TD3

➔ *Modéliser des connaissances linguistiques à l'aide d'automates à états finis*

➔ *Utiliser un outil pour cela : UNITEX*

Les objectifs de ce TD un peu plus en détail :

- Se familiariser avec **Unitex**, un outil d'annotation de textes à l'aide de transducteurs, sortes d'automates finis. L'intérêt de cet outil réside dans son interface graphique qui permet de constituer et de maintenir des automates complexes.

- Apprendre à **construire des ressources linguistiques** qui vont permettre le **repérage** et l'**annotation automatiques** d'unités textuelles. Ici, cette annotation concernera une partie de ce qui entre communément dans l'analyse temporelle de textes : l'analyse des **adverbiaux temporels** (ex : « le 4 mars 2012 », « depuis lundi 11 août », « aux alentours de la mi-décembre », « en début d'année »).

- Se confronter à la complexité de la modélisation linguistique impliquant une réflexion sur les notions et critères linguistiques centraux … ce qui implique parfois de revenir sur ses premières intuitions.

- Dans un premier temps, vous suivrez les étapes 1 à 8 scrupuleusement. L'étape 9 vous laissera libre ensuite de proposer une modélisation sous la forme de transducteurs qui permettent de reconnaitre (au moins) les adverbiaux temporels présents dans le texte nommé *TexteVillepin*.

1. Installer & lancer Unitex/GramLab 3.1

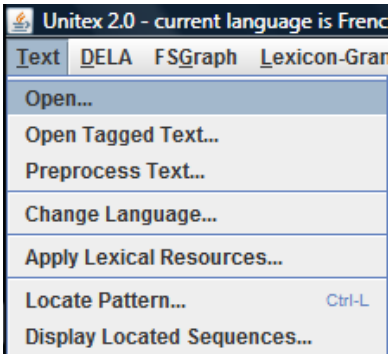
http://www-igm.univ-mlv.fr/~unitex/index.php?page=3&html=download2.html
ou
http://unitexgramlab.org/#downloads

- Télécharger l'application et le manuel utilisateur
- Lancer l'application
- Sélectionner la langue de travail (« French »)

Remarque : pour chaque langue de travail, Unitex crée un répertoire de travail contenant 3 sous répertoires importants : « Corpus » (pour les textes), « Dela » (pour les dictionnaires) et « Graphs » (pour les automates à états finis).

2. Charger un texte brut (pré-traitements : normalisation et segmentation)

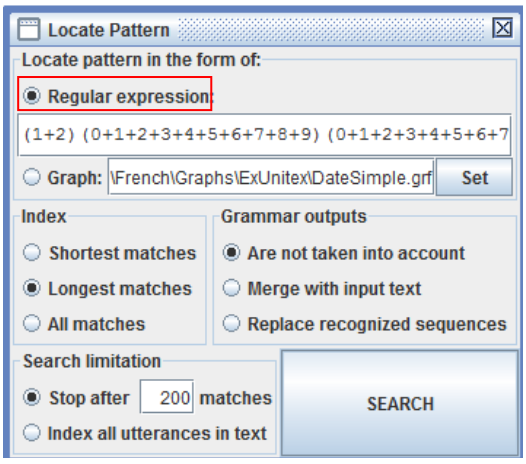
- copier le texte *TexteAFPCorpus* dans le répertoire « Corpus » dans l'arborescence du français



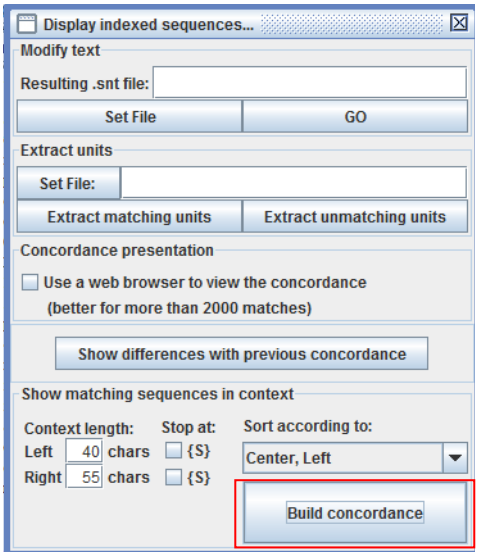
- aller à menu « Text/Open ». Dans la fenêtre qui s'ouvre, sélectionner la valeur « All files » pour le champ « Files of type »
- ouvrir le texte à traiter
- une fenêtre s'ouvre proposant d'effectuer sur le texte différents pré-traitements (tokenisation, normalisation des contractions, segmentation en phrases, dictionnaires de langue)
- lancer les pré-traitements
- le texte s'ouvre dans une fenêtre (au format .snt, propre à Unitex)

3. Tester des expressions régulières et la visualisation à l'aide d'un concordancier

- menu « Text/Locate Pattern »
- une fenêtre s'ouvre avec différents paramètres
- choisir le mode expression régulière



-
- saisir une expression régulière :
  - ex1 : un nom de mois (mars, par exemple)
  - ex2 : une expression régulière représentant des années :
    (1+2)(0+1+2+3+4+5+6+7+8+9)(0+1+2+3+4+5+6+7+8+9)(0+1+2+3+4+5+6+7+8+9)
- lancer la recherche du motif dans le texte (bouton « SEARCH »)
- une fenêtre apparaît proposant différents choix de traitement des résultats
- pour visualiser les résultats sous la forme d'un concordancier, cliquer sur le bouton en bas de la fenêtre « Build Concordance »



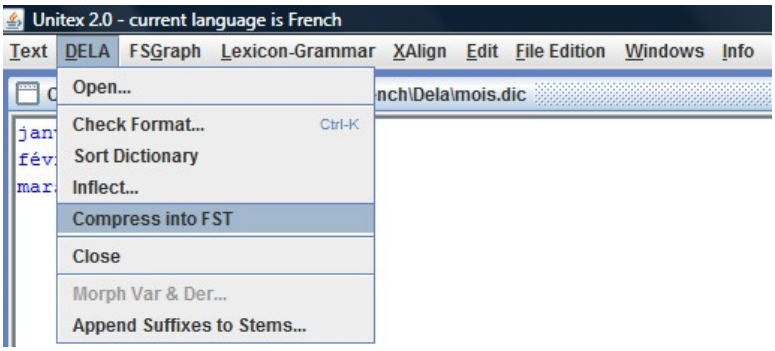
4. Constituer des dictionnaires simples (jour de la semaine, mois)

- menu « File Edition/New File »
- créer un dictionnaire des mois sur le modèle suivant :

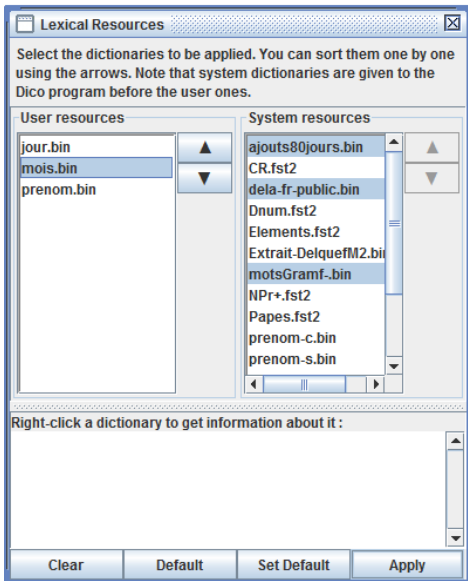
janvier,.MOIS
février,.MOIS
…
décembre,.MOIS

Pour info : une entrée de dictionnaire correspond au texte à reconnaître, la virgule permet de délimiter la partie décrivant des traits particuliers à l'entrée du dictionnaire. En majuscule figure l'annotation principale. Pour plus d'informations sur la syntaxe propre aux dictionnaires de type LADL, consulter le manuel d'Unitex.

- enregistrer le dictionnaire dans le répertoire « Dela » (format des dictionnaires du LADL) sous le nom « mois.dic »
- ouvrir le dictionnaire ainsi créé : menu « DELA/Open »
- compiler le dictionnaire sous la forme d'un automate à état fini : menu « DELA/Compress into FST »



- tester le dictionnaire sur le corpus de travail : menu « Text/Apply Lexical Resources »
- une fenêtre s'ouvre précisant les dictionnaires disponibles qui peuvent être appliqués au texte : sélectionner le dictionnaire nouvellement créé : « mois.bin »



- pour lancer la recherche des entrées des dictionnaires sur le texte, cliquer sur le bouton « Apply »
- tester l'expression régulière suivante « <MOIS> » (cf. reprendre au niveau de l'exercice 3)

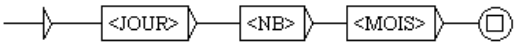
- en reprenant au début de l'exercice 4, créer un dictionnaire des jours de la semaine :

  lundi,.JOUR
  mardi,.JOUR
  mercredi,.JOUR
  jeudi,.JOUR
  vendredi,.JOUR
  samedi,.JOUR
  dimanche,.JOUR

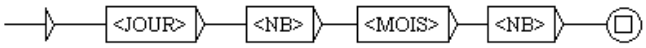
- enregistrer le dictionnaire dans le répertoire DELA sous le nom « jour.dic »
- tester l'expression régulière suivante « <JOUR><NB><MOIS> » (cf. reprendre au niveau de l'exercice 3). Pour info : dans la phase de tokenisation (étape du pré-traitements), les caractères présents dans le texte sont regroupés et répartis en différentes classes ou types de token. Sont ainsi distingués : les nombres <NB>, les « mots » <MOT>, la ponctuation <PNC> et les espaces

5. Créer une grammaire simple (repérage)

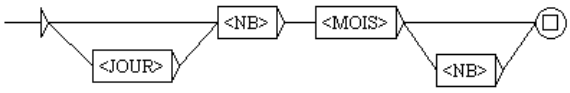
- menu « FS-Graph/New »
- créer la grammaire suivante (CTRL+clic pour créer un nœud ; sélectionner un nœud puis un second pour les relier ou les délier)



- enregistrer la grammaire (menu « FSGraph/Save ») sous le nom « DateSimple.grf » dans le répertoire dédié aux graphes (répertoire « Graphs »)
- lancer la grammaire sur le texte : menu « Text/Locate Pattern »
- dans la fenêtre qui s'ouvre, sélectionner le mode « Graph » et choisir la grammaire nouvellement créée
- observer le concordancier
- afin de repérer des expressions temporelles dans leur intégralité, modifier la grammaire pour capter les années et observer le concordancier obtenu



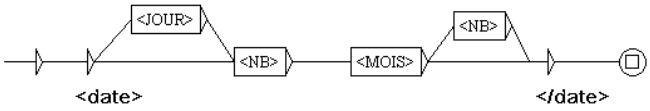
-
- afin de repérer davantage d'expressions temporelles, modifier la grammaire pour rendre la présence du jour et de l'année optionnelle (notion d'embranchement) et observer le concordancier



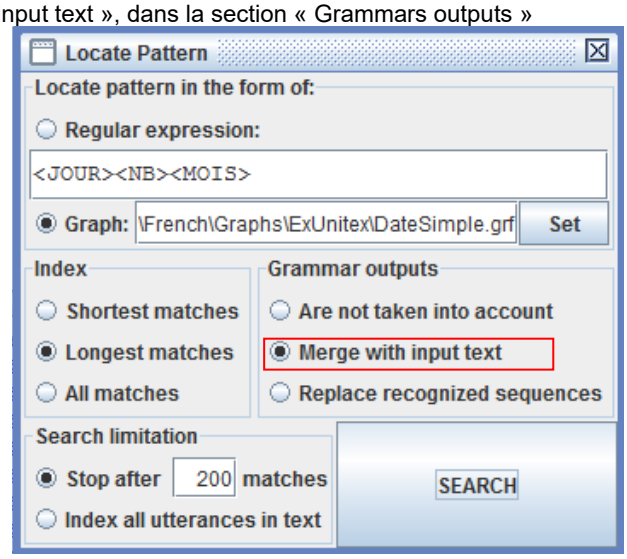
6. Créer un transducteur (sortie au format xml)

6.1 Une sortie simple

- modifier le graphe pour qu'il produise une annotation dans le texte, lorsqu'une expression temporelle est repérée
- pour cela, créer des nœuds pour les annotations (ex de nœud produisant une annotation : « <E>/<date> » (Pour info : <E> est un nœud vide)



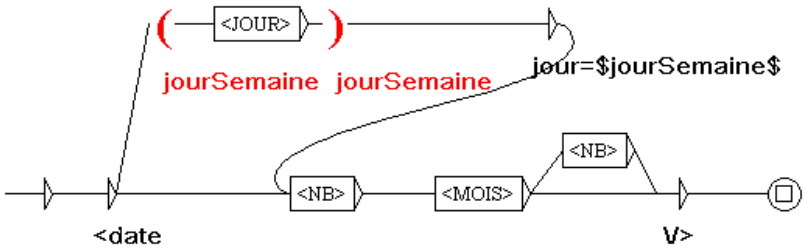
- menu « Text/Locate Pattern »
- dans la fenêtre qui s'ouvre, afin de visualiser les annotations, sélectionner le graphe modifié et choisir le mode « merge with input text », dans la section « Grammars outputs »



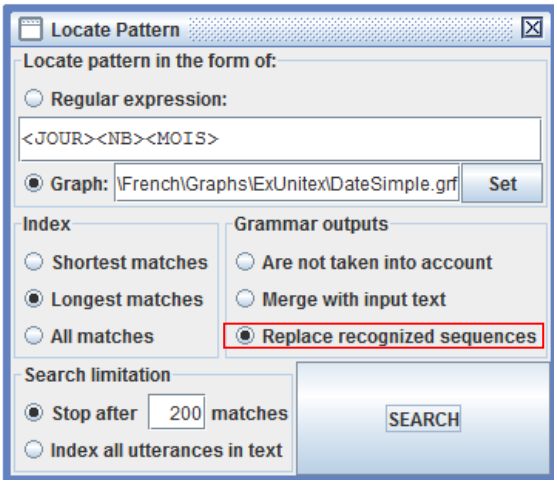
- observer le concordancier

6.2 Une sortie avec des variables

- dans la sortie, on peut souhaiter conserver la valeur de certaines informations repérées (le jour de la semaine, le jour du mois, le mois, l'année)
- pour conserver de telles valeurs, il faut créer des variables
- modifier le graphe de sorte à mettre la valeur du jour de la semaine dans une variable. Pour cela, entourer le jour par deux nœuds (« $jourSemaine(» pour la 1e balise et « $jourSemaine) » pour la seconde). On récupère la valeur de la variable en créant une annotation sur le modèle suivant : « <E>/jour=$jourSemaine$ »



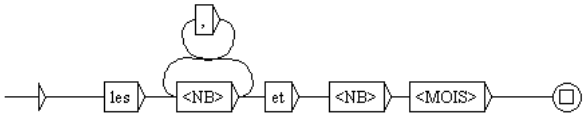
- modifier le graphe pour récupérer la valeur des jours du mois, du mois, de l'année et du texte (qui enserre tout le motif repéré)
- lancer l'annotation (menu « Text/Locate Pattern)
- dans la fenêtre qui s'ouvre sélectionner la valeur « Replace Recognize sequence »



- observer le concordancier

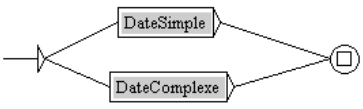
7. Créer une grammaire pour annoter des expressions calendaires complexes (motifs répétés)

- On peut souhaiter capter des expressions plus complexes (ex : les 27, 28 et 29 juin)
- Créer le graphe suivant (DateComplexe.grf) :



8. Pour aller plus loin

- Tester les appels de sous graphes dans un graphe. Syntaxe : deux points suivis du graphe appelé ex « :nomDuGraphe » (ex : « :DateSimple » pour appeler le graphe « DateSimple.grf »)



- Réfléchir à la notion d'ambiguïté sémantique et de contexte local. Ex : ambiguïté entre les références à des années et à des nombres (« en 1960 » et « 1960 personnes évacuées ».) D'où l'intérêt de capter des éléments de contexte local qui ne seront pas nécessairement annotés.
- Tester la négation de motifs (cf. manuel Unitex chap. 6.3)

9. Application à l'annotation de l'ensemble des adverbiaux temporels d'un texte

- Créez un ensemble de graphes permettant de repérer et d'annoter l'ensemble des adverbiaux temporels présents dans le texte biographique sur D. de Villepin selon la catégorisation que vous aurez retenue.