

Extraction d'information

Cours 2

Nassim ZELLAL

Installation d'Unitex

- Pour installer Unitex : <http://unitexgramlab.org/fr>
- Sous Linux :
 - 1- téléchargez le fichier :
 - Unitex-GramLab-3.2-linux-i686.run (32 bits)
 - ou bien
Unitex-GramLab-3.2-linux-x86_64.run (64 bits)
 - 2-donnez lui les droits d'exécution par exemple :
 - `chmod +x Unitex-GramLab-3.2-linux-x86_64.run`
 - 3-lancez le fichier :
 - `./Unitex-GramLab-3.2-linux-x86_64.run`
 - 4-lancez le jar "Unitex.jar" se trouvant dans le dossier "Unitex-GramLab>App".
- Sous Windows :
 - Téléchargez l'exécutable Unitex-GramLab-3.2_win64-setup.exe (version 64 bits) ou Unitex-GramLab-3.2_win32-setup.exe (version 32 bits).
 - Ensuite, lancez l'exécutable à partir du raccourci sur votre bureau ou bien à partir du dossier "Unitex-GramLab>App", en y ouvrant une invite de commandes puis tapez: `java -jar Unitex.jar`

Après l'installation d'Unitex

- Nous obtenons deux emplacements :
- Le premier emplacement contient les ressources du système, qui sont classées par langue (Arabic, Chinese, English, French, etc.) ainsi que le dossier « App ».
- Dans ce dossier: C:\.....\Unitex-GramLab**App**, il y a, entre autres, l'archive JAR exécutable « Unitex.jar » dont nous aurons besoin pour lancer Unitex.
- Le deuxième emplacement, C:\.....\Unitex**French**
- contient les ressources de l'utilisateur, comprenant, entre autres, les corpus à traiter (dossier « Corpus »), les dictionnaires DELA (dossier « Dela »), les graphes (grammaires) d'extraction (dossier « Graphs ») et les graphes flexionnels (dossier « Inflection »).

Emplacement des ressources linguistiques du système

Unitex-GramLab

C:\Users\user\Desktop\Unitex-GramLab

Fichier Edition Affichage Outils ?

Organiser Ouvrir Inclure dans la bibliothèque Partager avec Graver Nouveau dossier

Favoris

- Bureau
- Emplacements récents
- Téléchargements
- Logiciel

Bibliothèques

- Documents
- Images
- Musique
- Vidéos

Ordinateur

- Disque local (C:)
- Disque local (D:)

Réseau

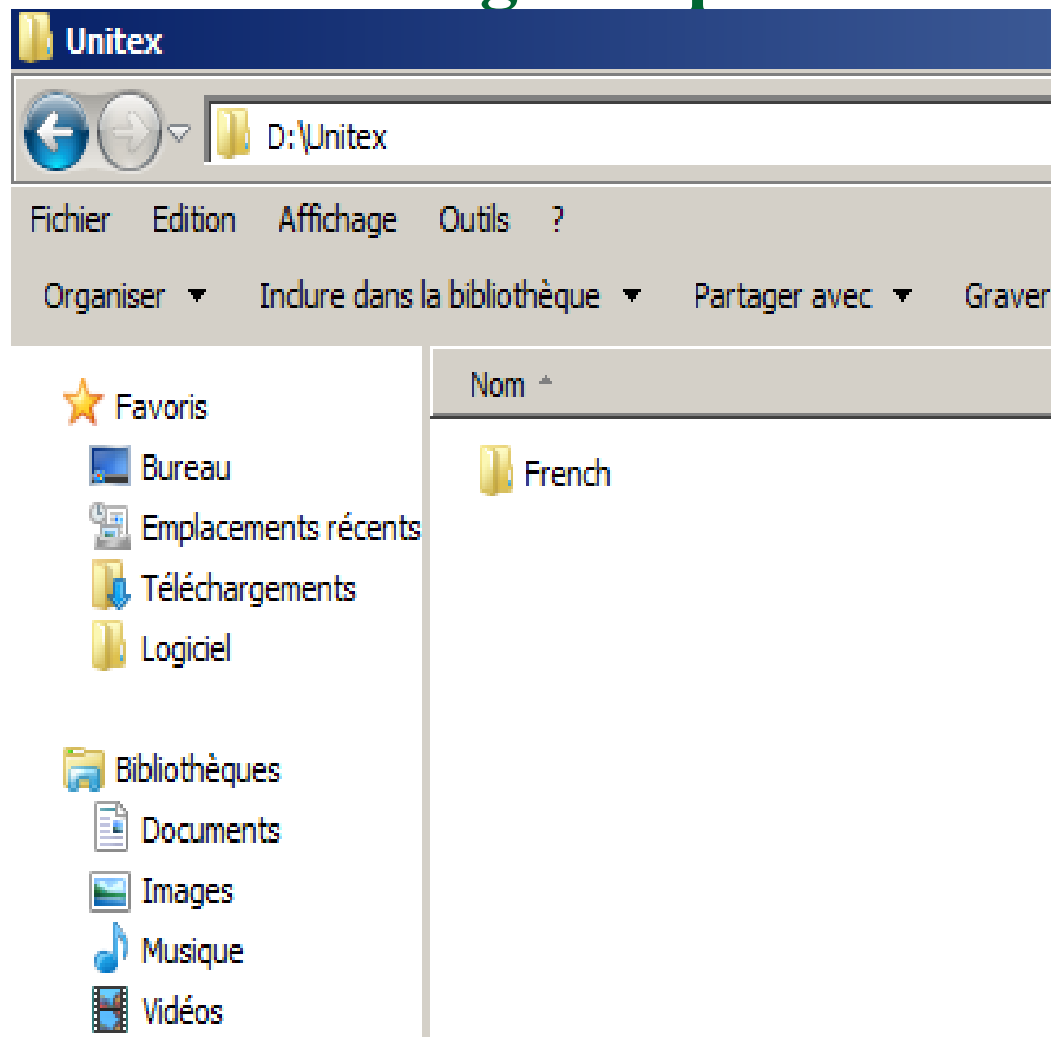
Nom ^	Modifié le	Type	Taille
App	24/12/2020 10:52	Dossier de fichiers	
Arabic	24/12/2020 10:51	Dossier de fichiers	
Chinese	24/12/2020 10:51	Dossier de fichiers	
English	24/12/2020 10:51	Dossier de fichiers	
Finnish	24/12/2020 10:51	Dossier de fichiers	
French	24/12/2020 10:52	Dossier de fichiers	
Georgian (Ancient)	24/12/2020 10:52	Dossier de fichiers	
German	24/12/2020 10:52	Dossier de fichiers	
Greek (Ancient)	24/12/2020 10:52	Dossier de fichiers	
Greek (Modern)	24/12/2020 10:52	Dossier de fichiers	
Italian	24/12/2020 10:52	Dossier de fichiers	
Korean	24/12/2020 10:52	Dossier de fichiers	
Latin	24/12/2020 10:52	Dossier de fichiers	
Malagasy	24/12/2020 10:52	Dossier de fichiers	
Norwegian (Bokmal)	24/12/2020 10:52	Dossier de fichiers	
Norwegian (Nynorsk)	24/12/2020 10:52	Dossier de fichiers	
Polish	24/12/2020 10:52	Dossier de fichiers	

Emplacement de l'archive JAR exécutable d'Unitex

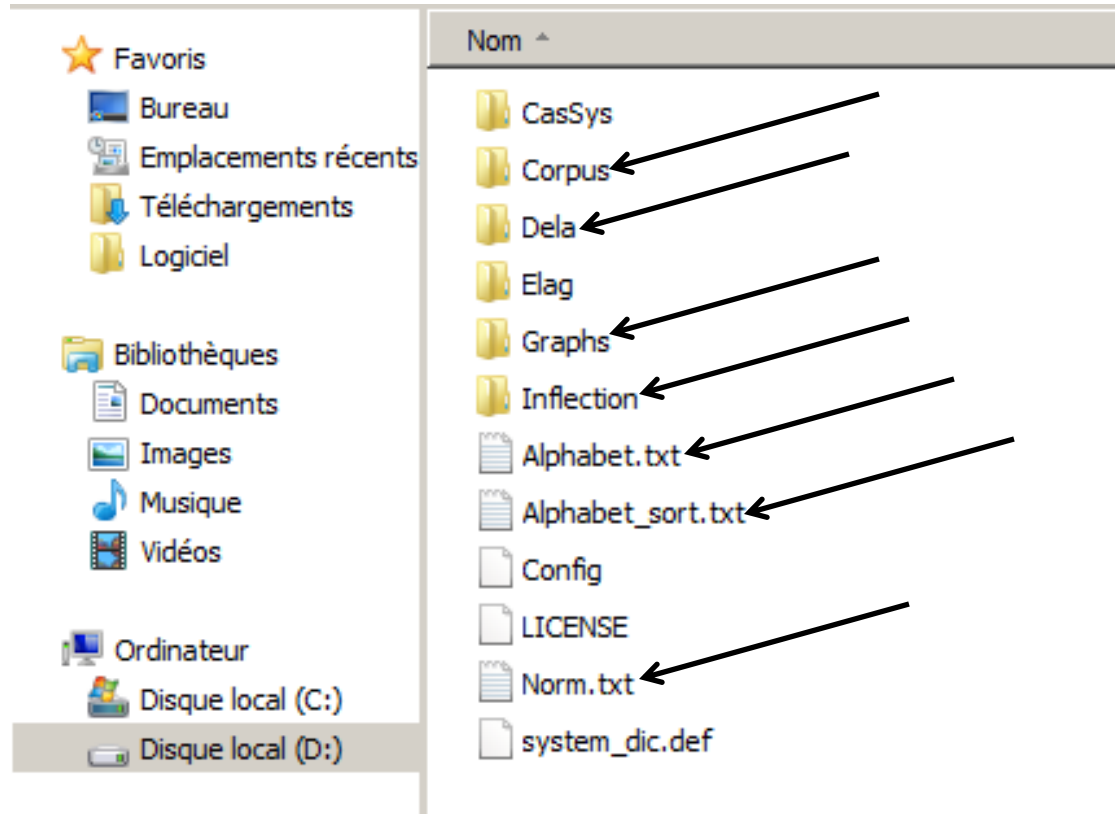
The screenshot shows a Windows File Explorer window titled 'App'. The address bar indicates the path is 'Unitex-GramLab > App'. The left sidebar shows the 'Favoris' (Favorites) section with 'Bureau', 'Emplacements récents', 'Téléchargements', and 'Logiciel'. The main pane displays a list of files and folders with columns for 'Nom', 'Modifié le', 'Type', and 'Taille'. A black arrow points to the 'Unitex.jar' file, which is an 'Executable Jar File' of 1 343 Ko.

Nom	Modifié le	Type ^	Taille
assembly	24/12/2020 10:51	Dossier de fichiers	
disclaimers	24/12/2020 10:51	Dossier de fichiers	
lib	24/12/2020 10:51	Dossier de fichiers	
licenses	24/12/2020 10:51	Dossier de fichiers	
manual	24/12/2020 10:52	Dossier de fichiers	
UnitexToolLogger.exe	17/06/2020 06:01	Application	2 962 Ko
Unitex.jar	14/01/2020 02:14	Executable Jar File	1 343 Ko
pom.xml	14/01/2020 20:03	Fichier XML	2 Ko
Unitex.ico	14/01/2020 02:14	Icône	4 Ko

Le répertoire « French » - ressources linguistiques de l'utilisateur



Le répertoire « French » - ressources linguistiques de l'utilisateur



Démarrer UNITEX

App

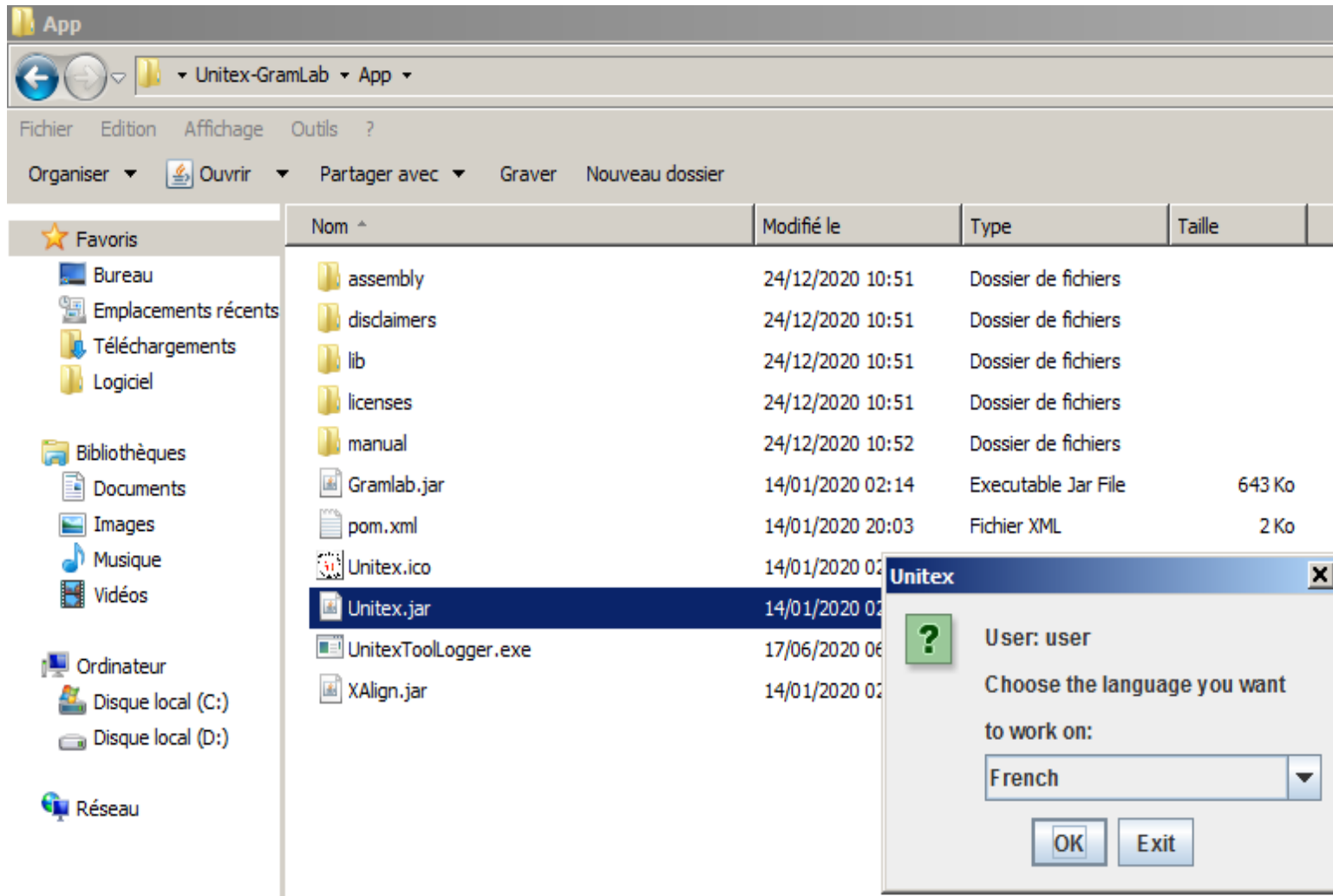
Unitex-GramLab App

Fichier Edition Affichage Outils ?


Organiser Ouvrir Partager avec Graver Nouveau dossier

	Nom	Modifié le	Type ^	Taille
★ Favoris				
Bureau	assembly	24/12/2020 10:51	Dossier de fichiers	
Emplacements récents	disclaimers	24/12/2020 10:51	Dossier de fichiers	
Téléchargements	lib	24/12/2020 10:51	Dossier de fichiers	
Logiciel	licenses	24/12/2020 10:51	Dossier de fichiers	
	manual	24/12/2020 10:52	Dossier de fichiers	
Bibliothèques	UnitexToolLogger.exe	17/06/2020 06:01	Application	2 962 Ko
Documents	Unitex.jar	14/01/2020 02:14	Executable Jar File	1 343 Ko
Images	pom.xml	14/01/2020 20:03	Fichier XML	2 Ko
Musique	Unitex.ico	14/01/2020 02:14	Icône	4 Ko
Vidéos				

Emplacement de l'archive JAR exécutable d'Unitex

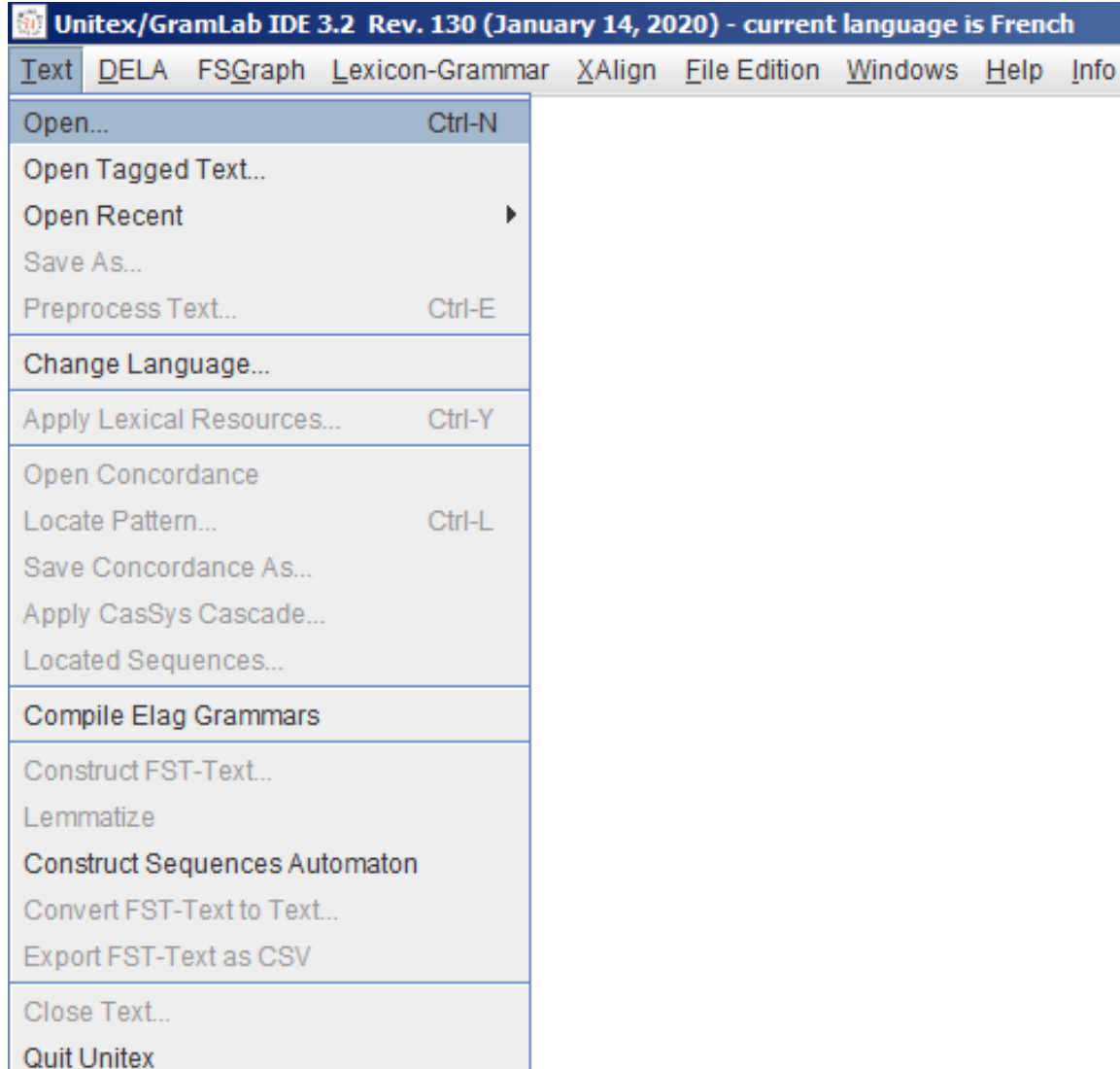


UNITEX/Gramlab IDE 3.2

 Unitex/GramLab IDE 3.2 Rev. 130 (January 14, 2020) - current language is French

Text DELA FSGraph Lexicon-Grammar XAlign File Edition Windows Help Inf

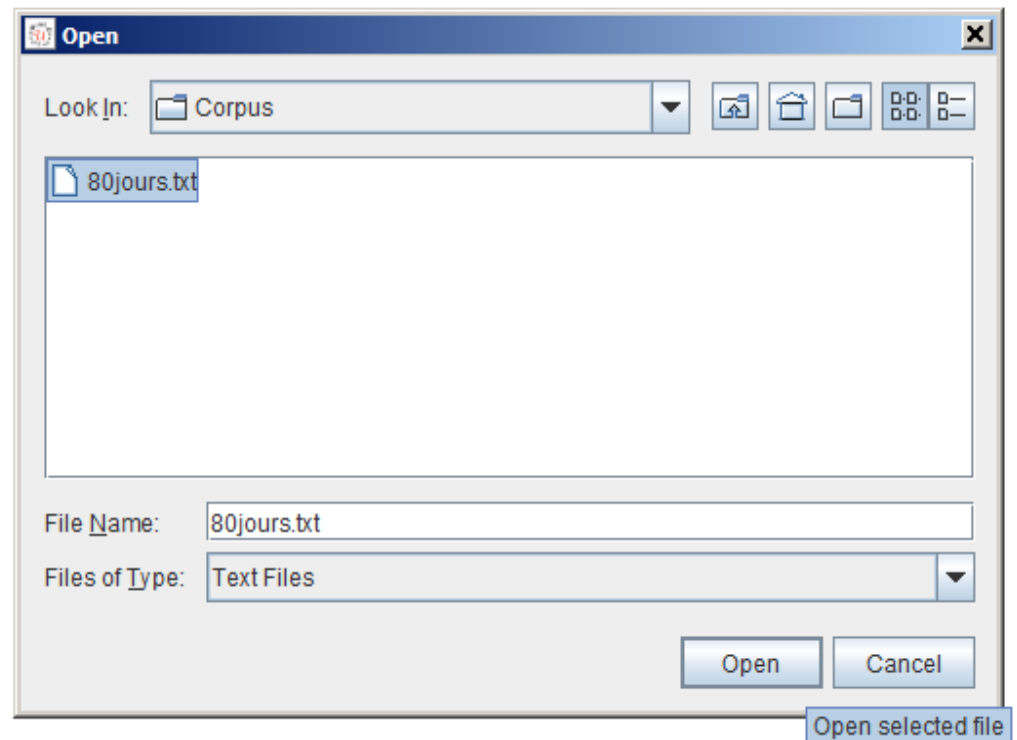
Ouvrir le texte « 80jours.txt »



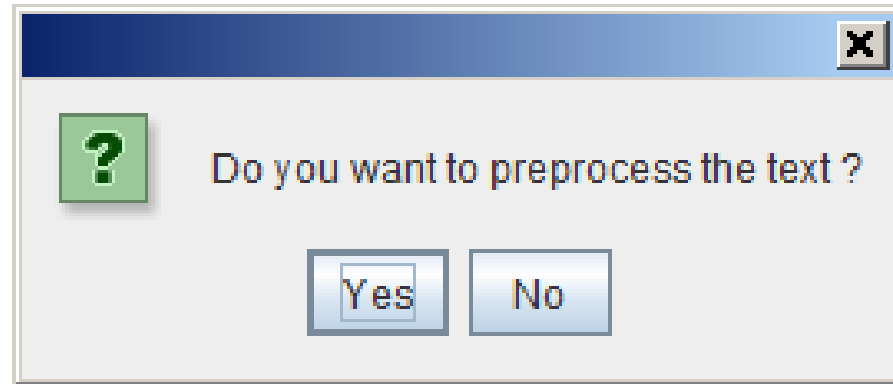
Ouvrir le texte « 80jours.txt »

Unitex/GramLab IDE 3.2 Rev. 130 (January 14, 2020) - current language is French

Text DELA FSGraph Lexicon-Grammar XAlign File Edition Windows Help Info



Prétraitement du texte « 80jours.txt »

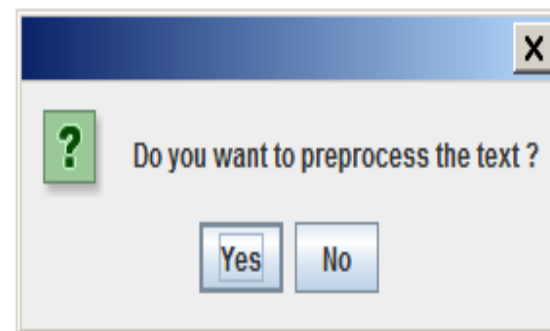


Prétraitement du texte « 80jours.txt »

Text DELA FSGraph Lexicon-Grammar XAlign File Edition Windows Help Info

Unitex propose à l'utilisateur de prétraiter le texte : découpage (segmentation) de texte en phrases, découpage (tokenisation) de texte en unités lexicales (Token list sur Unitex), normalisation de séparateurs et de certains caractères comme les accolades { } via la ressource de l'utilisateur « Norm.txt » et application des dictionnaires par défaut d'Unitex (ressources du système) qui se trouve dans l'emplacement : **C:\.....\Unitex-GramLab\French\Dela.**

Ces dictionnaires par défaut sont : **Dela fr.bin** et **motsGramf-.bin**.



Le fichier « Norm.txt »

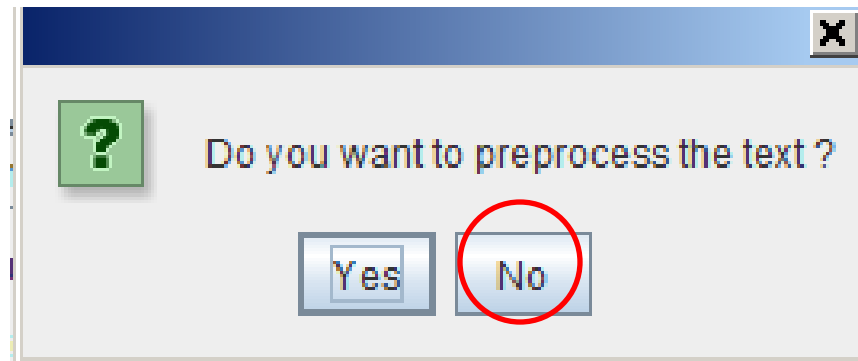
```
1 | —→ [ # ] CR LF
2 | | —→ [ # # # CR LF
3 { —→ [ [ ] ] CR LF
4 } —→ ( ( ) )
```

Le fichier « Norm.txt » se trouve dans le dossier « French » de l'utilisateur (C:\.....\Unitex\French).

C'est une ressource servant à normaliser certains caractères. Ce fichier peut être enrichi par l'utilisateur.

Prétraitement du texte « 80jours.txt »

- Si on ne souhaite pas effectuer le prétraitement, le texte sera quand même tokenisé et normalisé (séparateurs + certains caractères via la ressource de l'utilisateur « Norm.txt »).



Avant le phase de tokenisation et de normalisation du texte « 80jours.txt »

```
1 Chapitre · I · CRLF
```

2 DANS LEQUEL PHILEAS → → → → → → FOGG ET PASSEPARTOUT S'ACCEPTENT RÉCIPROQUEMENT L'UN COMME MAÎTRE, L'AUTRE COMME DOMESTIQUE CRLE

3 En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens - maison dans laquelle Sheridan mourut en 1814 -, était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarquables du Reform-Club de Londres, bien qu'il semblât prendre à tâche de ne rien faire qui pût { attirer } l'attention. Le | cœur des hommes | |. **CR LF**

4 CRLF

5 CRLF

6 CRLF

7 CRLF

8 CRLF

9 A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Phileas Fogg, personnage énigmatique, dont on ne savait rien, sinon que c'était un fort galant homme et l'un des plus beaux gentlemen de la haute société anglaise. CRLE

10 On disait qu'il ressemblait à Byron -- par la tête, car il était irréprochable quant aux pieds --, mais un Byron à moustaches et à favoris, Byron impassible, qui aurait vécu mille ans sans vieillir. **CRLE**

Résultat après la phase de tokenisation et de normalisation du texte « 80jours.txt »

```
80jours.snt (D:\Unitex\French\Corpus)
0 sentence delimiter, 161541 (9462 diff) tokens, 71826 (9432) simple forms, 438 (10) digits

Chapitre I
DANS LEQUEL PHILEAS FOGG ET PASSEPARTOUT S'ACCEPTENT RÉCIPROQUEMENT L'UN COMME MAÎTRE,
L'AUTRE COMME DOMESTIQUE
En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens - maison
dans laquelle Sheridan mourut en 1814 - , était habitée par Phileas Fogg, esq., l'un des
membres les plus singuliers et les plus remarqués du Reform-Club de Londres, bien qu'il
semblât prendre à tâche de ne rien faire qui pût [[]] attirer (()) l'attention. Le [#] cœur
des hommes [###.
A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Phileas Fogg,
personnage énigmatique, dont on ne savait rien, sinon que c'était un fort galant homme et
l'un des plus beaux gentlemen de la haute société anglaise.
```

Résultat après la phase de tokenisation et de normalisation du texte « 80jours.txt »

The screenshot displays a software interface with three main windows. The top window, titled '80jours.snt (D:\Unitex\French\Corpus)', shows the processed text of 'Chapitre I' from '80 Jours'. The text is tokenized, with punctuation and special characters enclosed in brackets, such as '[[]]' for ']] attirer ([)]' and '[#]' for 'Le cœur'. The status bar indicates '0 sentence delimiter, 161541 (9462 diff) tokens, 71826 (9432) simple forms, 438 (10) digits'.

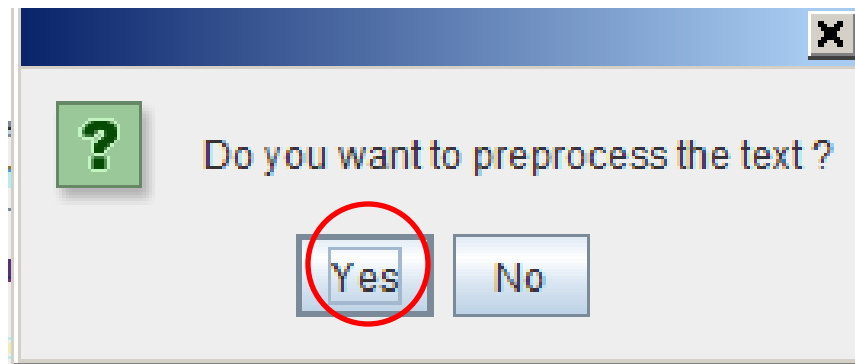
The bottom-left window, titled 'Token list', shows a list of tokens sorted by frequency. The top tokens are 'de', 'à', 'le', 'la', 'et', 'l', 'il', 'les', 'un', 'en', 'du', 'd', 'que', and 'Fogg'.

The bottom-right window, titled 'Word Lists in D:\Unitex\French\Corpus\80jours_snt', contains three panels: 'DLF: simple-word lexical entries', 'DLC: compound lexical entries', and 'ERR: unknown simple words'. All three panels display the message 'This file is empty.'.

Count	Token
69885	
6940	,
4567	.
3806	'
2798	de
2039	-
1666	à
1608	le
1488	la
1466	et
1137	l
1106	il
948	les
824	un
784	en
753	du
726	d
717	"
655	que
655	Fogg

Prétraitement du texte « 80jours.txt »

- Si on souhaite effectuer ce prétraitement, plusieurs fonctionnalités de prétraitement de texte seront proposées.



Prétraitement du texte « 80jours.txt »

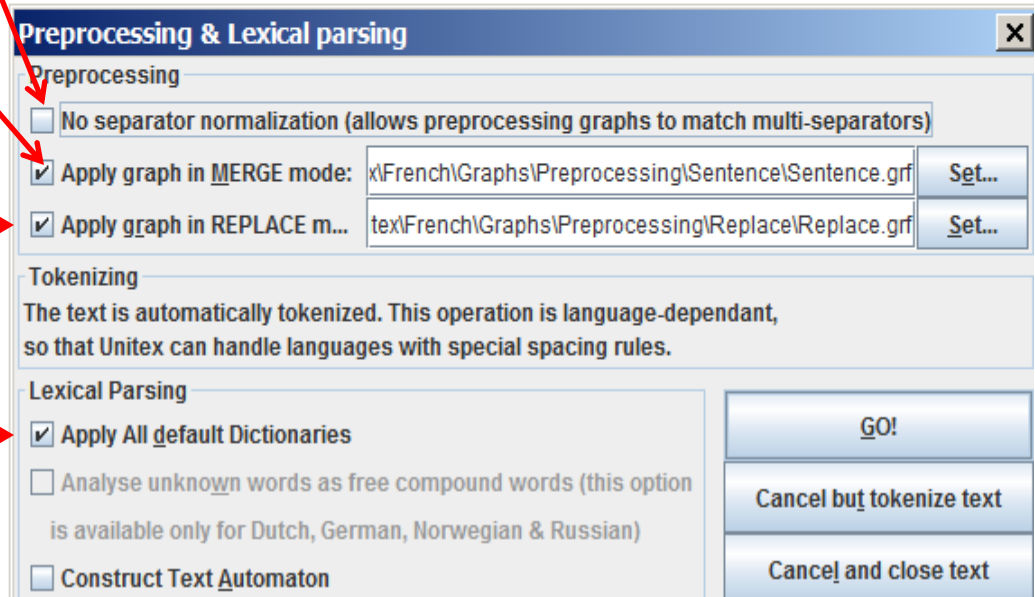
Text DELA FSGraph Lexicon-Grammar XAlign File Edition Windows Help Info

Empêcher la normalisation des séparateurs.

Découpage d'un texte en phrases.

Normalisation de certaines formes, e.g. cœur (cette forme n'est pas présente dans le DELAF du système) → coeur

Ces dictionnaires (ressources du système) se trouvent dans le répertoire : C:\.....\Unitex-GramLab\French\Dela



Informations obtenues après le prétraitement du texte « 80jours.txt »

80jours.snt (D:\Unitex\French\Corpus)

3653 sentence delimiters, 165268 (9453 diff) tokens, 71863 (9422) simple forms, 438 (10) digits
67468 occurrences (12178 DLF entries) simple words, 1617 occurrences (1544 DLC entries) compound words, 4233 occurrences (477 ERR li...

Chapitre I
{S}DANS LEQUEL PHILEAS FOGG ET PASSEPARTOUT S'ACCEPTENT RÉCIPROQUEMENT L'UN COMME MAÎTRE,
L'AUTRE COMME DOMESTIQUE
{S}En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens _ maison
dans laquelle Sheridan mourut en 1814 _ , était habitée par Phileas Fogg, esq., l'un des
membres les plus singuliers et les plus remarquables du Reform-Club de Londres, bien qu'il
semblât prendre à tâche de ne rien faire qui pût [[]] attirer (()) l'attention.{S} Le [#]
coeur des hommes [###.
{S}A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Phileas Fogg,
personnage énigmatique, dont on ne savait rien, sinon que c'était un fort galant homme et
l'un des plus beaux gentlemen de la haute société anglaise.

Token list

By Frequency By Char Order

69932
6940 ,
4567 .
3797 '
3653 {S}
2807 de
1695 à
1608 le
1488 la
1466 et
1161 -
1137 l
1106 il
948 les
878 -
824 un
784 en
753 du
726 d
717 "

Word Lists in D:\Unitex\French\Corpus\80jours_snt

DLF: 12178 simple-word lexical entries

a, .N+z1:ms:mp
à, .PREP+z1
a, avoir.V+z1:P3s
abaissait, abaisser.V+z1:I3
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:

DLC: 1544 compound lexical entries

à base de, .PREP+z1
à bon droit, .ADV+z1
à bord d, à bord de, .PREP+z1
à bord de, .PREP+z1
à bord des, à bord de, .PREP+z1
à bord du, à bord de, .PREP+z1
à califourchon sur, .PREP+z1

ERR: 477 unknown simple words

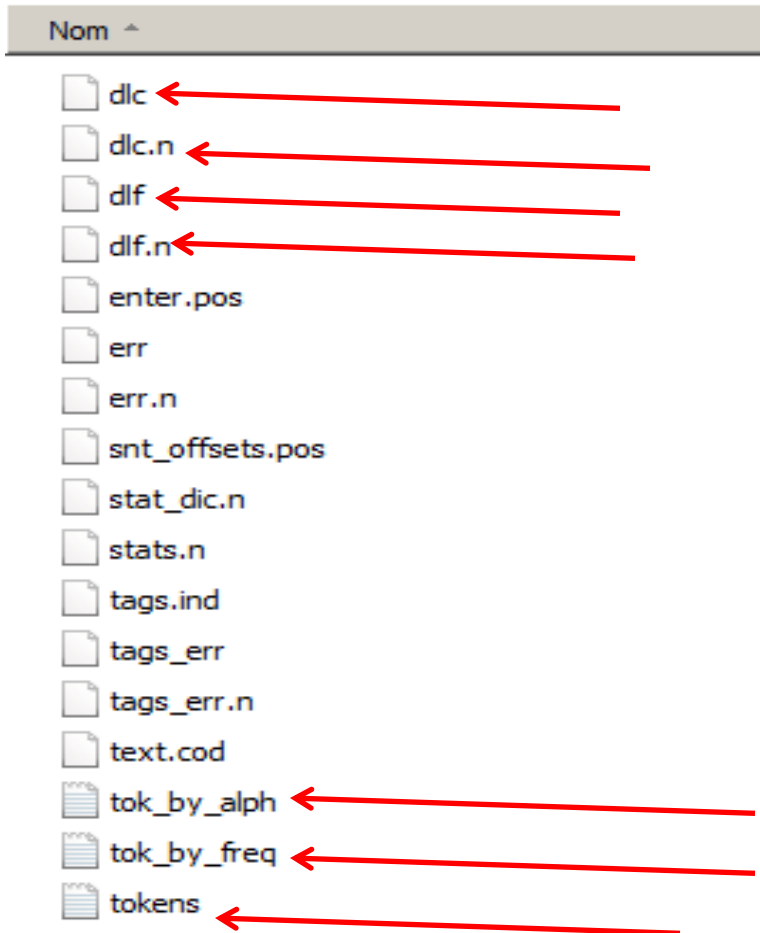
☐ Filter unknown words with tags.ind

Abraham
Aden
Afrique
Agra
Ahméhnagara
Alabama
Albermale
Allahabad
Allemagne
American
and
Andaman
Andrew
Angelica
Angleterre
Annam
Aouda
Arkansas
Armonica

Token list

Word list

Fichiers générés après le prétraitement



dlc : entrées lexicales composées.

dlf : entrées lexicales simples.

dlc.n : nombre d'entrées lexicales composées.

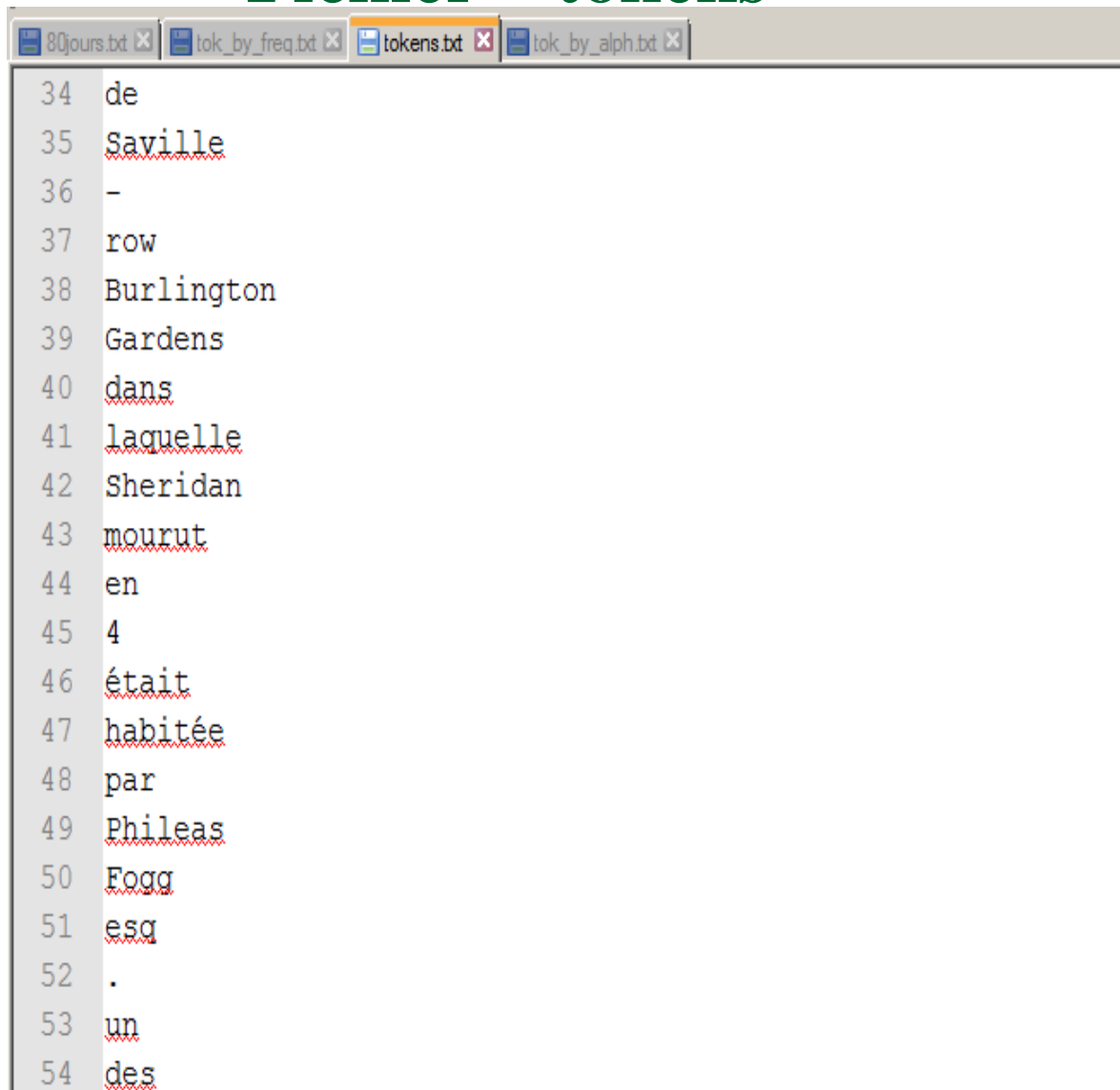
dlf.n : nombre d'entrées lexicales simples.

tok_by_freq : liste des unités lexicales triée par ordre de fréquence.

tok_by_alph : liste des unités lexicales triée par ordre alphabétique.

tokens : liste des tokens dans l'ordre d'apparition dans le texte.

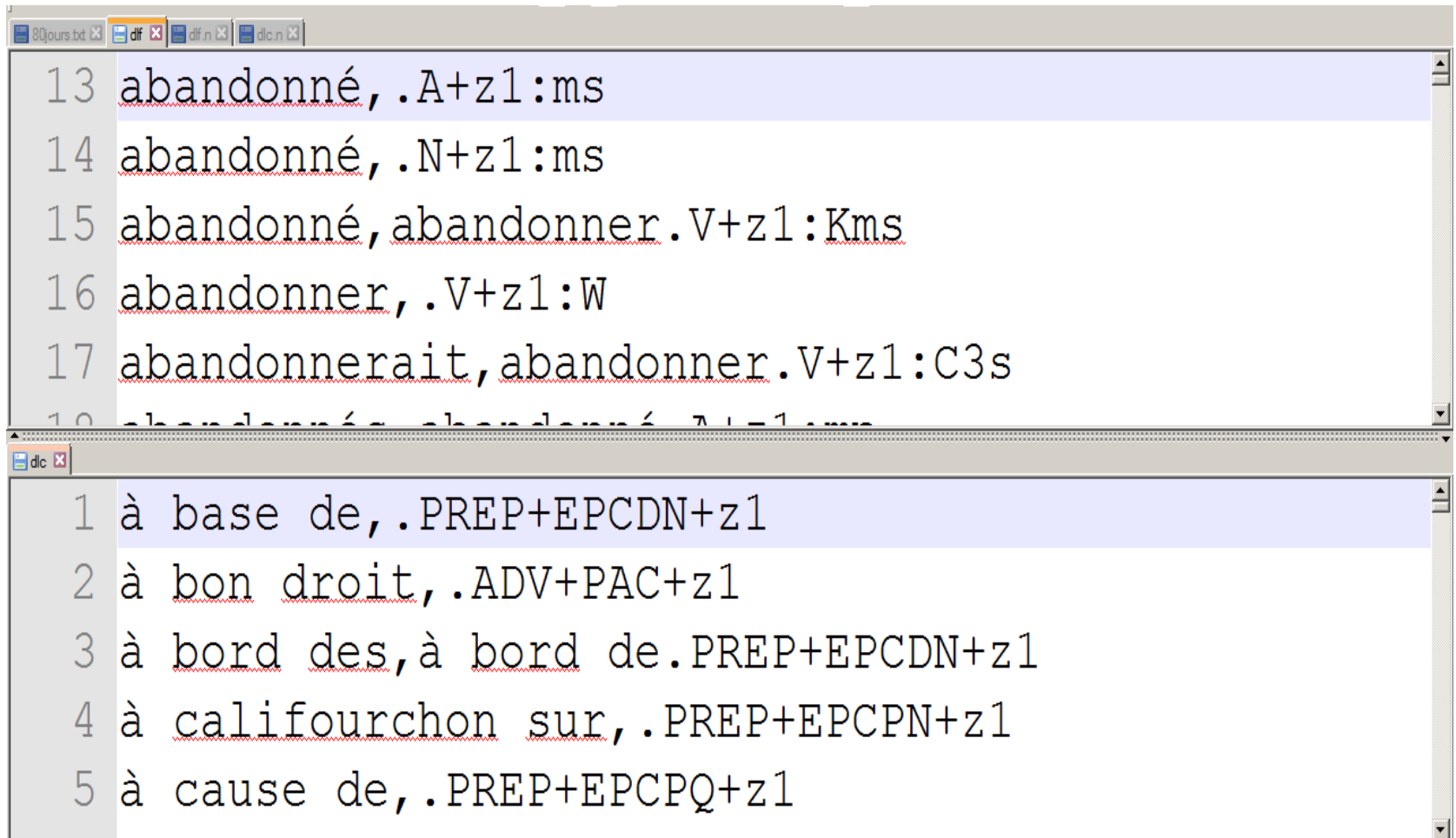
Fichier « tokens »



The screenshot shows a text editor window with four tabs: '80jours.bt', 'tok_by_freq.bt', 'tokens.bt' (which is the active tab), and 'tok_by_alph.bt'. The active tab displays a list of tokens, each preceded by a line number from 34 to 54. The tokens are: 'de', 'Saville', '-', 'row', 'Burlington', 'Gardens', 'dans', 'laquelle', 'Sheridan', 'mourut', 'en', '4', 'était', 'habitée', 'par', 'Phileas', 'Fogg', 'esq', '.', 'un', and 'des'. The words 'Saville', 'dans', 'laquelle', 'mourut', 'était', 'habitée', 'Fogg', 'esq', 'un', and 'des' are underlined with red dashed lines.

```
34 de
35 Saville
36 -
37 row
38 Burlington
39 Gardens
40 dans
41 laquelle
42 Sheridan
43 mourut
44 en
45 4
46 était
47 habitée
48 par
49 Phileas
50 Fogg
51 esq
52 .
53 un
54 des
```


Fichiers générés après l'application des dictionnaires du système



The image shows a screenshot of a text editor window with two tabs. The top tab, labeled '80jours.txt', contains a list of linguistic entries numbered 13 to 18. The bottom tab, labeled 'dic', contains a list of entries numbered 1 to 5. Each entry consists of a French phrase followed by a morphological analysis in a specific notation.

```
13 abandonné, .A+z1:ms
14 abandonné, .N+z1:ms
15 abandonné, abandonner.V+z1:Kms
16 abandonner, .V+z1:W
17 abandonnerait, abandonner.V+z1:C3s
18 abandonné, abandonné, .N+z1:ms

1 à base de, .PREP+EPCDN+z1
2 à bon droit, .ADV+PAC+z1
3 à bord des, à bord de.PREP+EPCDN+z1
4 à califourchon sur, .PREP+EPCPN+z1
5 à cause de, .PREP+EPCPQ+z1
```

Exercice 1

- Utiliser UNITEX pour remplacer dans le texte « 80jours.txt » tous les tirets - par un \$.
 - Utiliser UNITEX pour connaître le nombre de phrases dans ce texte.
 - Écrire un script Python permettant de connaître le nombre de phrases de ce texte, à partir du résultat généré par UNITEX.
-

Exercice 2

- Écrire un script Python permettant de découper le texte suivant en phrases.

```
1 Ce Phileas Fogg était-il riche ? Incontestablement. Mais comment il
  avait fait fortune, c'est ce que les mieux informés ne pouvaient
  dire, et Mr. Fogg était le dernier à lequel il convint de
  s'adresser pour l'apprendre. En tout cas, il n'était prodigue de
  rien, mais non avare, car partout où il manquait un appoint pour
  une chose noble, utile ou généreuse, il l'apportait silencieusement
  et même anonymement.
2 Anglais, à coup sûr, Phileas Fogg n'était peut-être pas Londonner.
  On ne l'avait jamais vu ni à la Bourse, ni à la Banque, ni dans
  aucun des comptoirs de la Cité.
3 _ Monsieur Fogg, répondit Mrs. Ada, c'est à moi de vous faire cette
  question. Vous étiez ruiné, vous voici riche...
4 _ Pardonnez-moi, madame, cette fortune vous appartient. Si vous
  n'aviez pas eu la pensée de ce mariage, mon domestique ne serait
  pas allé chez le révérend Samuel Wilson, je n'aurais pas été averti
  de mon erreur, et...
5 _ Cher monsieur Fogg..., dit la jeune femme.
6 _ Chère Ada... , répondit Phileas Fogg.
7 On comprend bien que le mariage se fit quarante-huit heures plus
  tard, et Passepartout, superbe, resplendissant, éblouissant, y
  figura comme témoin de la jeune femme. Ne l'avait-il pas sauvée, et
  ne lui devait-on pas cet honneur ?
```