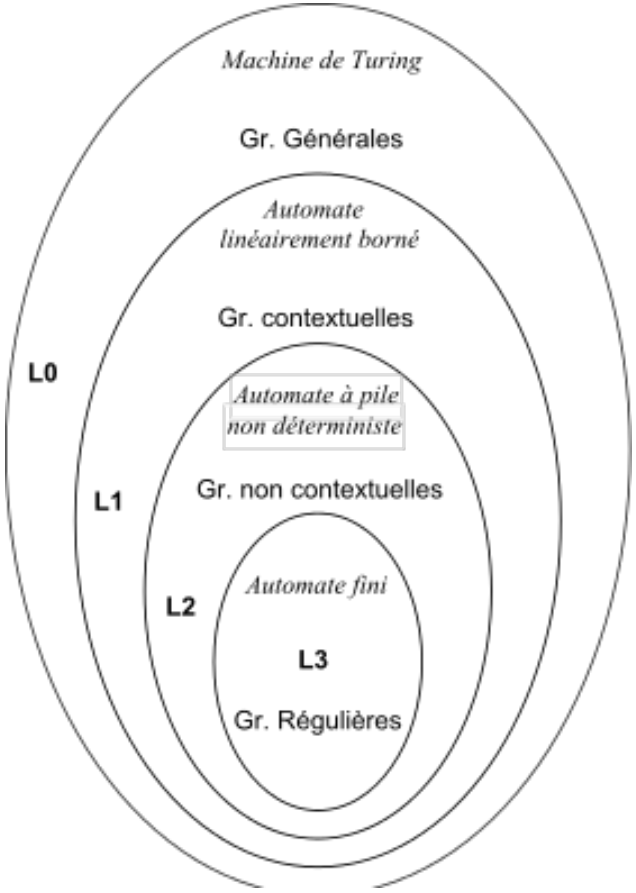
- tous les adjectifs qui ne sont pas très spécialisés
- tous les verbes un peu ou très spécialisés, soit au participe passé, soit à l'infinitif
- toutes les séquences
  - commençant par le verbe avoir (et)
  - se terminant par un verbe au participe passé (et)
  - dans lesquelles peuvent s'insérer des séquences quelconques de mots entre virgules (*eût, au contraire, perdu*)
    $< avoir >, < MOT > *, < V : K >$
- tous les verbes au subjonctif passé ou présent, contenant *uiss*

Expressions régulières avancées

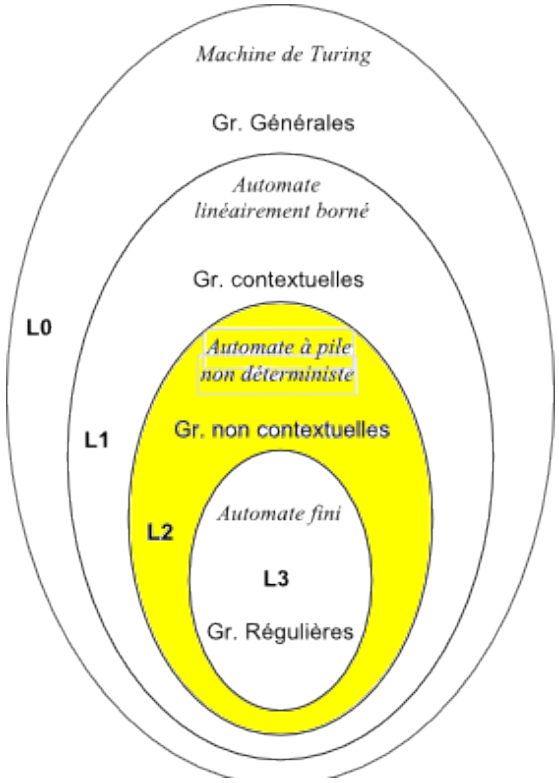
Rechercher dans le TDM

- tous les adjectifs qui ne sont pas très spécialisés
- tous les verbes un peu ou très spécialisés, soit au participe passé, soit à l'infinitif
- toutes les séquences
  - commençant par le verbe avoir (et)
  - se terminant par un verbe au participe passé (et)
  - dans lesquelles peuvent s'insérer des séquences quelconques de mots entre virgules (*eût, au contraire, perdu*)
- tous les verbes au subjonctif passé ou présent, contenant *uiss*
  $< V : S : T >< < uiss > >$ (p. 83 du Manuel d'Unitex)

Hiérarchie des grammaires formelles



Grammaires hors contexte ou algébriques



Définition

Ce sont des grammaires contextuelles où le contexte est vide, ce qui signifie que les symboles non terminaux sont traités indépendamment de la place où ils apparaissent.
[Wikipédia, Grammaires formelles, consultée le 21/09/2014]

Grammaires manipulées par Unitex

Grammaires algébriques étendues

Définition

Les grammaires algébriques étendues sont des grammaires algébriques où les membres droits des règles ne sont plus des suites de symboles mais des expressions rationnelles. [Manuel d'Unitex, p. 94]

$S \rightarrow aS$ devient $S \rightarrow a*$
$S \rightarrow \varepsilon$

Les grammaires (ou graphes) d'Unitex intègrent également la notion de transduction (elles peuvent produire des sorties)

Créer un graphe simple

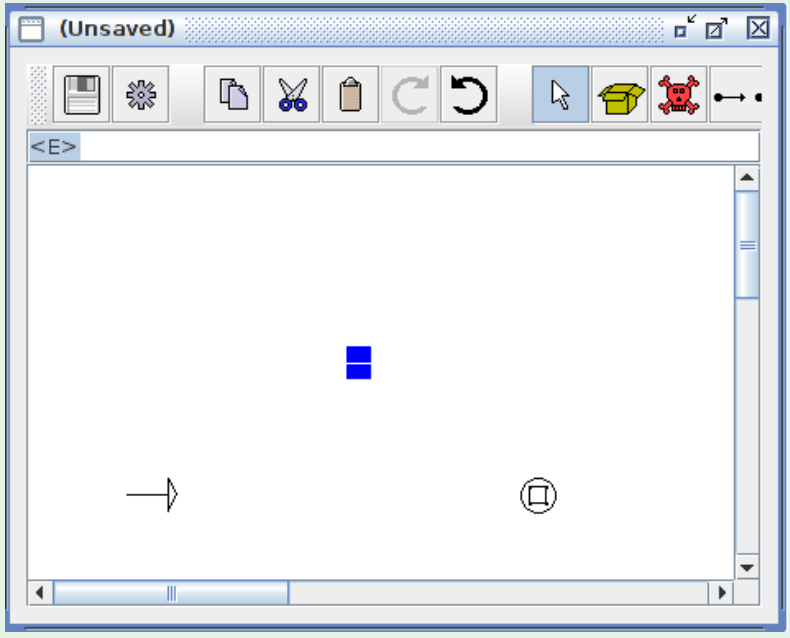
Premier graphe

- lancer `Unitex`
- dans le menu `FSGraphe`, sélectionner `New`
- faire un CTRL+clic quelque part entre l'état initial et l'état final

Créer un graphe simple

Premier graphe

- lancer Unitex
- dans le menu FSGraphe, sélectionner New
- faire un CTRL+clic quelque part entre l'état initial et l'état final

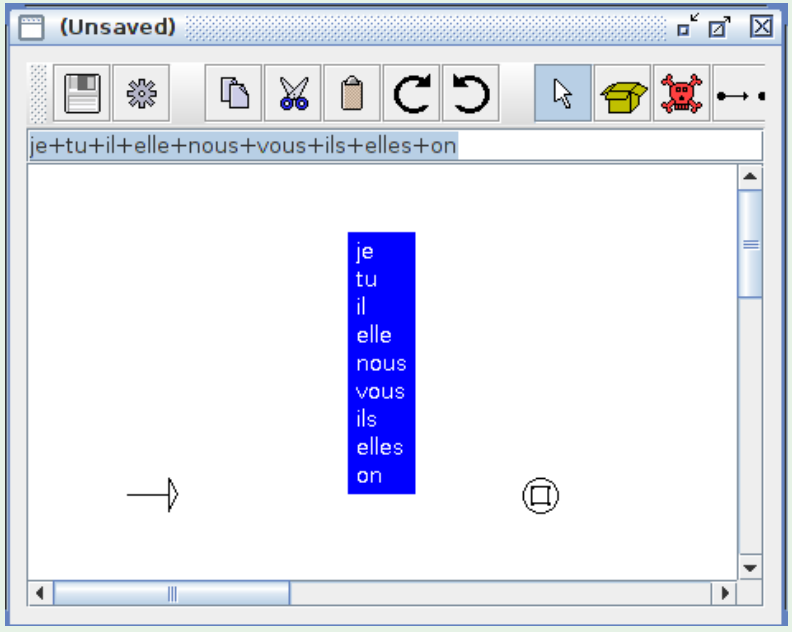


pour supprimer un état, cliquer sur la tête de mort

Créer un graphe simple

Premier graphe

- dans la barre de texte, à la place de <*E*>, taper : *je+tu+il+elle+on+nous+vous+ils+elles*
- taper *Entrer*

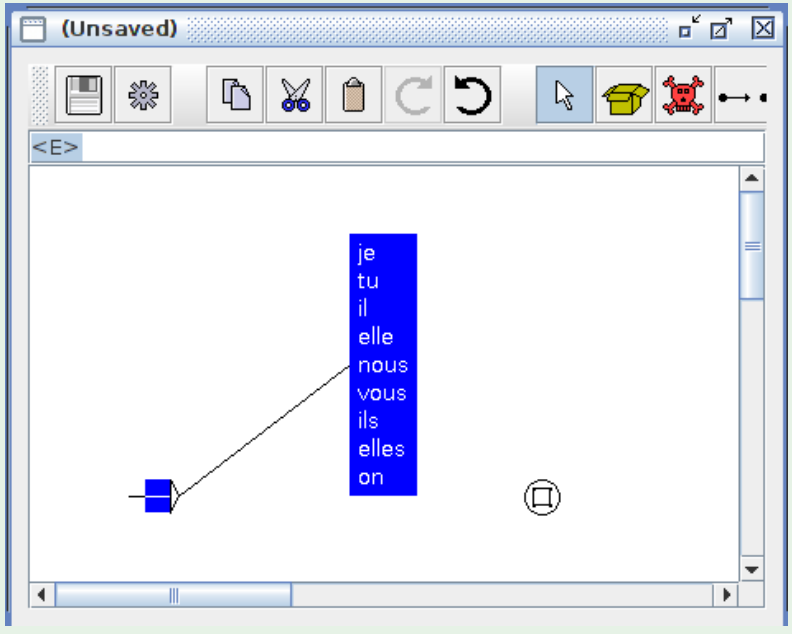


Créer un graphe simple

Premier graphe

- cliquer sur la flèche de l'état initial, puis sur l'état *je+tu...* (une transition apparaît)

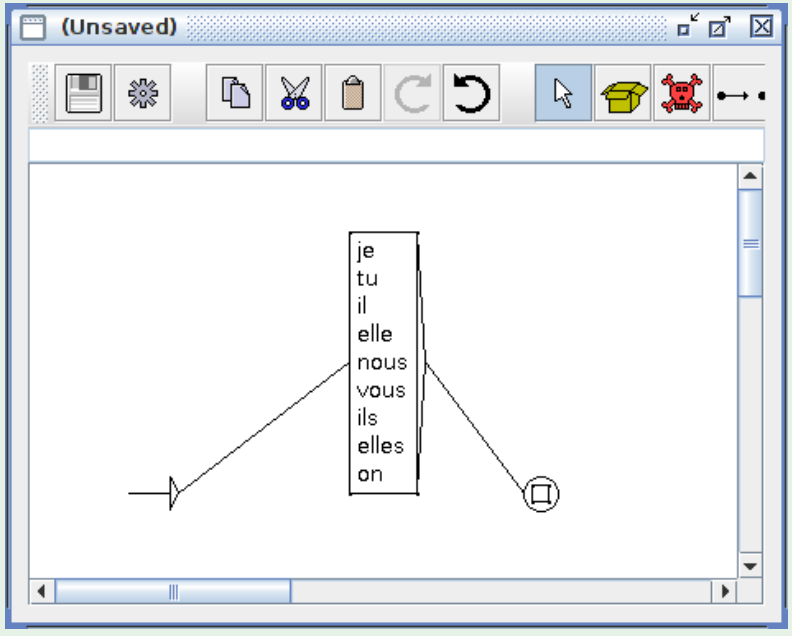
pour supprimer une transition, refaire la même manipulation



Créer un graphe simple

Premier graphe

- cliquer sur l'état *je+tu...* puis sur l'état final (une transition apparaît)



- dans le menu `FSGraphe`, sélectionner `Save as...` (pour enregistrer le graphe)

Que fait ce graphe ?

Appliquer un graphe à un texte

Application

Pour appliquer le graphe au texte :

- ouvrir un texte (par exemple le *Tour du monde en 80 jours*)
- puis, menu `Text` / `Locate Pattern`
- dans `Graph`, indiquer (*set*) le chemin vers le graphe enregistré précédemment

Exercice

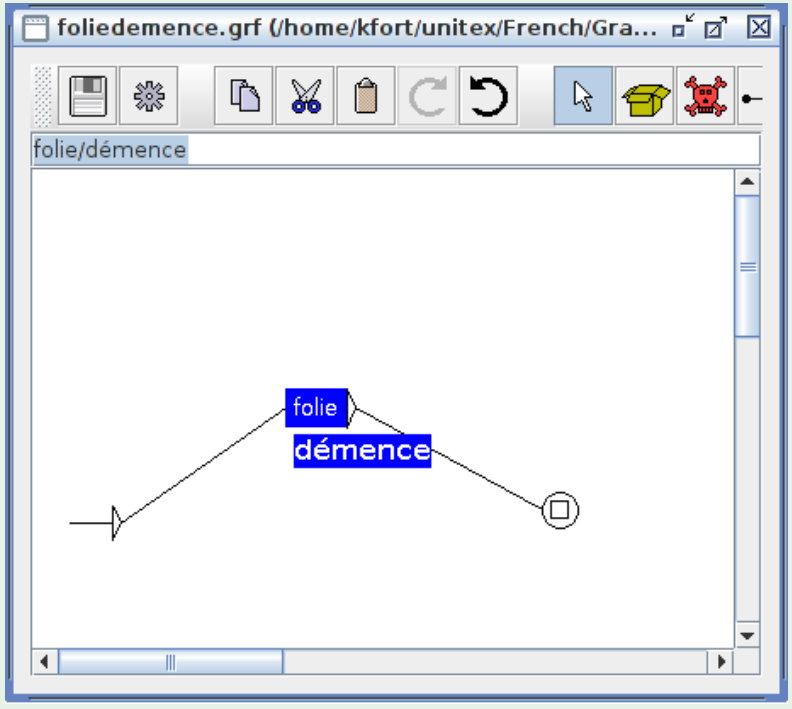
Manipulation de graphe

- compléter ce graphe pour qu'il reconnaisse un pronom personnel suivi d'un verbe (n'importe lequel, sous n'importe quelle forme)
- l'appliquer sur le texte

Création d'une substitution

Remplacer une chaîne par une autre

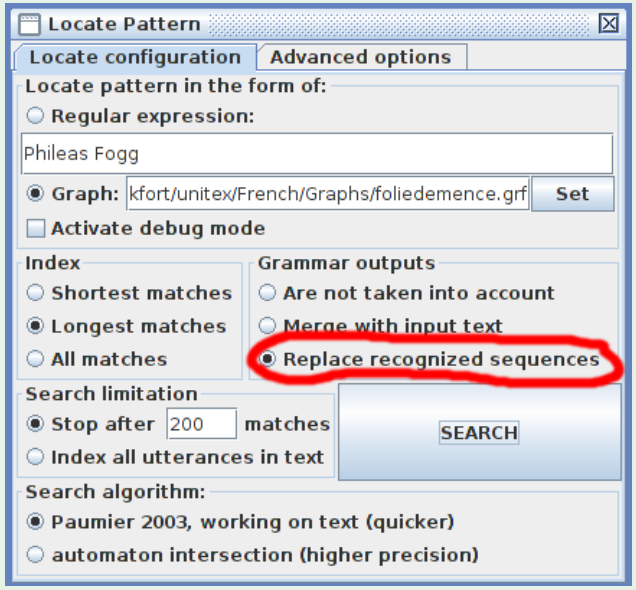
- créer un graphe à un état intermédiaire
- dans cet état, écrire : *folie/démence*
- enregistrer le graphe



Application d'un graphe avec substitution

Remplacer une chaîne par une autre

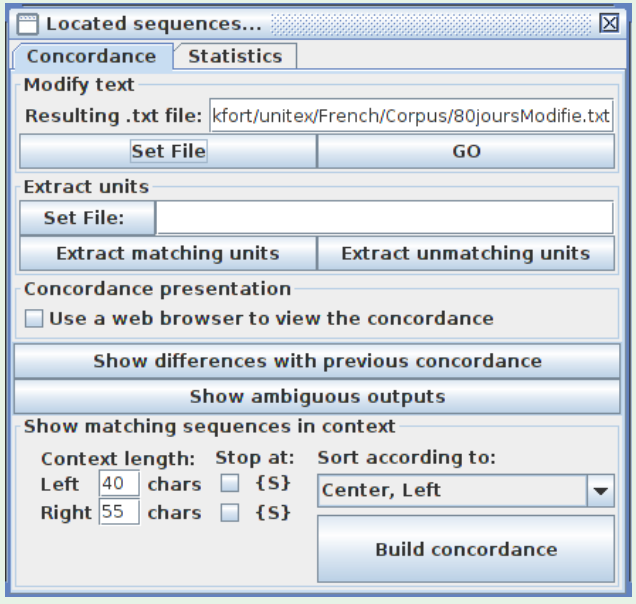
- ouvrir un texte (par exemple le *Tour du monde en 80 jours*)
- puis, menu `Text` / `Locate Pattern`
- dans `Graph`, indiquer (*set*) le chemin vers le graphe enregistré précédemment
- sélectionner `Grammar Outputs` / `Replace recognized sentences`



Appliquer les remplacements dans le texte

Écriture dans le texte

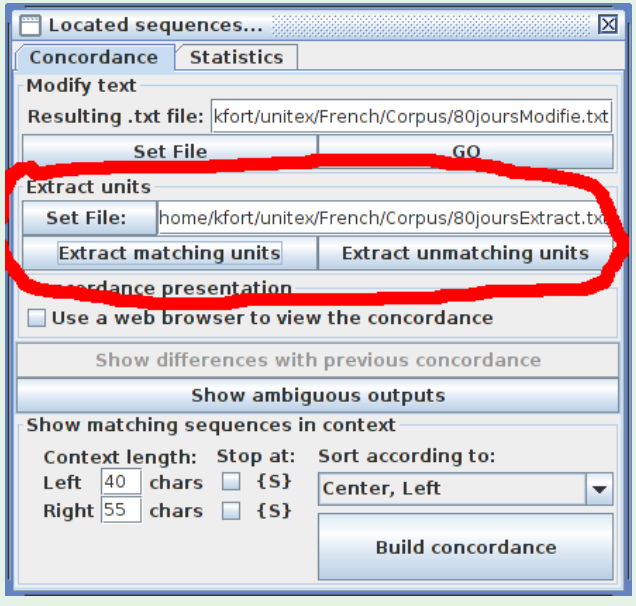
- dans `Located sequences...`, `Concordance`, `Modify text`
- spécifier (`set file`) le fichier résultant (`Resulting txt file`)
- `Go` !



Extraire les unités remplacées

Création d'un fichier des phrases contenant les unités remplacées (ou non remplacées)

- dans `Located sequences...`, `Concordance`, `Extract units`
- spécifier (`Set file`) le fichier résultant (`Resulting txt file`)
- `Extract matching units` !



Exercice

Transformer le texte

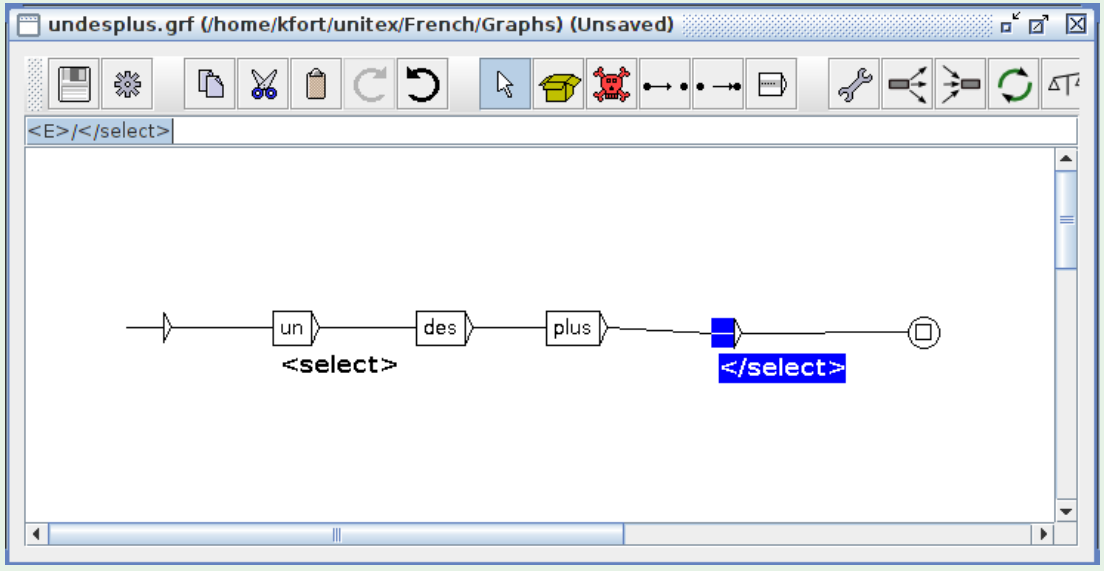
Créer un graphe qui réalise un début d'anonymisation du *Tour du monde en 80 jours* :

- remplacer toutes les occurrences de *Phileas Fogg* par *Pers1*
- remplacer toutes les occurrences de *Passepartout* par *Pers2*
- l'appliquer sur le texte et vérifier les résultats obtenus

Création d'une annotation

Annoter une chaîne de caractères

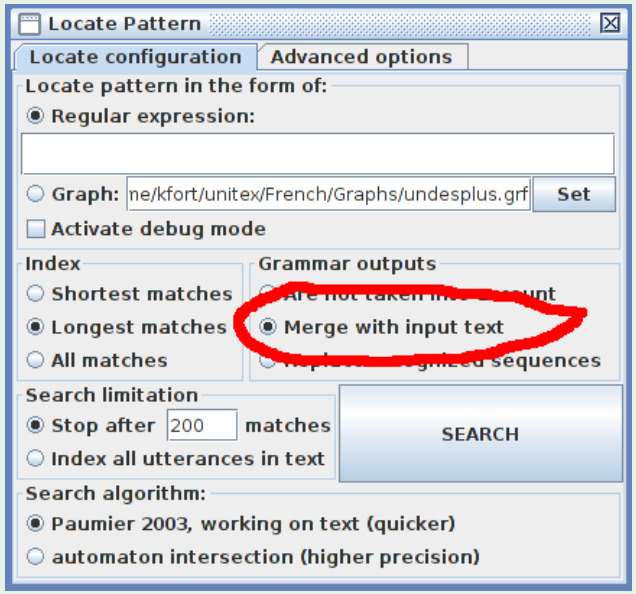
- créer un graphe reconnaissant l'expression *un des plus*
- écrire : *un*/< *select* >
- ajouter après l'expression l'état contenant : < *E* >/< /*select* >
- enregistrer le graphe



Appliquer les annotations dans le texte

Annoter le texte

- ouvrir un texte (par exemple le *Tour du monde en 80 jours*)
- puis, menu `Text` / `Locate Pattern`
- sélectionner `Grammar Outputs / Merge with input text`



Exercice

Annoter le texte

Créer un graphe qui réalise un début d'annotation en entités nommées du *Tour du monde en 80 jours* :

- annoter toutes les occurrences de *Phileas Fogg* et de *Passepartout* par des balises $<$ *pers* $><$ /*pers* $>$
- appliquer le graphe sur le texte et vérifier les résultats obtenus

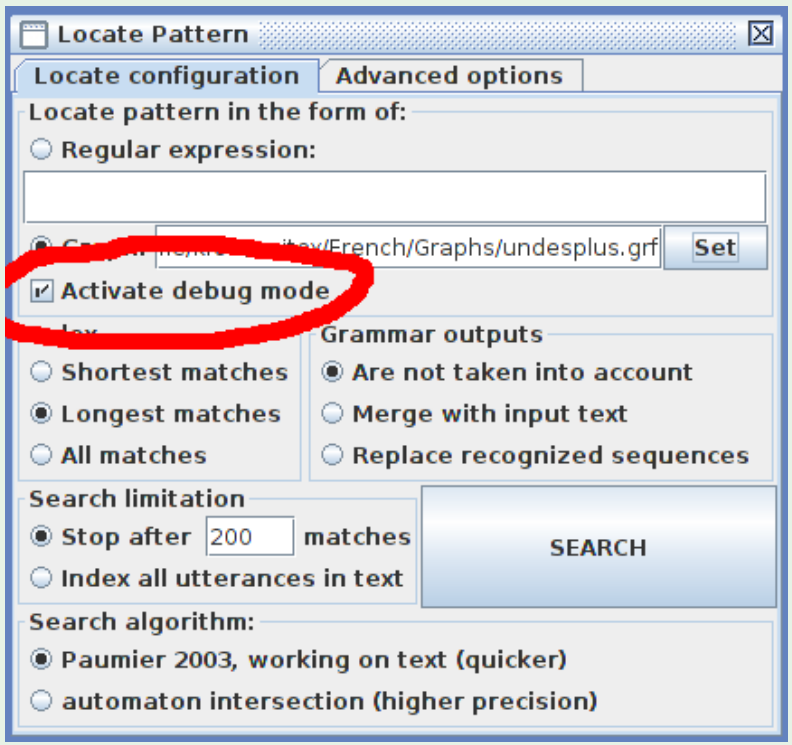
Transducteurs : `replace` *vs* `merge`

- `replace` : les sorties remplacent les séquences identifiées
  folie/démence $\rightarrow$ *démence* remplace *folie*
- `merge` : les sorties sont insérées à gauche des séquences reconnues
  un/$<$ *select* $>$ $\rightarrow$ $<$ *select* $>$ *un*

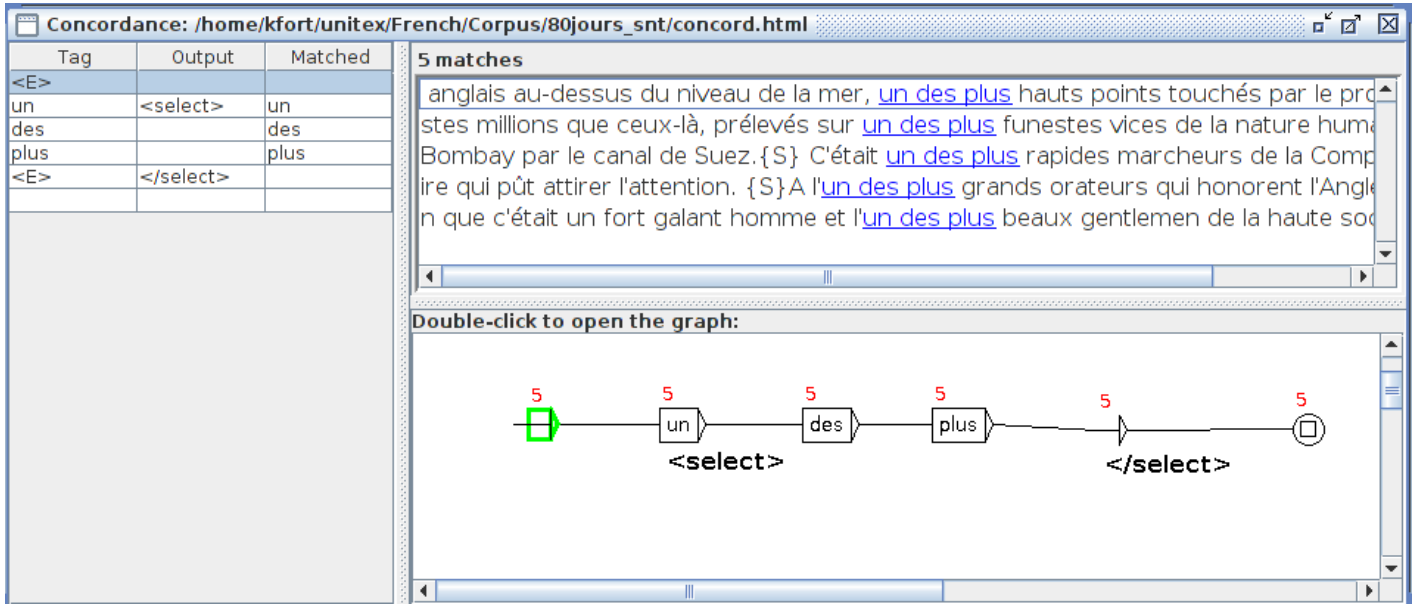
Déboguer des graphes

Mode *debug*

- ouvrir un texte (par exemple le *Tour du monde en 80 jours*)
- puis, menu `Text` / `Locate Pattern`
- sélectionner le graphe *un des plus*
- cocher la case `Activate debug mode`



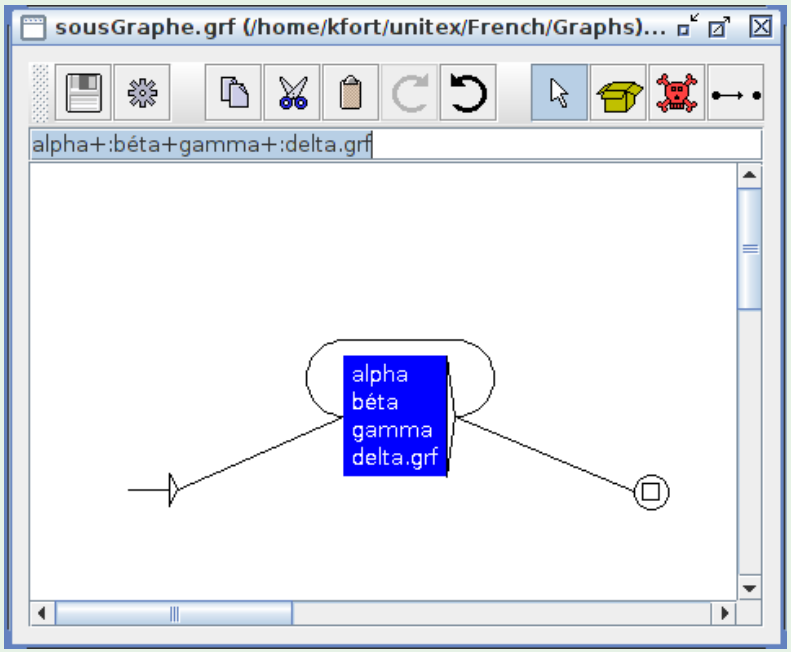
Mode *debug* : affichage



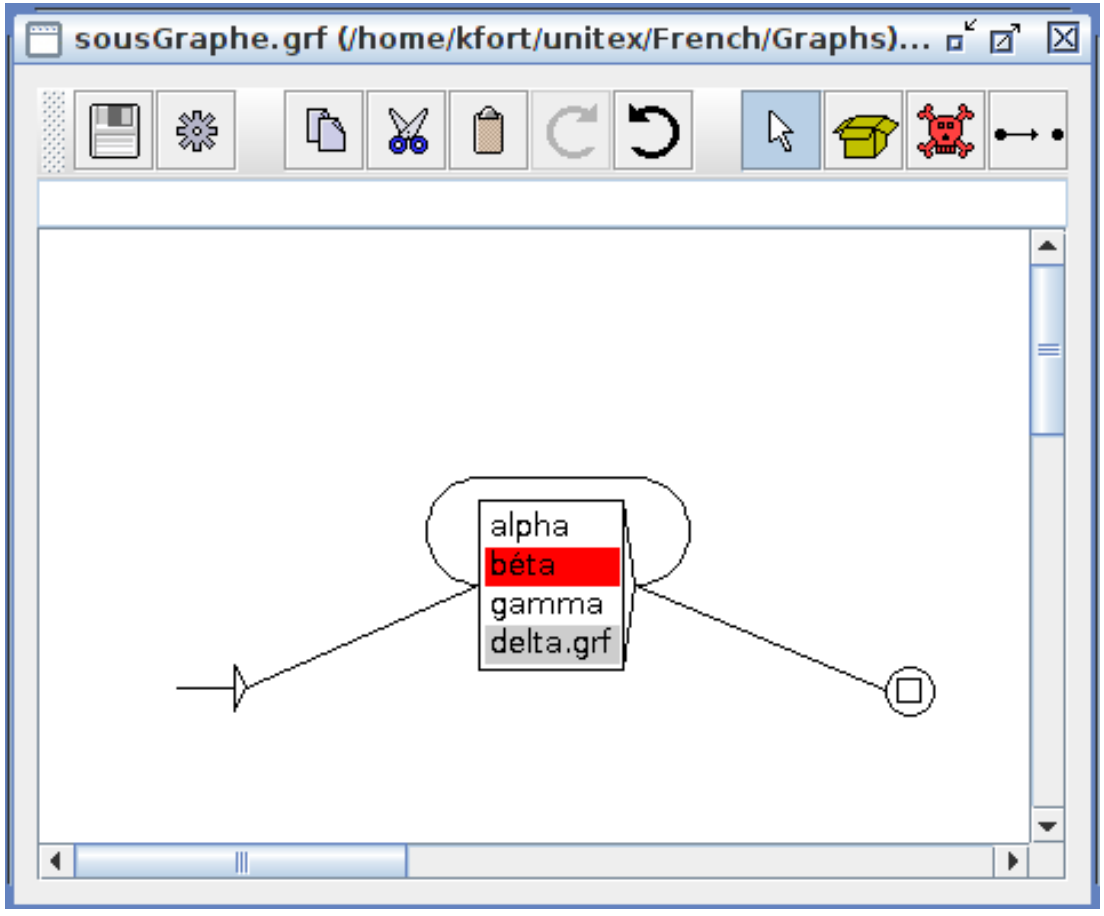
Faire appel à un sous-graphe

Sous-graphe

- créer un graphe vide
- ajouter un état contenant *alpha+beta+gamma+delta*
- modifier l'état en appelant des sous-graphes
  alpha+ :beta+gamma+ :CheminVersFichier delta.grf



Affichage des sous-graphes



Exercice

Utilisation de sous-graphes

Pour créer un graphe reconnaissant les dates :

- créer un sous-graphe reconnaissant les jours de la semaine
- créer un sous-graphe reconnaissant les mois
- intégrer les sous-graphes dans un graphe reconnaissant les expressions de type *jour de la semaine numéro du jour mois année* :
  lundi 22 septembre 1997

Savoir :

- créer un graphe
- insérer des sous-graphes
- transformer et annoter un texte
- utiliser des variables

Comprendre :

- le type de grammaire utilisé
- le fonctionnement des transducteurs

Recherches avancées

Exercices 1

1. Modifiez la grammaire des dates pour extraire des résultats intéressants sur le corpus du *Tour du Monde en 80 jours*
2. Construisez une grammaire
   - reconnaissant des groupes nominaux simples, en tenant compte des accords en genre et en nombre
   - insérez des sorties dans la grammaire afin qu'à partir du texte « après tout, son énorme gaffe n'est pas sérieuse. », on puisse obtenir la concordance suivante :
     son énorme gaffe,.GN