CC BY NC SA

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

Karën Fort

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

30 novembre 2018

LETTRES SORBONNE UNIVERSITÉ

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

Grammaire des dates

Modifiez la grammaire des dates pour extraire des résultats intéressants sur le corpus du *Tour du Monde en 80 jours*

Groupes nominaux

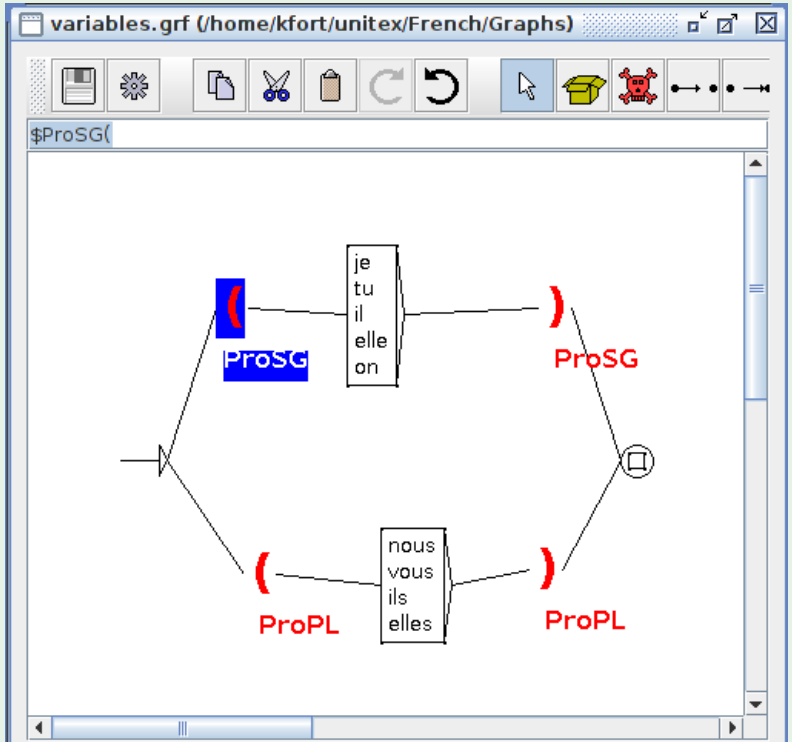
Construisez une grammaire :

- 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

Créer une variable (d'entrée)

Création d'une variable

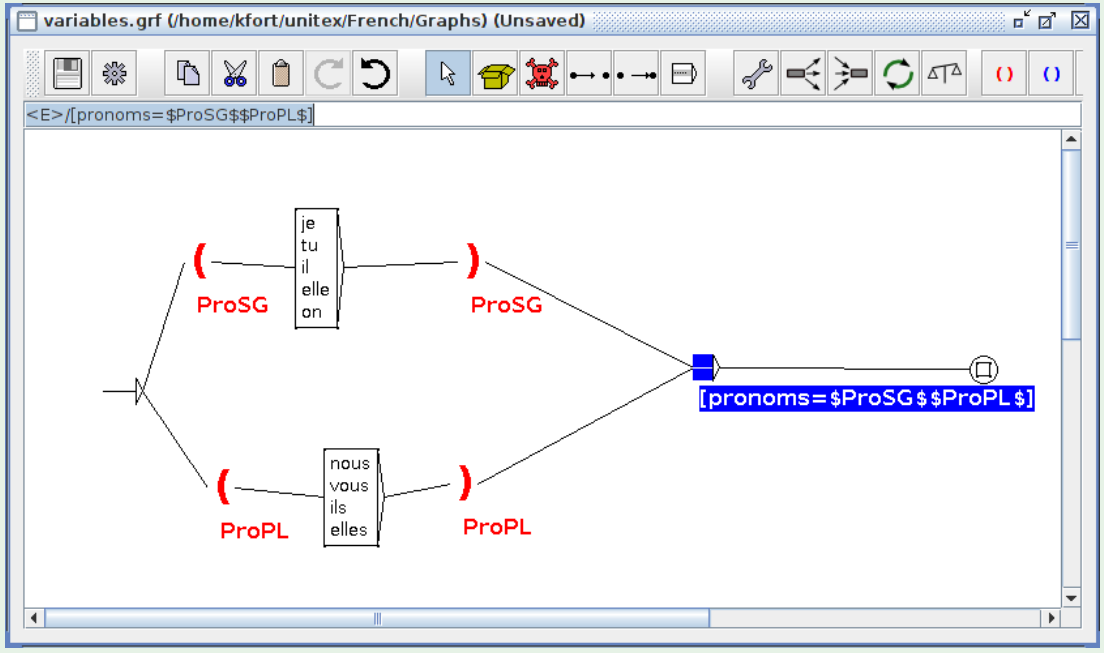
- créer un graphe qui reconnaît :
  - les pronoms personnels singuliers (je, tu, il, elle, on) ou
  - les pronoms personnels pluriels (nous, vous, ils, elles)
- insérer des états englobant définissant la variable nomDeVariable : $nomDeVariable(et $nomDeVariable)



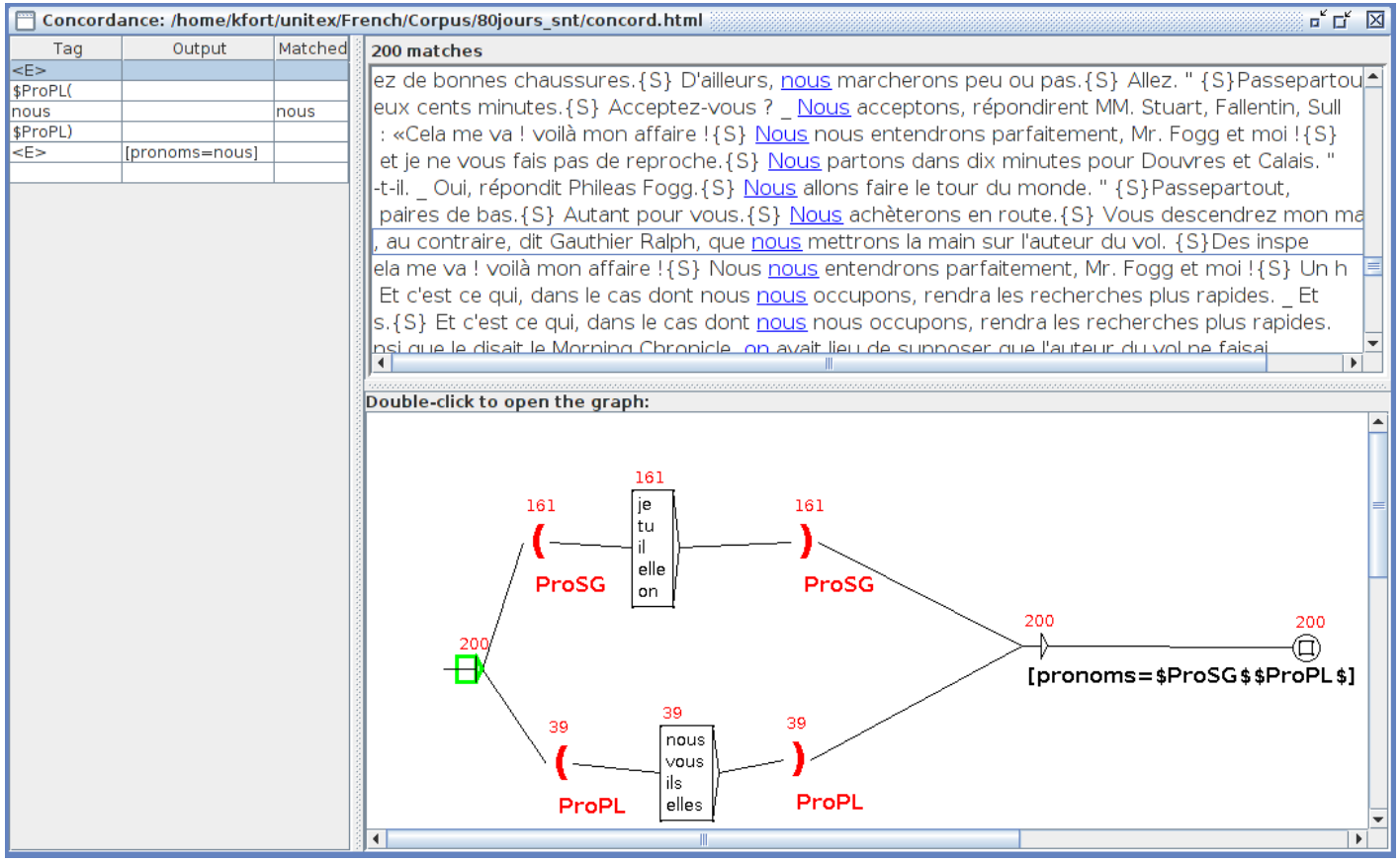
Utiliser les variables

Fusionner des résultats

- ouvrir le graphe précédemment créé
- ajouter, juste avant l'état final, un état contenant
  $< E >$/[nom de l'élément qu'on décrit=$nomDeVariable$]



Variables en mode *debug*



les variables sont globales

Exercice : une autre utilisation des variables

Inverser des motifs

- créer un graphe qui reconnaît les mois/années
- ajouter les variables $mois et $annee
- inverser les mois et les années

Annotation *vs* variable

Annotation :

- ajoutée à la bande de sortie du FST
- ne peut pas être testée ou comparée

Variable :

- permet de stocker une chaîne de la bande d'entrée du FST
- peut être testée :
  - en insérant $xxx.SET$ à la sortie d'une boîte
  - si une variable dénommée xxx a été définie, cette séquence est ignorée et la reconnaissance continue, sinon, elle s'arrête et le programme repart en arrière
- peut être comparée :
  - en insérant $abc.EQUAL=xyz$ à la sortie d'une boîte
  - agit comme un interrupteur qui permet de bloquer l'exploration de grammaire si la valeur de la variable *abc* est différente de la valeur de la variable *xyz*

Test

cf. Manuel d'Unitex p. 148

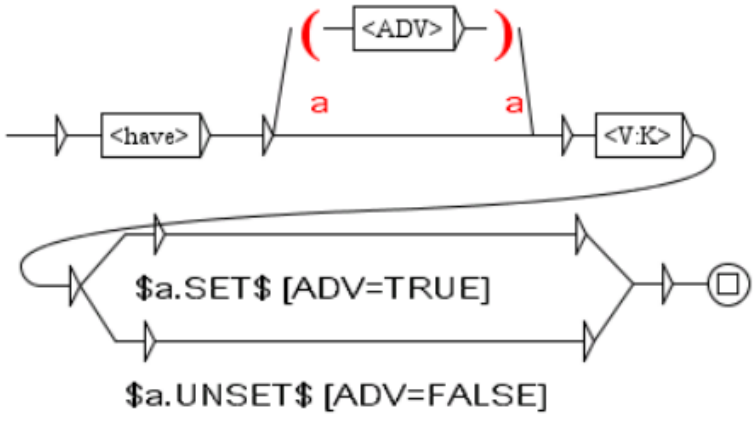
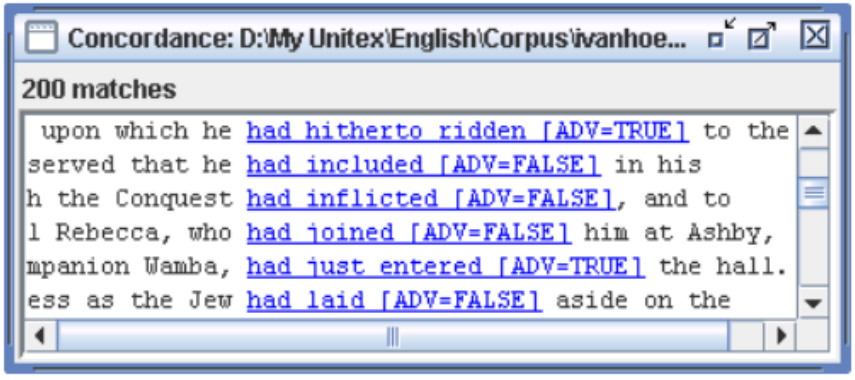


FIGURE 6.50 – Test d'une variable



Construction des dictionnaires

1. construction d'un dictionnaire de formes canoniques (ou formes de base)
2. construction de modules de flexion automatique (transducteurs)
3. à chaque forme de base, on associe une classe flexionnelle (un ensemble de règles)

Étape 1 : créer le fichier DELAS (formes non fléchies)

Menu File Edition / New File

Ajouter (1 par ligne) des mots (unités lexicales simples) qui sont toujours au masculin :

- ballon
- livre
- (votre exemple)

Quelle flexion ? On va la créer : N1000

Ce qui donne :
ballon,N1000
livre,N1000
RETOUR À LA LIGNE

Enregistrer le fichier **sous Dela** avec une extension .dic

Étape 2 : créer le graphe de flexion

Menu FS Graph / New

Créer un graphe permettant :

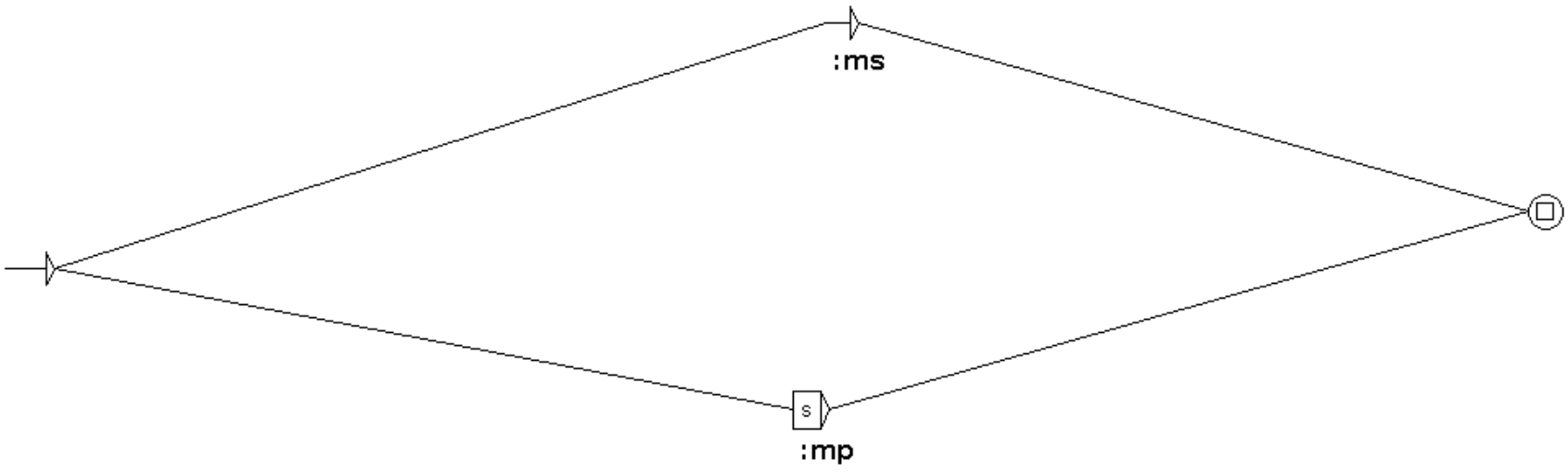
- d'ajouter un s au (masculin) pluriel : s/ :mp
- de ne rien ajouter au (masculin) singulier : <E>/ :ms

L'enregistrer **sous Inflection** avec le nom N1000 (.grf). Le compiler (Unitex va créer un .fst2).

! pas d'espace

Étape 2 : résultat

ballon

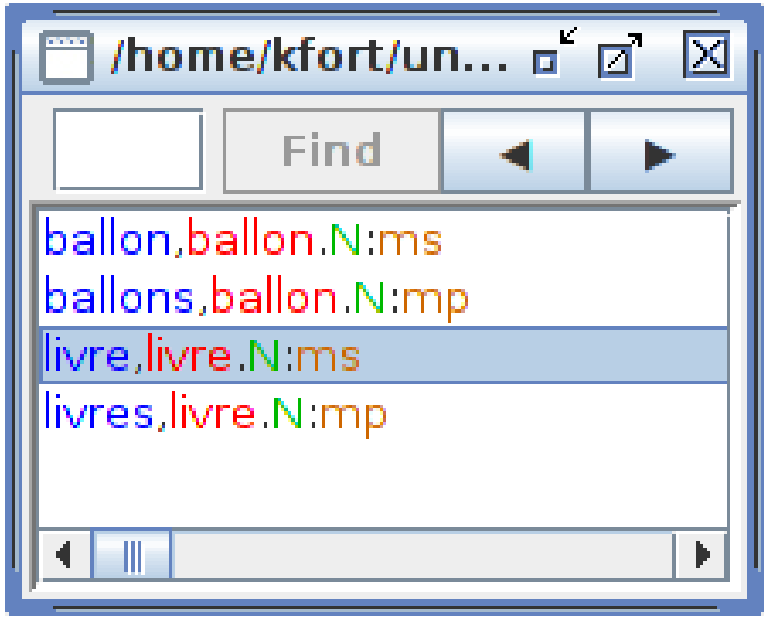


Étape 3 : créer le dictionnaire fléchi

Menu DELA / Open

- sélectionner le fichier dictionnaire précédemment créé
- DELA / Inflect...
- Inflect Dictionary

Étape 3 : résultat



Mode console : UnitexTool

- permet d'exécuter les programmes externes d'Unitex
- permet d'enchaîner les commandes

Exemple : faire un `locate` et lancer la concordance

```
 UnitexTool {
 Locate "-tD:\My Unitex\English\Corpus\ivanhoe.snt"
"D:\My Unitex\English\regexp.fst2"
"-aD:\My Unitex\English\Alphabet.txt" -L -I -n200
"--morpho=D:\Unitex2.0\English\Dela\dela-en-public.bin" -b -Y
}
{
Concord "D:\My Unitex\English\Corpus\ivanhoe_snt\concord.ind"
"-fCourier new" -s12 -l40 -r55 --CL --html
"-aD:\My Unitex\English\Alphabet_sort.txt"
}
```

Symboles spéciaux

Caractère	Signification	Codage
"	les guillemets délimitent des séquences qui ne doivent ni être interprétées par Unitex, ni subir de variantes de casse	\"
+	+ sépare les différentes lignes boîtes	"+"
:	: sert à introduire à appel à un sous-graphe	":" or \:
/	/ indique le début de la sortie d'une boîte	\/
<	< indique le début d'un motif ou d'un méta	"<" or \<
>	> indique la fin d'un motif ou d'un méta	">" or \>
#	# sert à interdire la présence de l'espace	"#"
\	\ sert à déspécialiser la plupart des caractères spéciaux	\\

Savoir :

- utiliser des variables
- construire un dictionnaire fléchi de mots simples

Ajouter des entrées au dictionnaire

Créer les entrées fléchies des mots suivants

- twittos
- followeur
- dégagisme
- frotteur
- grossophobie