Linguistique de corpus

Outils et méthodes de traitement de corpus
Filtrage de corpus avec Unitex

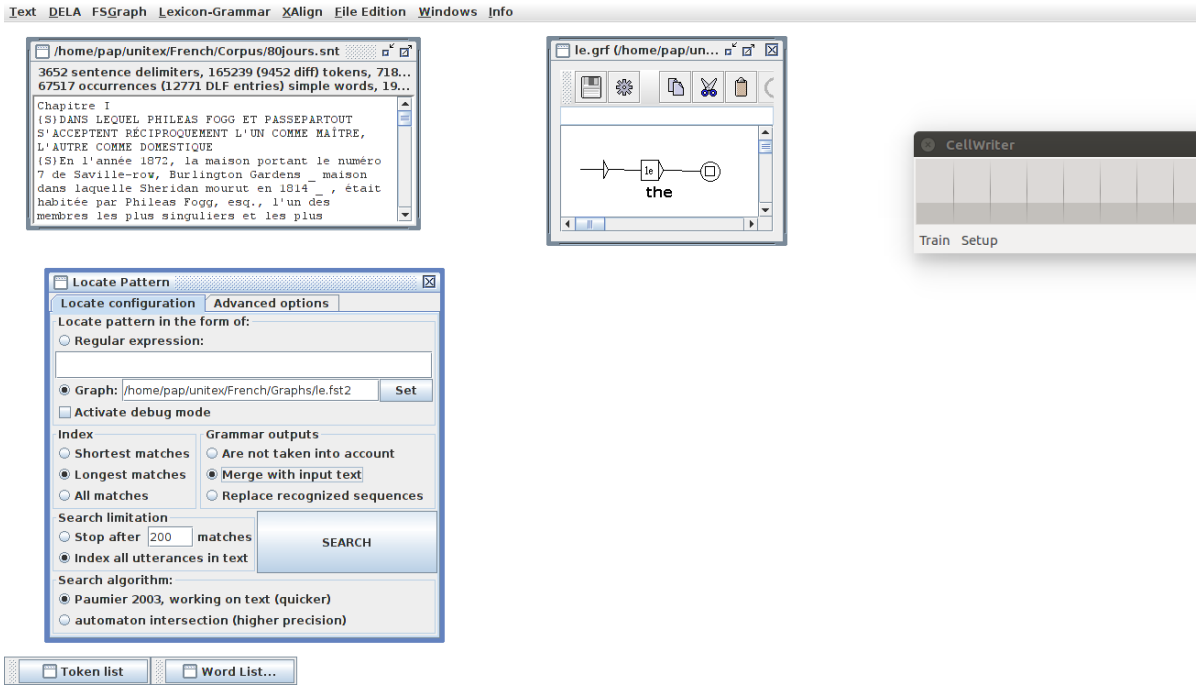
Patrick Paroubek

LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris XI, 91403 Orsay Cedex
pap@limsi.fr

mercredi 22 février 2017 / L3 - Cours S2-3

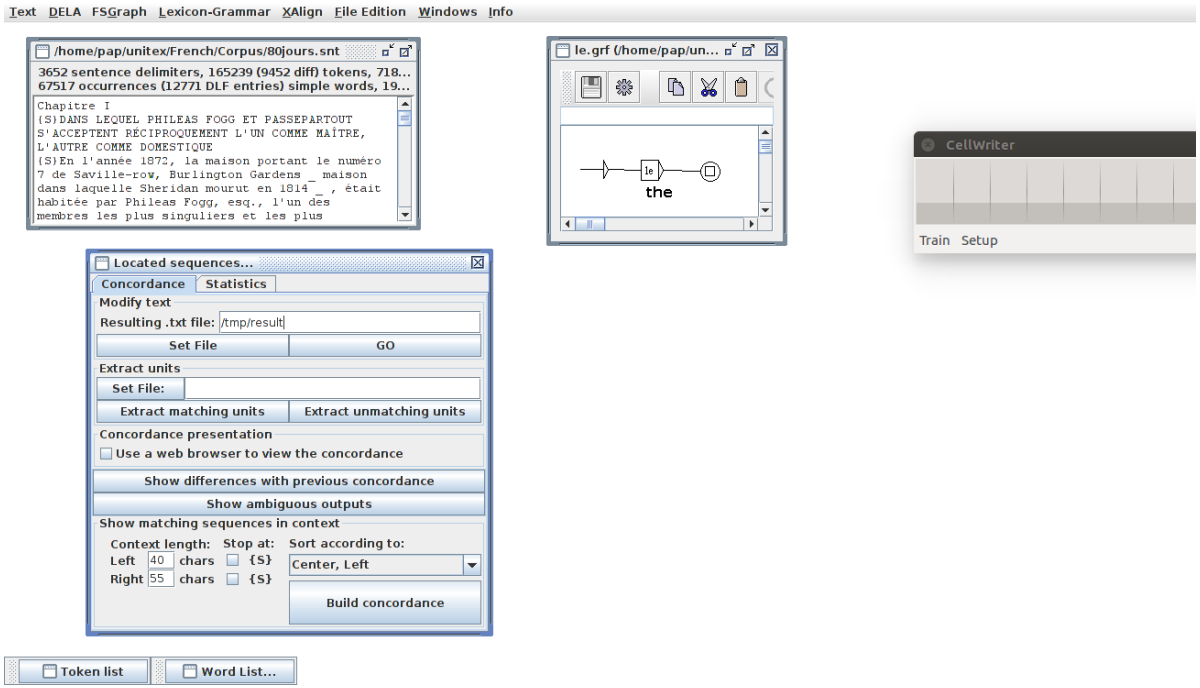
Pratique Unitex
Reconnaissance et transcription

On veut reconnaître un motif et le remplacer par un autre mais l'on souhaite récupérer dans un fichier le résultat de cet traitement (uniquement les mots concernés).



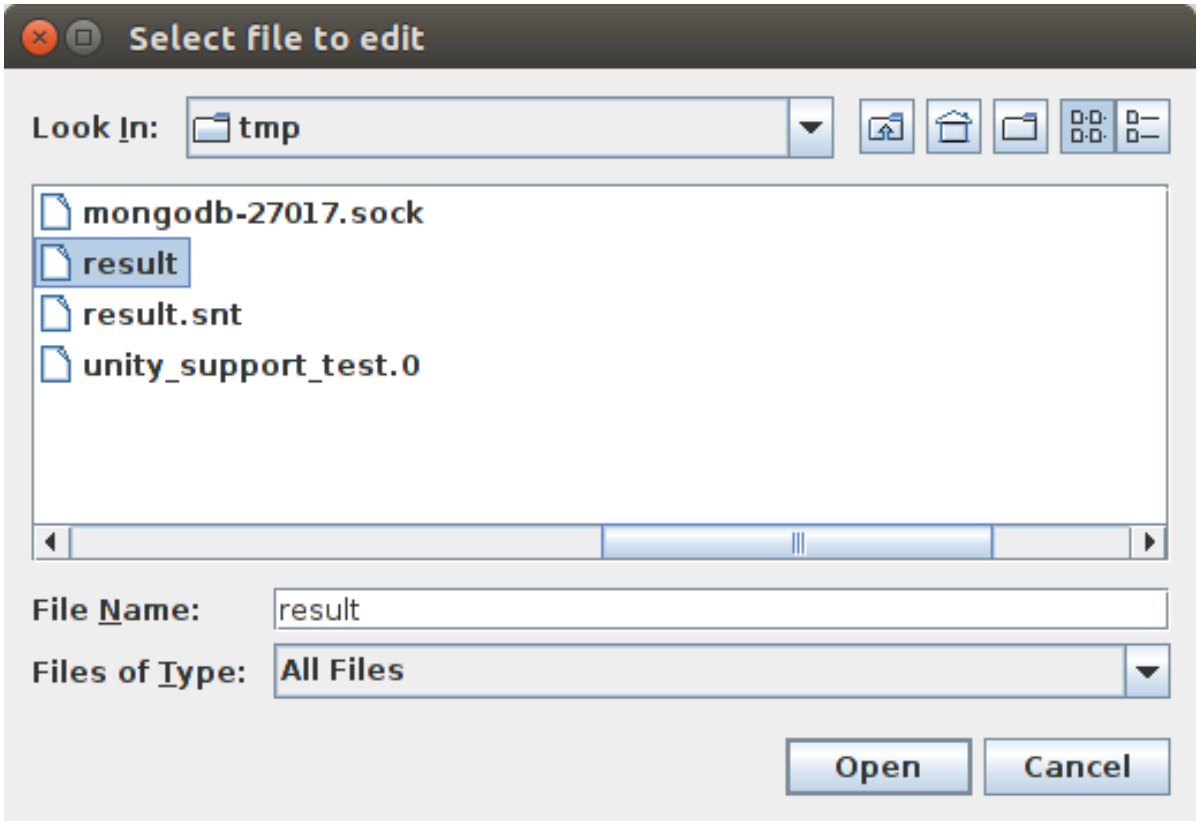
Pratique Unitex
Reconnaissance et transcription

on sauvegarde le résultat dans un fichier



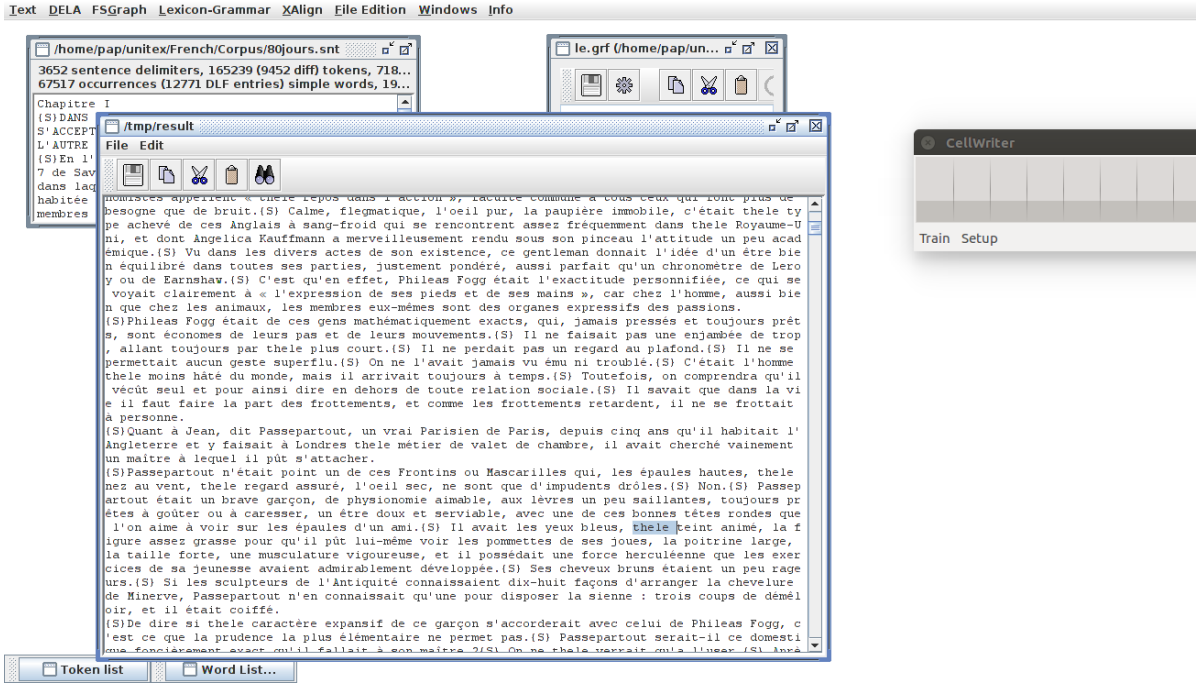
Pratique Unitex
Reconnaissance et transcription

On charge le fichier résultat dans un éditeur.



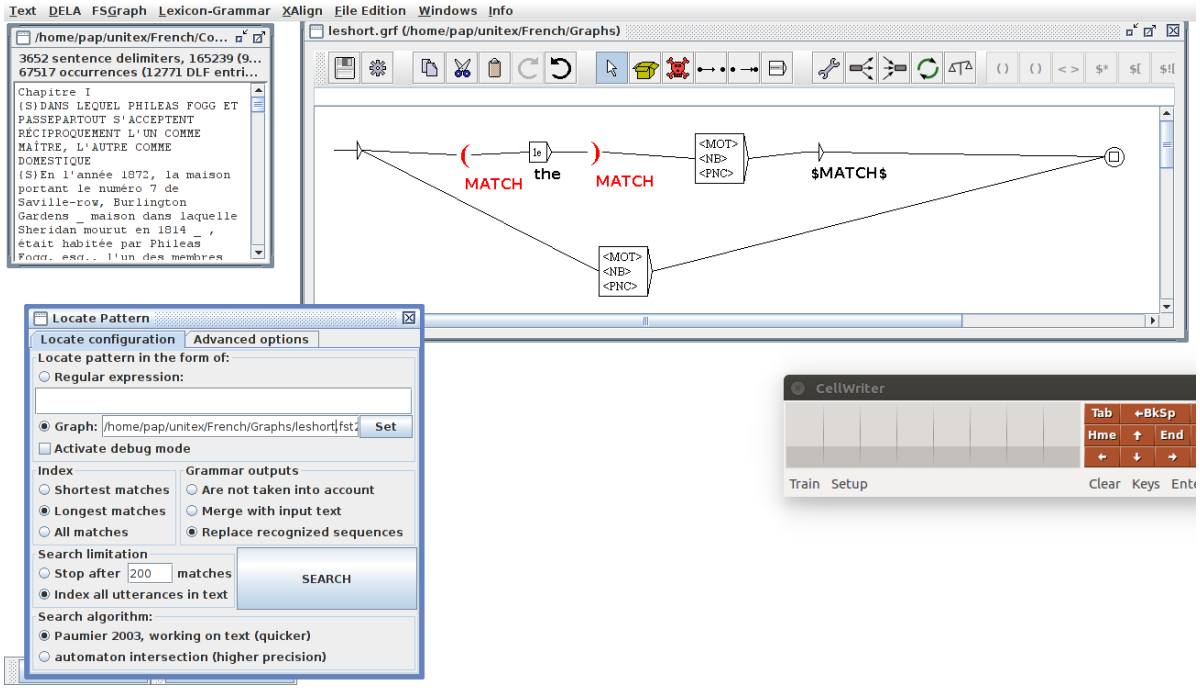
Pratique Unitex
Reconnaissance et transcription

Problème : on récupère les phrases dans lesquelles le motif apparaît MAIS avec tous leurs mots.



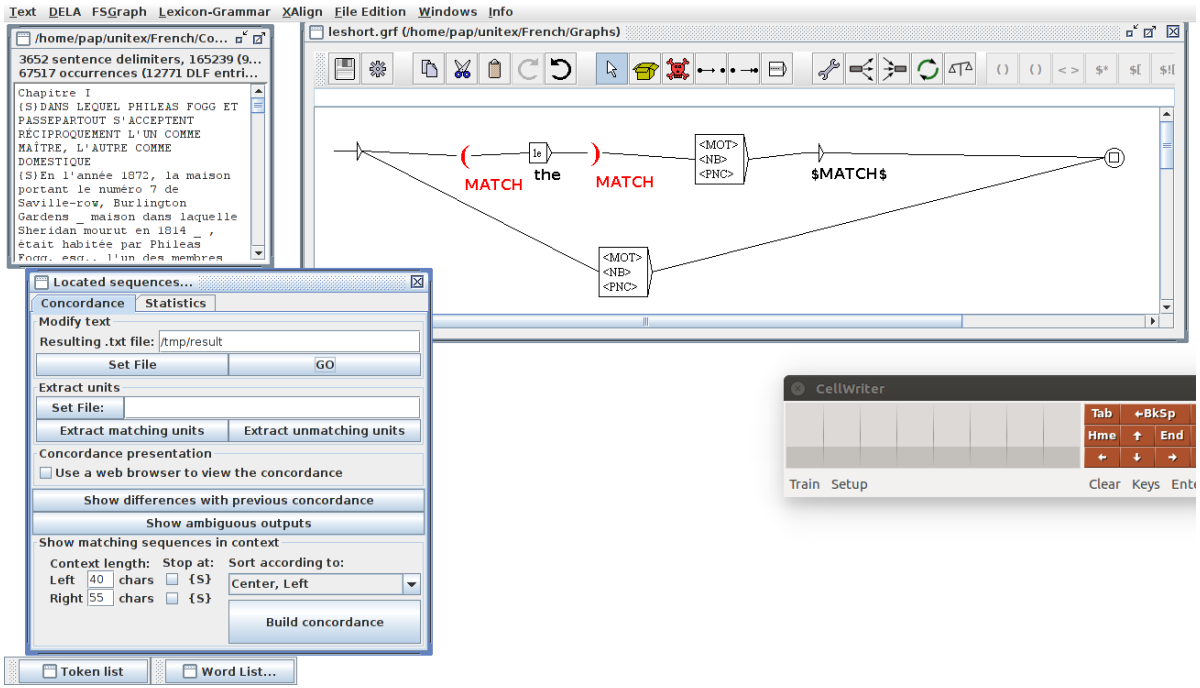
Pratique Unitex
Reconnaissance et transcription

Solution, on ajoute une transition “poubelle” qui va se déclencher sur tous les motfis complémentaires de celui que l’on recherche ; et demander un reconnaissance avec remplacement du texte reconnu ; bien sur en remplaçant les texte reconnu par la transition “poubelle” par rien.

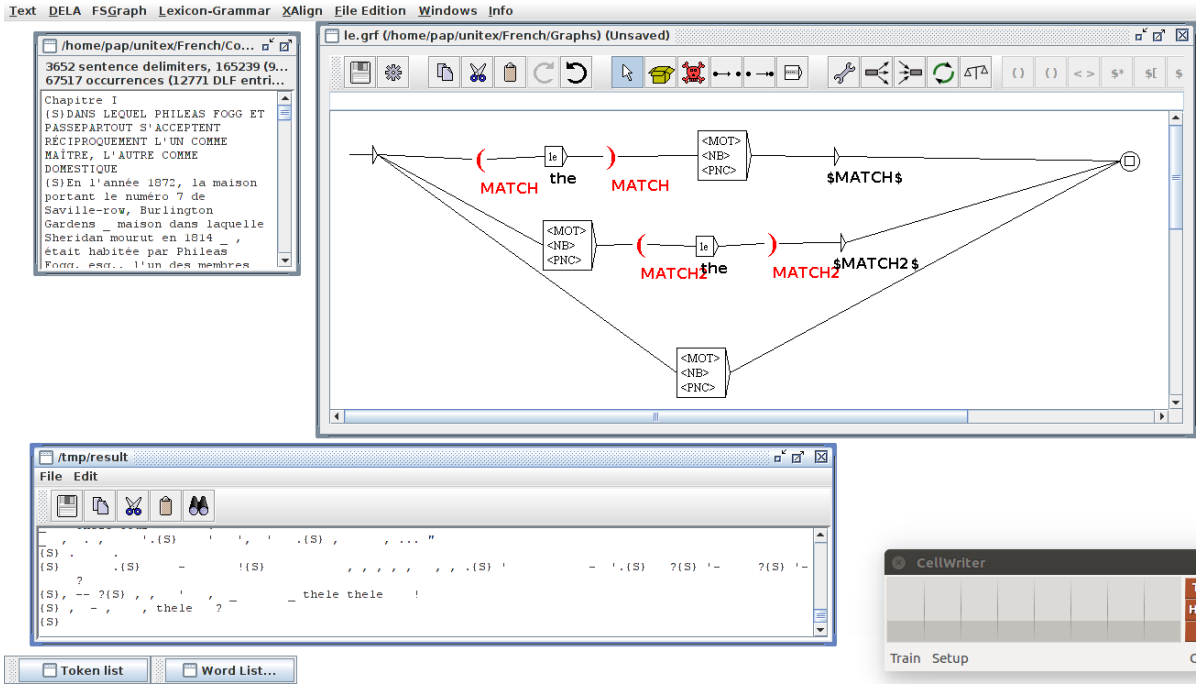


Pratique Unitex

Reconnaissance et transcription



Souvent les motifs génériques comme <MOT>, <NB>, <PNC>, ne suffisent pas pour éliminer le texte complémentaire



du motif recherché

Il suffit d'ajouter dans la transition "poubelle" des motifs pour couvrir les charactéres et mots non encore filtrés, en plusieurs itération ; à la fin il ne reste que les motifs générées et des balises `{S}`