



Plate-formes logicielles pour le TAL 3 : Unitex - Flexion(s) automatique(s)

Karën Fort

karen.fort@sorbonne-universite.fr / <https://www.schplaf.org/kf/>

2 octobre 2020



Quelques sources d'inspiration

- ▶ Manuel d'Unitex : <http://www-igm.univ-mlv.fr/~unitex/index.php?page=4>
- ▶ Denis Maurel : son aide et ses conseils

Sources

Rappel sur les dictionnaires d'Unitex

Créer un dictionnaire fléchi

Flexions plus complexes

Flexions des mots composés

Pour finir

Application des dictionnaires

```
Dico "-tFrench/Corpus/80jours.snt"  
"-aFrench/Alphabet.txt" "French/Dela/dela-fr-public.bin"  
"French/Dela/ajouts80jours.bin" "French/Dela/motsGramf-.bin"
```

- ▶ .bin : format compressé
- ▶ possibilité d'utiliser des graphes dictionnaires (.fst2)

Application des dictionnaires du français sur le TDM

The screenshot displays a software window titled "Word Lists in /home/fortkare/unitex/French/Corpus/80jours_snt". The window is divided into three main sections:

- DLF: 12771 simple-word lexical entries**: A list of words with their grammatical tags, such as "a, .N+z1:ms:mp", "à, .PREP+z1", "a, avoir, V+z1:P3s", "abaissait, abaisser, V+z1:I3s", "abaissant, .A+z2:ms", "abaissant, abaisser, V+z1:G", "abaissé, .A+z1:ms", "abaissé, abaisser, V+z1:Kms", "abaissement, .N+z2:ms", "abandonna, abandonner, V+z1:J3s", and "abandonnait, abandonner, V+z1:I3s".
- DLC: 2055 compound lexical entries**: A list of compound words with their grammatical tags, such as "à base de, .PREP+EPCDN+z1", "à bon droit, .ADV+PAC+z1", "à bord de, .PREP+EPCDN+z1", "à bord des, à bord de, .PREP+EPCDN+z1", "à califourchon sur, .PREP+EPCPN+z1", "à cause de, .PREP+EPCPQ+z1", "à cause de, .PREP+PCDN+z1", "à cause de, .PREP+PCDN1+z1", "à cause de, à cause, .PREP+Prépconjs+1", "à cause, .ADV+PCDN+z1", "à ces mots, .ADV+PDETC+z1", "à cet effet, .ADV+PDETC+z1", "à cet égard, .ADV+PDETC+z1", "à chaque instant, .ADV+PDETC+z1", "à cheval, .A+EPC+z1", "à condition que, à condition, .CONJS+6", "à coup sûr, .ADV+PCA+z1", "à coups de, .PREP+PCDN+z1", "à coups de, .PREP+PCDN1+z1", and "à coups, .ADV+PCDN+z1".
- ERR: 449 unknown simple words**: A list of words that were not found in the dictionaries, including "Abraham", "Aden", "afin", "Afrique", "Agra", "Ahmémnagara", "Alabama", "Albermale", "Allahabad", "Allemagne", "Andaman", "Andrew", "Angelica", "Angleterre", "Annam", "Aouda", "Arkansas", "Armonica", "Arrien", "Arthémidore", "Asie", "Assurghur", "Athènes", "Aureng", "Aurangabad", "bambousiers", "Bank", "Baring", "BATULCAR", "Batulcar", "Béhar", and "Bénarès". There is a checkbox labeled "Filter unknown words with tags.ind" which is currently unchecked.

Contenu d'un dictionnaire Unitex

Dictionnaire (Unitex)

un ensemble d'entrées lexicales

- ▶ exemple d'entrée lexicale :

institutrice, instituteur.N+Hum :fs

- ▶ forme de base (ou canonique, ou lemme) : *instituteur*
- ▶ catégorie grammaticale : nom (*N*)
- ▶ informations flexionnelles (genre, nombre) : *fs*
- ▶ forme fléchie : *institutrice*
- ▶ traits syntactico-sémantiques : *Humain*

Mots simples vs mots composés (pour Unitex)

Mot simple

une séquence de lettres : délimitation par des séparateurs (espaces, ponctuation, etc.)

Mot composé

une séquence de mots simples, dont le sens est non compositionnel : *cordons bleus*, *pomme de terre*, *belle famille*, *porte-manteau*, ...

Les dictionnaires Unitex

Deux types :

1. dictionnaires de formes simples (DELAS)
2. dictionnaires de formes fléchies (DELAF)

qui comprennent des formes simples ou composées

DELAS :

`cheval,N4+An1`

DELAF :

`mercantiles,mercantile.A+z1:mp:fp/ceci est un exemple
grand=mères,grand=mère.N:fp`

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)

DELAS → Flexion automatique → DELAF

Traitement des dictionnaires

Compression automatique des dictionnaires (en transducteurs)

Avantages :

- ▶ taille mémoire
- ▶ accès à l'information

Sources

Rappel sur les dictionnaires d'Unitex

Créer un dictionnaire fléchi

Flexions plus complexes

Flexions des mots composés

Pour finir

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 :

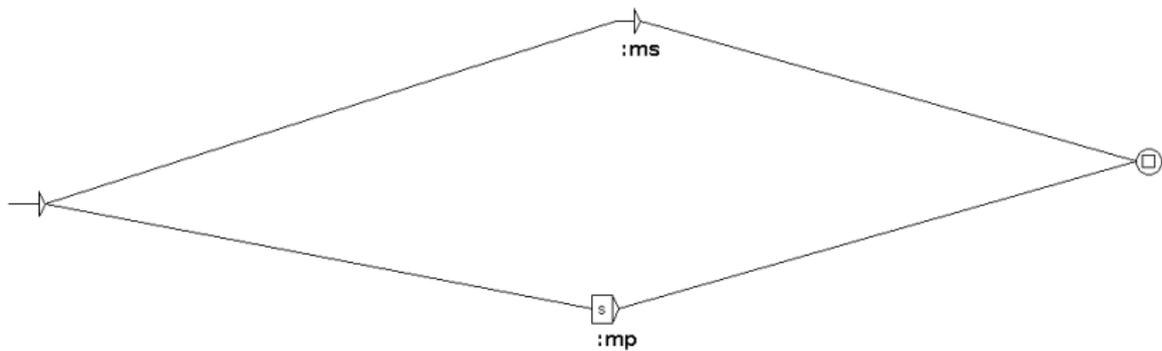
- ▶ d'ajouter un s au (masculin) pluriel : s/ :mp
- ▶ de ne rien ajouter au (masculin) singulier : <E>/ :ms

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

! pas d'espace

Étape 2 : résultat

ballon

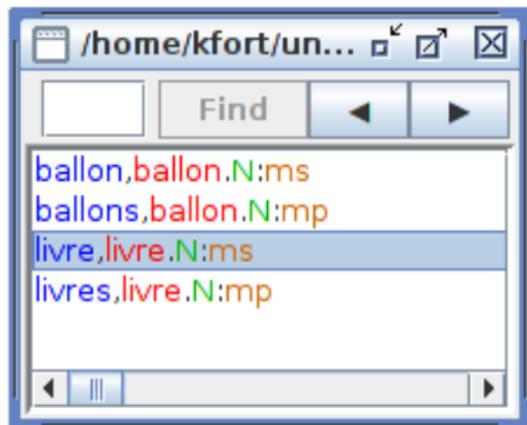


Étape 3 : créer le dictionnaire fléchi

Menu DELA / Open

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

Étape 3 : résultat



Exercice

Ajouter :

- ▶ le verbe twitter : le graphe de flexion est déjà dans Unitex (V3)
- ▶ amour, délice, orgue (masculin au singulier, féminin au pluriel) : créer le graphe de flexion

Tester le dictionnaire créé

Sur le Tour du monde en 80 jours (par exemple)

- ▶ Dela/Inflect (générer les formes fléchies)
- ▶ Dela/Compress into FST (à partir du dico fléchi)
- ▶ Text/Open le Tour du monde en 80 jours (PAS de preprocessing)
- ▶ Text/Apply lexical resources
 - ▶ Clear
 - ▶ sélectionner le dico
 - ▶ Apply

Tadaaaa !

Sources

Rappel sur les dictionnaires d'Unitex

Créer un dictionnaire fléchi

Flexions plus complexes

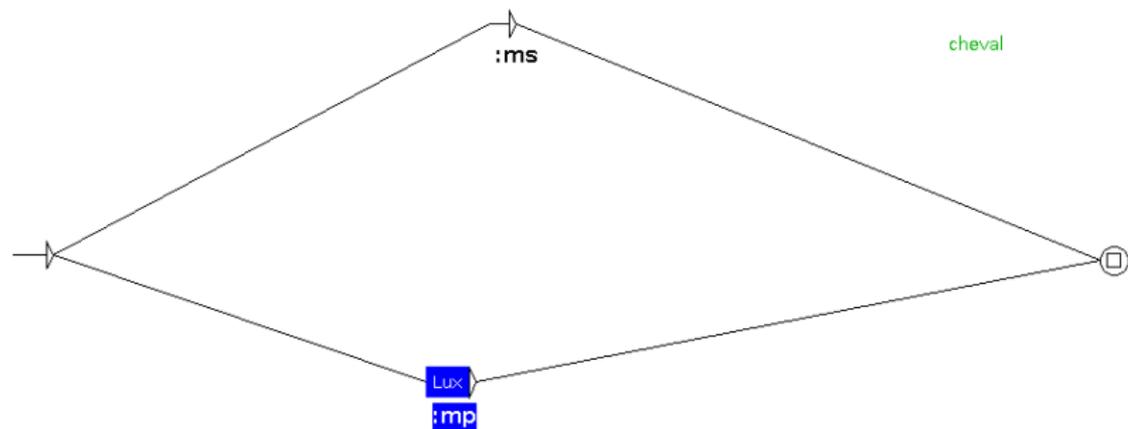
Flexions des mots composés

Pour finir

Exercice à cheval

- ▶ ajouter *cheval*, *N1001* dans le dictionnaire
- ▶ créer un graphe de flexion pour *cheval* :
 - ▶ L permet le déplacement d'une lettre à gauche
 - ▶ Lux permet donc de remplacer l par ux

Cheval : graphe de flexion



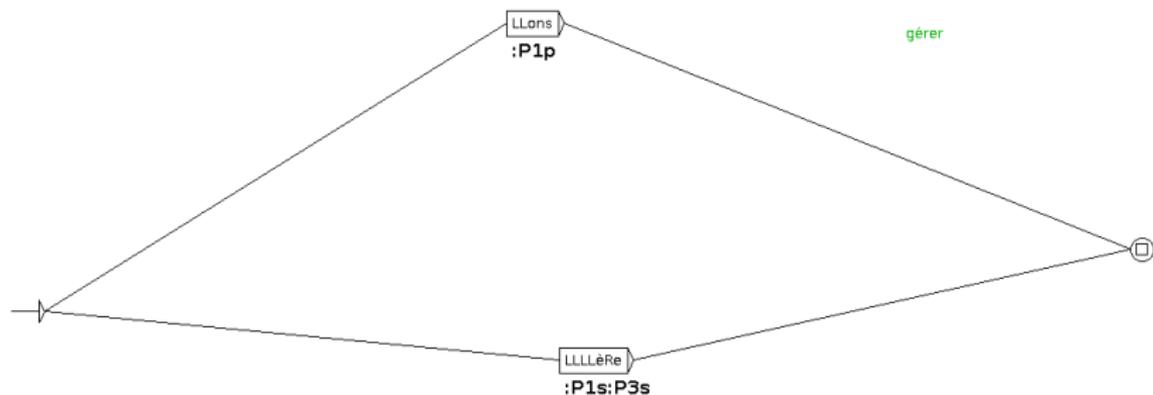
Lux/ :mp

Exercice à cheter

peler, acheter, gérer :

- ▶ ajouter-les dans le dictionnaire (V1000)
- ▶ créer un (et pas trois) graphe de flexion pour ces trois verbes :
 - ▶ R permet le déplacement d'une lettre à droite (quelle qu'elle soit)
 - ▶ C permet la copie d'une lettre (il est toujours avec R : RC)
 - ▶ LLLLèRe donc...

Gérer, peler, acheter : graphe de flexion



LLLLèRe/ :P1s :P3s

Sources

Rappel sur les dictionnaires d'Unitex

Créer un dictionnaire fléchi

Flexions plus complexes

Flexions des mots composés

Pour finir

Les mots composés en français

Peuvent être composés très diversement, par exemple :

- ▶ N-N : *porte-fenêtre* (à la fois une porte et une fenêtre)
- ▶ V-N : *cure-dent* (qui cure les dents)
- ▶ V-V : *garde-manger* (qui garde ce qui se mange)
- ▶ Prep ou Adv - N : *arrière-boutique*
- ▶ Adj-N : *basse-cour*
- ▶ etc.

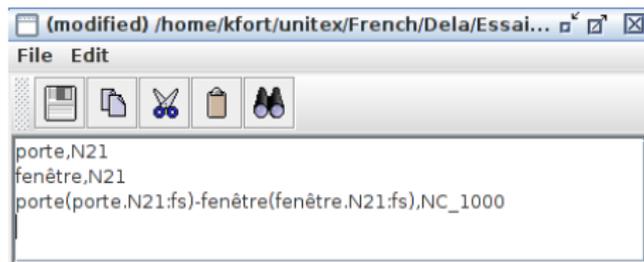
Leur flexion dépend de leur composition :

- ▶ N-N : les 2 au pluriel
- ▶ V-N : le nom seul prend le pluriel
- ▶ V-V : pas de marque du pluriel
- ▶ etc

Ajout de *porte-fenêtre* dans le dico

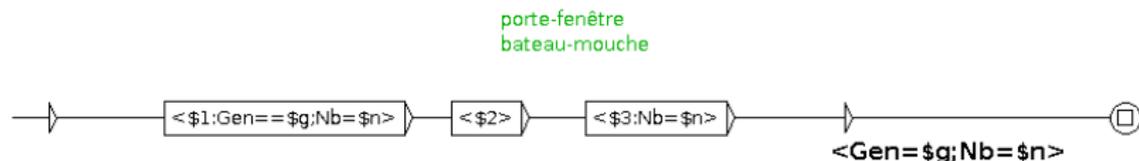
porte-fenêtre étant composé de *porte* et de *fenêtre*, il faut les ajouter dans le dictionnaire, AVANT *porte-fenêtre*

- ▶ leur flexion est définie dans le graphe N21
- ▶ il faut indiquer, pour le composé, quelle forme de ces mots il doit prendre (ici, le féminin singulier)
- ▶ les graphes de flexion des mots composés doivent avoir un nom commençant par NC_



```
(modified) /home/kfort/unitex/French/Dela/Essai...
File Edit
[Icons: Save, Copy, Paste, Undo, Redo]
porte,N21
fenêtre,N21
porte(porte.N21:fs)-fenêtre(fenêtre.N21:fs),NC_1000
```

Graphe de flexion pour *porte-fenêtre*

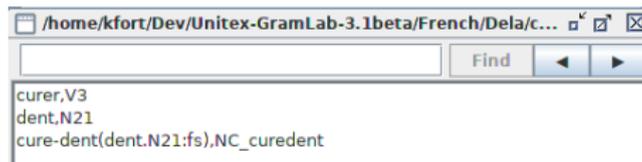


- ▶ on utilise des variables (voir cours de M1)
 - ▶ \$1 pour le premier élément du mot composé
 - ▶ \$2 pour le deuxième (-)
 - ▶ \$3 pour le troisième
- ▶ == signifie qu'on hérite de (ici le mot composé hérite du genre du premier élément, un seul genre donc)
- ▶ Nb=\$n signifie que Nb prend toutes les valeurs de nombre possibles (sg et pl)

Ajout de *cure-dent* dans le dico

cure-dent : composé de *curer* et de *dent*

- ▶ *curer* : flexion en V3
- ▶ *dent* : flexion en N21
- ▶ *cure-dent* : cure est invariable



The screenshot shows a window titled `/home/kfort/Dev/Unitex-GramLab-3.1beta/French/Dela/c...`. It features a search bar with the text "Find" and navigation arrows. Below the search bar, the following text is displayed:

```
curer,V3  
dent,N21  
cure-dent(dent.N21:fs),NC_curedent
```

Graphe de flexion pour *cure-dent*



Sources

Rappel sur les dictionnaires d'Unitex

Créer un dictionnaire fléchi

Flexions plus complexes

Flexions des mots composés

Pour finir

CQFR : Ce Qu'il Faut Retenir



- ▶ créer un dictionnaire
- ▶ créer un graphe de flexions
- ▶ utiliser les règles de déplacement