

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

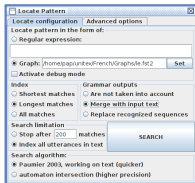
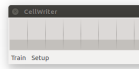
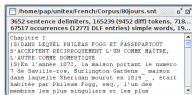
mercredi 22 février 2017 / L3 - Cours S2-3

Pratique Unix

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).

Text DDLA FSGraph Lexicon-Grammar XAlign Elle Edition Windows Info



Pratique Unix

Reconnaissance et transcription

on sauvegarde le résultat dans un fichier

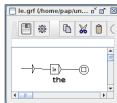
Text DELA F5Graph Lexicon-Grammar XAlign Elle Edition Windows Info

```

/home/pap/unite/frenchCorpus/9jours.snt
3652 sentence delimiters, 165239 19452 dHD tokens, 718...
67917 occurrences (12773 DLF entries) single words, 18...

Chapitre 1
(8)DANS LAQUEL PHILEAS FOGG ET PASSEPARTOUT
9 ACCEPTENT RECIPROQUEMENT L'UN COMME MAÎTRE,
L'AUTRE COMME DOMESTIQUE
(18)En l'année 1872, in maison portait le numéro
7 de Saville-row, Burlington Gardens, maison
dans laquelle Sheridan mourut en 1816, était
habitée par Phileas Fogg, qui, l'un des
meilleurs jeu plus singuliers et les plus

```



Cellwriter

Train Setup

Located sequences...

Concordance Statistics

Modify text

Resulting txt file: /tmp/result

Set File GO

Extract units

Set File:

Extract matching units Extract unmatching units

Concordance presentation

☐ Use a web browser to view the concordance

Show differences with previous concordance

Show ambiguous outputs

Show matching sequences in context

Context length: Stop at: Sort according to:

Left ☐ chars ☐ (63) Center, Left

Right ☐ chars ☐ (53)

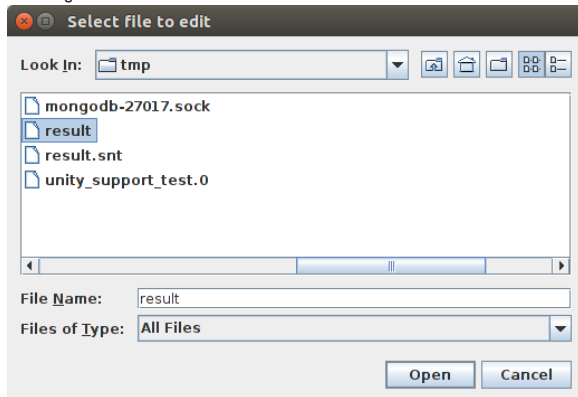
Build concordance

Token list Word List...

Pratique Unix

Reconnaissance et transcription

On charge le fichier résultat dans un éditeur.



Pratique Unix

Reconnaissance et transcription

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

Text DELA FSGraph Lexicon-Grammar XAlign Elle Edition Windows Info

3652 sentence delimiters, 16529 19452 dtd tokens, 718...
67917 occurrences (12771 dtd entries) single words, 18...

le.grf (home/pap/unix...)

File Edit

Word List...

Pratique Unix

Reconnaissance et transcription

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

The screenshot displays a Unix-like environment with three main windows:

- Text Editor (Left):** Shows a file named 'Chapter 1' with French text. The first line is '3652 sentence delimiters, 165239 (9... 87512 occurrences (1277) DLS enti...'. Below it, a list of words is shown: 'Chapitre 1', '1918 DANS LEQUEL, PRILEAS FOUG ET PASSEPARTOUT S'ACCEPTENT RECIPROQUEMENT L'EN CORNE MATRE, L'AUTRE CORNE ORESTEUSE', '1918 S'ANNEE 1972, LA MAISON PORTAIT LE NUMERO 7 de Seville-row, Burlington Gardens - maison dans laquelle Sheridan mourut en 1814 - était habitée par Philéas Fogg, etc. Une des scènes'.
- Graph Editor (Center):** Shows a graph with two nodes. The top node has a transition labeled 'MATCH' with a red 'the' in the middle. The bottom node has a transition labeled 'MATCH' with a red 'the' in the middle. The graph is titled 'leshort.grf (home/papunitex/French/Graphs)'.
- Search Tool (Bottom Left):** A 'Locate Pattern' dialog box. It has tabs for 'Locate configuration' and 'Advanced options'. The 'Locate pattern in the form of:' section has a radio button for 'Regular expression'. The 'Graph:' field is set to '(home/papunitex/French/Graphs/eshof5tc)'. The 'Activate debug mode' checkbox is checked. The 'Index' section has radio buttons for 'Shortest matches', 'Longest matches', and 'All matches'. The 'Search limitation' section has a 'Stop after' field set to '200 matches'. The 'Search algorithm:' section has radio buttons for 'Paumier 2003, working on text (quicker)' and 'automaton intersection Dighier precision'.
- CellWriter (Bottom Right):** A small window titled 'CellWriter' with a grid of cells and buttons for 'Tab', 'BkSp', 'Hme', 'End', 'Train', 'Setup', 'Clear', 'Keys', and 'End'.

Pratique Unix

Reconnaissance et transcription

The screenshot displays the XAlign software interface, which is used for creating and analyzing word lists and concordances. The main window shows a concordance search for the word "the". The search results are displayed in a list, with the word "the" highlighted in red. The search criteria are set to "MATCH" and "the". The search results are displayed in a list, with the word "the" highlighted in red. The search criteria are set to "MATCH" and "the".

The interface includes a menu bar (Text, DELA, F5Graph, Lexicon-Grammar, XAlign, File, Edition, Windows, Info) and a toolbar with various icons for file operations and editing. A file list on the left shows the current project files. The main window displays a concordance search for the word "the", with the search results displayed in a list. The search criteria are set to "MATCH" and "the".

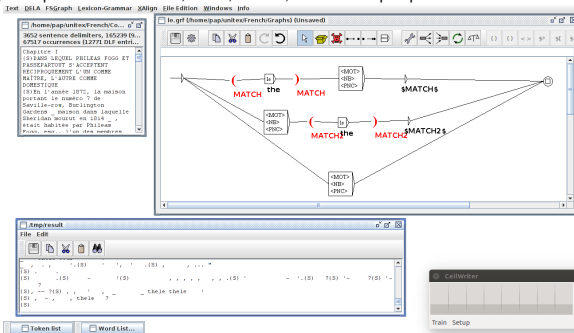
A "Located sequences..." dialog box is open, showing the "Concordance" tab. It includes fields for "Modify text", "Resulting .txt file", and "Set File". The "Extract units" section has buttons for "Extract matching units" and "Extract unmatching units". The "Concordance presentation" section has a checkbox for "Use a web browser to view the concordance". The "Show differences with previous concordance" section has a checkbox for "Show ambiguous outputs". The "Show matching sequences in context" section has a "Build concordance" button.

A "CellWriter" window is also visible, showing a table with columns for "Tab", "BkSp", "Home", and "End". The table contains the following data:

Tab	BkSp	Home	End

The "CellWriter" window also includes a "Train Setup" button and a "Clear Keys" button.

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 caractères et mots non encore filtrés, en plusieurs itération ; à la fin il ne reste que les motifs générés et des balises { S }

