



Introduction à Unitex - Prétraitements

Karën Fort

karen.fort@sorbonne-universite.fr / <https://members.loria.fr/KFort/>



Quelques sources d'inspiration

- ▶ <https://unitexgramlab.org/fr>
- ▶ Manuel d'Unitex :
<https://unitexgramlab.org/releases/3.2/man/Unitex-GramLab-3.2-usermanual-fr.pdf>
- ▶ Cours de M. Constant, Université de Marne-la-Vallée

Sources

Introduction

Présentation

Fonctionnalités

Prise en main d'Unitex

Derrière le rideau

Pour finir

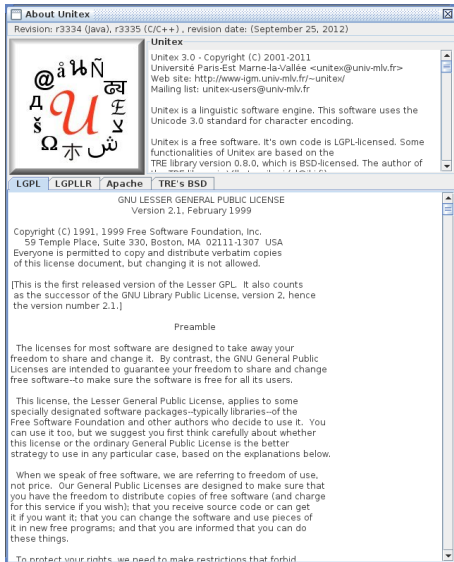
Installer Unitex

C'est ici

<https://unitexgramlab.org/fr>

Ce que dit Unitex à propos de d'Unitex

Menu Info / About Unitex...



About Unitex

Revision: r3334 (Java), r3335 (C/C++) , revision date: (September 25, 2012)

Unitex

Unitex 3.0 - Copyright (C) 2001-2011
Université Paris-Est Marne-la-Vallée <unitex@univ-mlv.fr>
Web site: <http://www-igm.univ-mlv.fr/~unitex/>
Mailing list: unitex-users@univ-mlv.fr

Unitex is a linguistic software engine. This software uses the Unicode 3.0 standard for character encoding.

Unitex is a free software. It's own code is LGPL-licensed. Some functionalities of Unitex are based on the TRE library version 0.8.0, which is BSD-licensed. The author of

LGPL **LGPLLR** **Apache** **TRE's BSD**

GNU LESSER GENERAL PUBLIC LICENSE
Version 2.1, February 1999

Copyright (C) 1991, 1999 Free Software Foundation, Inc.
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA
Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

[This is the first released version of the Lesser GPL. It also counts as the successor of the GNU Library Public License, version 2, hence the version number 2.1.]

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public Licenses are intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users.

This license, the Lesser General Public License, applies to some specially designated software packages--typically libraries--of the Free Software Foundation and other authors who decide to use it. You can use it too, but we suggest you first think carefully about whether this license or the ordinary General Public License is the better strategy to use in any particular case, based on the explanations below.

When we speak of free software, we are referring to freedom of use, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish); that you receive source code or can get it if you want it; that you can change the software and use pieces of it in new free programs; and that you are informed that you can do these things.

To protect your rights, we need to make restrictions that forbid

Unitex

- ▶ Université de Marne-la-Vallée
- ▶ à l'origine, version libre d'INTEX :
<http://mshe.univ-fcomte.fr/intex/>
- ▶ licence LGPL (code) et LGPL-LR (ressources linguistiques)
- ▶ Java pour l'interface, C++ dessous (efficacité)
- ▶ Unicode, support de nombreuses langues (Info / Preferences / Encoding : mettre UTF8)
- ▶ projet GramLab (2010 à 2013) : surcouche d'Unitex
- ▶ utilisé par :
 - ▶ de très nombreuses universités
 - ▶ des entreprises du TAL (Kwaga, CEA, Sinequa, Systran, Viavoo, ...)

À remarquer

INTEX a également donné naissance à NooJ, qui est maintenant libre

Caractéristiques linguistiques

- ▶ Unitex repose sur l'utilisation de données linguistiques
- ▶ dépendantes des langues
- ▶ trois types de données :
 - ▶ dictionnaires électroniques
 - ▶ grammaires locales
 - ▶ tables lexico-syntaxiques (lexique-grammaire)

Ressources

- ▶ site Internet officiel
<https://unitexgramlab.org/fr>
- ▶ site Internet de l'équipe TLN de l'Université de Tours <https://tln.lifat.univ-tours.fr/tln/version-francaise/navigation/ressources/tutoriels-unitex/>
- ▶ Manuel d'Utilisation (Paumier - Martineau 2006)
- ▶ ateliers réguliers
- ▶ liste de diffusion

À quoi ça sert ?

- ▶ recherche de motifs complexes dans des textes
 - ▶ concordance (visualisation des résultats en contexte)
 - ▶ annotation
 - ▶ analyse
- par la création de grammaires locales ou de transducteurs
- *via* une interface graphique

Exemple d'utilisation (1)

- ▶ rédaction de grammaires locales
- ▶ pour la recherche de motifs

Par exemple :

Loto à 19 h. Ouverture à 18 h

Exemple d'utilisation (1)

- ▶ rédaction de grammaires locales
- ▶ pour la recherche de motifs

Par exemple :

Loto à 19 h . Ouverture à 18 h

Exemple d'utilisation (2)

- ▶ développement de transducteurs
- ▶ pour l'annotation de textes

Par exemple :

Loto à 19 h. Ouverture à 18 h

Exemple d'utilisation (2)

- ▶ développement de transducteurs
- ▶ pour l'annotation de textes

Par exemple :

```
<EVENT eid="e0" eiid="ei0" class="OCCURENCE">Loto</EVENT > à  
<TIMEX3 tid="t1" type="TIME" value="T19:XX">19 h</TIMEX3>  
< TLINK lid="l1" relType="BEGUN_BY" eventInstanceID="ei1" relatedToTime="t1"/>.  
<EVENT eid="e1" eiid="ei1" class="OCCURENCE">Ouverture</EVENT >  
<TLINK lid="l1" relType="IDENTITY" eventInstanceID="ei1" relatedToTime="t1"/> à  
<TIMEX3 tid="t1" type="TIME" value="T18:XX" >18 h</TIMEX3>  
<TLINK lid="l1" relType="BEGUN_BY" eventInstanceID="ei1" relatedToTime="t1"/>
```

Ce qu'Unitex ne fait pas

- ▶ des traitements statistiques (\neq GATE) - à l'exception d'un module de désambiguïsation
- ▶ des traitements sur corpus (\neq texte), mais lancé en mode console. . .
- ▶ la désambiguïsation (par défaut)

Sources

Introduction

Prise en main d'Unitex

- Premier pas

- Un pas de côté

- Un pas plus loin

- Multilinguisme

Derrière le rideau

Pour finir

Manipulations de base sur Unix (1)

Premiers pas

- ▶ lancer Unix
- ▶ vérifier que la liste de langues est correcte
- ▶ ouvrir le *Tour du monde en 80 jours* (TDM) avec le prétraitement par défaut :
 - ▶ quels traitements ont été effectués ?
 - ▶ que signifie le S dans le texte ?

Manipulations de base sur Unix (2)

D'une langue à l'autre

- ▶ changer de langue, passer en allemand
- ▶ ouvrir le texte *KafkaProzess* avec le prétraitement par défaut :
 - ▶ quelles différences avec le français ?
- ▶ changer de langue, passer en thaï
- ▶ ouvrir le texte *SiPhanDin3* avec le prétraitement par défaut :
 - ▶ que constatez-vous ?

Manipulations de base sur Unix (3)

Compter avec Unix

Sur le fichier TDM, combien :

- ▶ de tokens ?
- ▶ de mots simples ?
- ▶ de locutions ?
- ▶ de mots inconnus ? (pourquoi n'ont-ils pas été reconnus ?)
- ▶ comment les visualiser ?

Gestion du multilinguisme

Les traitements sont tous **dépendants des langues** :

- ▶ avantages : précision, adaptation aux spécificités
- ▶ inconvénients : lourdeur, maintenance compliquée

(petit) exercice

- ▶ ouvrir l'alphabet du français
- ▶ que manque-t-il ? Comment est-ce géré ?

Sources

Introduction

Prise en main d'Unitex

Derrière le rideau

- Prétraitements

- Découpage en phrases

- Normalisations

- Découpage de base

- Dictionnaires

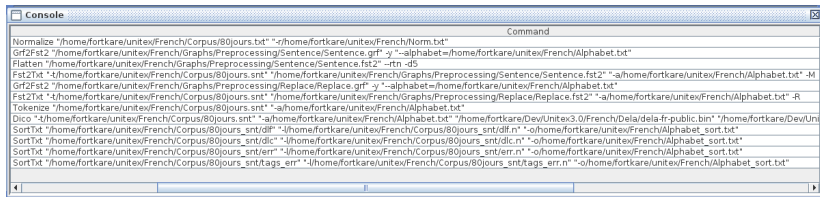
Pour finir

Prétraitements appliqués

1. découpage du texte en **phrases**
2. **normalisations** (\neq lemmatisation) : puisqu' \rightarrow puisque
3. découpage en **unités (lexicales)** (tokenisation)
4. application des **dictionnaires**
5. construction de l'**automate** du texte

La console Unitex

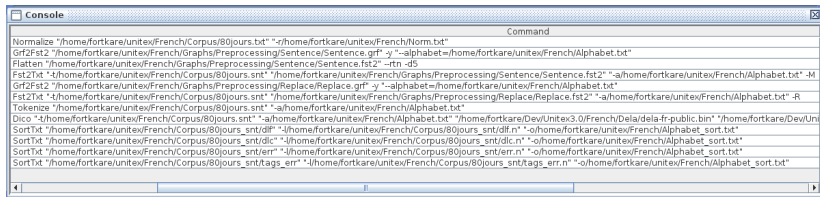
Menu Info / Console



```
Command
Normalize "/home/fortkare/unitex/French/Corpus/80jours.txt" -z/home/fortkare/unitex/French/Norm.txt
Grf2Fst2 "/home/fortkare/unitex/French/Graphs/Preprocessing/Sentence/Sentence.grf" -y "--alphabet=/home/fortkare/unitex/French/Alphabet.bt"
Flatten "/home/fortkare/unitex/French/Graphs/Preprocessing/Sentence/Sentence.fst2" --rtm -d5
Fst2Txt "-t/home/fortkare/unitex/French/Corpus/80jours.snt" "/home/fortkare/unitex/French/Graphs/Preprocessing/Sentence/Sentence.fst2" "-a/home/fortkare/unitex/French/Alphabet.txt" -M
Grf2Fst2 "/home/fortkare/unitex/French/Graphs/Preprocessing/Replace/Replace.grf" -y "--alphabet=/home/fortkare/unitex/French/Alphabet.txt"
Fst2Txt "-t/home/fortkare/unitex/French/Corpus/80jours.snt" "/home/fortkare/unitex/French/Graphs/Preprocessing/Replace/Replace.fst2" "-a/home/fortkare/unitex/French/Alphabet.txt" -R
Tokenize "/home/fortkare/unitex/French/Corpus/80jours.snt" "-a/home/fortkare/unitex/French/Alphabet.txt"
Dico "-t/home/fortkare/unitex/French/Corpus/80jours.snt" "-a/home/fortkare/unitex/French/Alphabet.txt" "/home/fortkare/Dev/Unitex3.0/French/Dela/dela-fr-public.bin" "/home/fortkare/Dev/Unitex3.0/French/Alphabet.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/dlc" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/dlc.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/dlc" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/dlc.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/err" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/err.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/tags_err" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/tags_err.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
```

La console Unitex

Menu Info / Console

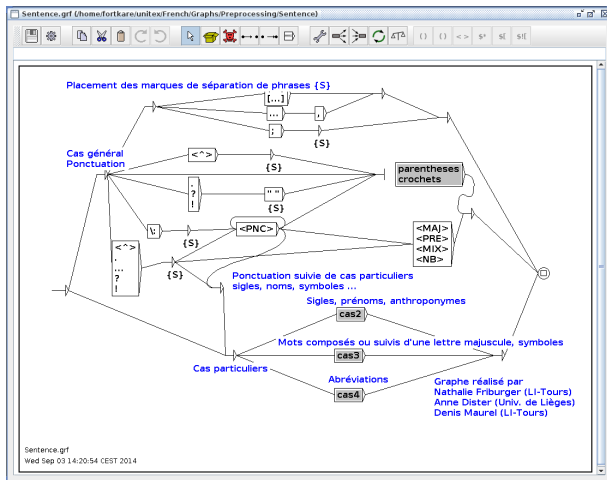


```
Command
Normalize "/home/fortkare/unitex/French/Corpus/80jours.txt" -x/home/fortkare/unitex/French/Norm.txt
Grf2fst2 "/home/fortkare/unitex/French/Graphs/Preprocessing/Sentence/Sentence.grf" -y "--alphabet=/home/fortkare/unitex/French/Alphabet.txt"
Flatten "/home/fortkare/unitex/French/Graphs/Preprocessing/Sentence/Sentence.fst2" --ftn -d5
Fst2Txt "t:/home/fortkare/unitex/French/Corpus/80jours_snt" "/home/fortkare/unitex/French/Graphs/Preprocessing/Sentence/Sentence.fst2" "-a/home/fortkare/unitex/French/Alphabet.txt" -M
Grf2fst2 "/home/fortkare/unitex/French/Graphs/Preprocessing/Replace/Replace.grf" -y "--alphabet=/home/fortkare/unitex/French/Alphabet.txt"
Fst2Txt "t:/home/fortkare/unitex/French/Corpus/80jours_snt" "/home/fortkare/unitex/French/Graphs/Preprocessing/Replace/Replace.fst2" "-a/home/fortkare/unitex/French/Alphabet.txt" -R
Tokenize "/home/fortkare/unitex/French/Corpus/80jours_snt" "-a/home/fortkare/unitex/French/Alphabet.txt"
Dico "t:/home/fortkare/unitex/French/Corpus/80jours_snt" "-a/home/fortkare/unitex/French/Alphabet.txt" "/home/fortkare/Dev/Unitex3.0/French/Dela/dela-fr-public.bin" "/home/fortkare/Dev/Unitex3.0/French/Dela/dela-fr-public.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/dif" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/dif.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/dlc" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/dlc.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/err" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/err.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
SortTxt "/home/fortkare/unitex/French/Corpus/80jours_snt/tags_err" "-l/home/fortkare/unitex/French/Corpus/80jours_snt/tags_err.n" "-o/home/fortkare/unitex/French/Alphabet_sort.txt"
```

- ▶ Normalize : remplace chaque séquence de séparateurs par un seul séparateur
- ▶ Grf2Fst2 : compile le(s) graphe(s) de la grammaire en .fst2
- ▶ Flatten : (essaye de) transforme(r) le .fst2 en transducteur
- ▶ Fst2Txt : applique un transducteur à un texte
- ▶ Tokenize : découpe le texte en unités (lexicales)
- ▶ Dico : applique des dictionnaires à un texte
- ▶ SortTxt : tri le texte selon le fichier paramètre

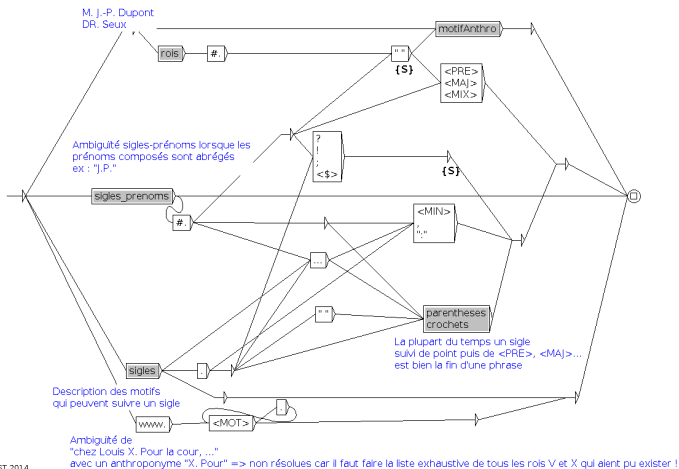
Découpage en phrases

Grf2Fst2 "French/Graphs/Preprocessing/Sentence/Sentence.grf"
-y "--alphabet=French/Alphabet.txt"



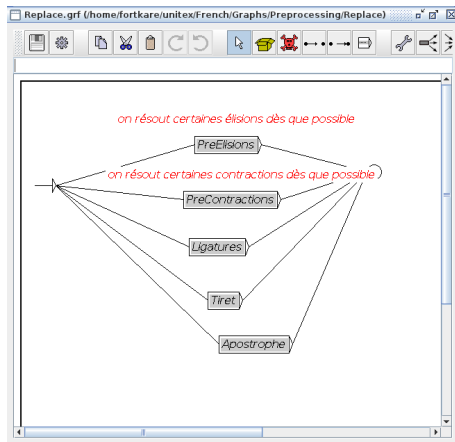
Découpage en phrases : à l'intérieur

Cas 2 : sigles, prénom, anthroponymes



Normalisations diverses

```
Grf2Fst2 "French/Graphs/Preprocessing/Replace/Replace.grf"  
-y "--alphabet=French/Alphabet.txt"
```



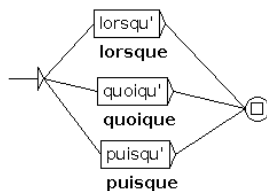
Normalisations diverses : à l'intérieur

Pré-élisions

c' ç' j' jusqu' m' n' presqu' qu' quelqu'
apparaissent dans des expressions figées

"t" est ambigu : "tu" ou "te" l' est ambigu

"s" est ambigu : "se" ou "si"



Découpage en unités

```
Tokenize "French/Corpus/80jours.snt"  
"-aFrench/Alphabet.txt"
```

Pour le français, une unité est :

- ▶ $\{S\}$
- ▶ une étiquette lexicale : ADV
- ▶ une séquence de lettres contiguës
- ▶ un (et un seul) caractère différent d'une lettre

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, such as "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

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

Mots simples vs mots composés

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 bleu, 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 (M2)

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 \rightarrow Flexion automatique \rightarrow DELAF

Traitement des dictionnaires

Compression automatique des dictionnaires (en transducteurs)

Avantages :

- ▶ taille mémoire
- ▶ accès à l'information

Sources

Introduction

Prise en main d'Unitex

Derrière le rideau

Pour finir

CQFR : Ce Qu'il Faut Retenir



Unitex est un outil :

- ▶ dépendant des langues
- ▶ qui permet
 - ▶ de faire des recherches de motifs
 - ▶ de visualiser les résultats et des stats de base
 - ▶ d'annoter
- ▶ qui traite des textes et non des corpus
- ▶ qui applique des prétraitements