

Unitex Getting Started

Denis Maurel, Cvetana Krstev

▶ To cite this version:

Denis Maurel, Cvetana Krstev. Unitex Getting Started. Independently published, 2024, 9798338757444. hal-04769290

HAL Id: hal-04769290 https://hal.science/hal-04769290v1

Submitted on 6 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



Language Resources and Technologies Society JeRTeh and University of Belgrade

Summary

Chap	oter 1: introduction	1
1 2 3	A short history A quick overview of the book Organization of correct files	1 1 2
5	Acknowledgement	6
Chap	ter 2: first steps	7
1 2 3	Discovering Unitex A detailed example The .snt file	7 16 35
Chap	ter 3: corpus annotation	45
1 2 3	First example: compound verbs using the verb <i>will</i> Second example: numbers written with words Third example: annotating Roman numerals	45 59 63
Chap	oter 4: CasSys	71
1 2 3 4	First example: named entity recognition in a plain text Second example: named entity recognition in an XML text Third example: measure recognition Additional possibilities (only for confident users)	71 84 88 94
Chap	ter 5: dictionary creation	107
1 2 3 4 5 6	Introduction Inflection of monolexical words Inflection of multi-word units Some additional remarks Dictionary graphs. Additional possibilities (only for confident users)	107 107 118 123 128 132
Chap	oter 6: other tools	137
1 2	Lexicon-grammar XAlign	137 141
Chap	oter 7: scripts	149
1 2 3 4 5	Introduction Creation of the <i>linguistic package</i> The command line An inventory of tag occurrences Two small remarks on the optimization of a cascade	149 149 157 158 164

1 A short history

Before introducing the Unitex software, we must go back in time and talk about the LADL, Laboratoire d'automatique documentaire et linguistique, established in 1968 by the late Professor Maurice Gross.¹

Maurice Gross introduced the notion of *lexicon-grammar*: he asserted that grammar could not be formalized without its dependence on the lexicon.² In collaboration with Professor Dominique Perrin,³ he proposed the use of the theory of finite state automata (FSA) for some aspects of linguistic description.⁴ At the same time, he had entrusted one of his doctoral students, Max Silberztein, with the task of creating the computer tool allowing such use of FSA. As part of his thesis, this student designed software called Intex.⁵

Another doctoral student of Maurice Gross,⁶ Sébastien Paumier, wanted to test in his thesis⁷ a different algorithm⁸ from the one implemented in Intex. As was customary in the nineties, the license to use Intex was not free and its code was not open. Therefore Sébastien Paumier had the idea of creating the free and open software, Unitex, based on this new algorithm, in addition to the features already developed by Max Silberztein.

Since then, Unitex has been enhanced by additions and improvements made by many different developers and users. The main contributors are listed in the Unitex user manual.⁹

2 A quick overview of the book

The aim of this book is to help its readers to get started with Unitex. In this, it differs from the Unitex user manual whose objective is to list the different possibilities of use that Unitex

¹ <u>https://en.wikipedia.org/wiki/Maurice_Gross</u>

² Gross M. (1975), *Méthodes en syntaxe*, Hermann, Paris.

³ <u>https://en.wikipedia.org/wiki/Dominique Perrin</u>

⁴ Gross M., Perrin D. (1989), *Electronic Dictionaries and Automata in Computational Linguistics*, LNCS 377.

⁵ Silberztein M. (1989), *Dictionnaires électroniques et reconnaissance lexicale automatique*, Thèse de doctorat (Université Paris VII).

⁶ Unfortunately, Maurice Gross passed away before the defense of the Sébastien Paumier's thesis.

⁷ Paumier S. (2003), *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*, Thèse de doctorat (Université de Marne-la-Vallée).

⁸ Paumier S. (2003), A time-efficient token representation for parsers, *FSMNLP 2003*, Budapest.

⁹ The manual is automatically integrated when downloading Unitex (see Chapter 2, Section 1.1.2, page 7). It is also available at the address: <u>https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf</u>

offers. The objective of this work is an educational approach to Unitex. It covers most of its possibilities, although not all of them. Namely:

```
Chapter 2 (first steps), page 7:
```

Installing and starting to work with Unitex; use of dictionaries; creation of simple graphs.

Chapter 3 (corpus annotation), page 45:

Creation of complex graphs with output; use of weights; morphological mode.

Chapter 4 (CasSys), page 71:

Creation of cascades of graphs; generalization graphs, negative right context, use of variables; testing variables.

Chapter 5 (dictionary creation), page 107:

Creation of dictionaries; inflectional graphs; dictionary graphs.

Chapter 6 (other tools), page 137:

Creation of graphs based on a lexicon-grammar; aligning texts.

<u>Chapter 7</u> (*scripts*), page 149: Creation of scripts.

This book proposes many exercises and micro-projects that can be implemented step by step by following the text.¹⁰ The different chapters are relatively independent. Screenshots are taken from the Unitex Windows interface. Some minor differences are possible with Mac or Linux interfaces.

3 Organization of correct files

The book comes with a zipped file (*MyUnitex.zip*) including all correct files (graphs, cascades, dictionaries, scripts and so on) answering the exercises. If the reader does not wish to construct her/his own graphs or wishes to verify them, she/he may consult them. The early chapters also include videos detailing the step-by-step construction of the graphs.



¹⁰ Work with Elag Grammars and text automata (or FST) is not covered.

Since these files often have the same name in one section and another, they are stored in folders named by the number of the section. Also, before consulting them, the reader must copy an entire section folder and paste it into the folder hierarchy just above.

For example, the first graph built is in the folder named:

MyUnitex\English\Graphs\UnitexGettingStarted\Will\3.1.1

with an illustrative video. The reader must copy this graph and paste it in the folder:

MyUnitex\English\Graphs\UnitexGettingStarted\Will

before opening it with Unitex.

4 A workbook?

For a user already familiar with Unitex, this book can be used differently, as a book of correct examples.

4.1 Graph exercises

Before starting graph exercises Open the text:*MyUnitex**English**Corpus**UnitexGettingStarted**DombeyAndSon**DombeyAndSon.txt*

Exercise 1

Create a graph that recognizes all forms of the verb *meet*, including verbs with auxiliaries and modal auxiliaries, also including negative and interrogative forms.

The correct solution is in Chapter 2, Section 2.3.6, page 28.

Exercise 2

Create a graph that recognizes and annotates compound verbs using the verbs *will* and *shall*. Use *Verb*, *Adverb* and *Pronoun* tags. *Verb* tags can have an attribute *type* for annotation.¹¹

The correct solution is in <u>Chapter 3, Section 1.9.3</u>, page 58.

Exercise 3

Create a graph that recognizes and annotates numbers written with words, from 2 to 999 999. Use non-nested *Number* tags.

The correct solution is in <u>Chapter 3, Section 2.4.2</u>, page 62.

Exercise 4

Create a graph that recognizes and annotates Roman numerals from 1 to 3 999. Use nonnested *RomanNumeral* tags.

The correct solution is in <u>Chapter 3, Section 3.5.2</u>, page 69.

4.2 Cascade exercises

Exercise 5

Before starting exercise 5

 $Open \ the \ text: {\it MyUnitex} \ {\it English} \ {\it Corpus} \ {\it UnitexGettingStarted} \ {\it DombeyAndSon} \ {\it DombeyAndSon.txt}$

¹¹ Affirmative, negative, interrogative and interro-negative forms.

Create a cascade of graphs that recognizes and annotates measures with six different types: *currencies* (pence, sixpence, ¹² shilling, pound), *duration* (minute, hour, day, week, month, year), *lengths* (mile, foot, yard), *speed* (mile an hour), *temperature* (degree) and *weight* (ton). Use *Number* and *Measure* tags.

Modify the synthesis cascade to have the *Number* tags only inside the *Measure* tags.

The correct solution is in <u>Chapter 4, Section 3</u>, page 88.

Exercise 6

Before starting exercise 6 Open the text:*MyUnitex\English\Corpus\UnitexGettingStarted\jubileePlainText\jubileePlainText.txt*

Create a cascade of graphs that recognizes and annotates absolute dates (the year is specified), relative dates (the year is not specified), names of towns or cities and names of streets or avenues. Use *AbsoluteDate*, *RelativeDate*, *Town* and *Address* tags.

Modify the synthesis cascade to have non-nested Address tags.

The correct solution is in <u>Chapter 4, Section 1</u>, page 71.

Exercise 7

```
Before starting exercise 7
Open the text:MyUnitex\English\Corpus\UnitexGettingStarted\jubileePlainText\jubileeXmlText.xml
```

Modify the previous cascades so that the annotation is only in the *<body></body>* part.

The correct solution is in <u>Chapter 4, Section 2</u>, page 84.

4.3 Dictionary exercises

```
Before starting Dictionary exercises
Open the text:MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon.txt
```

Create a dictionary that contains all forms of the following nouns and verbs. Use it to parse the text.

Exercise 8

The beginning of the list:

- *advantage, agent, firm*: nouns with suffix *s*.
- *branch, bus, box:* nouns with suffix *es*.
- *accept, connect, work*: verbs with suffixes s, ed and ing.

The correct solution is in <u>Chapter 5, Section 2.1</u>, page 107.

Continuation of the list:

- *company, complexity, family*: nouns with suffix *ies*.
- *leaf, thief, calf*: nouns with suffix *ves*.
- *wife, knife, life*: nouns with suffix *ves*.
- *mouse, louse*: nouns with suffix *ice*.
- *narcissus, terminus, tumulus*: nouns with suffix *i*.

¹² Sixpence is the name of an old coin.

- *phenomenon, criterion, datum*: nouns with suffix *a*.
- *love, live, give*: verbs with suffixes s, d and ing.
- *accompany, try, worry*: verbs with suffixes *ies, ied* and *ing*.

The correct solution is in <u>Chapter 5, Section 2.2</u>, page 113.

Continuation of the list:

- goose, tooth: nouns with suffix eese or eeth.
- *foot*: noun with suffix *eet*.
- *man, woman, hypothesis*: nouns in which the penultimate letter changes to *e*.
- *jog, plan, stop, prefer*: verbs with double consonant.
- *slit, split*: verbs with suffixes *s* and *ting*.
- *quiz, gas*: nouns with suffix *zzes* or *sses*.
- *shoot, speed, meet*: verbs for which the double letter (*oo, ee*) disappears in the simple past and the past participle.
- *kneel, sleep, creep*: verbs with suffixes *s, elt* or *ept* and *ing.*

The correct solution is in <u>Chapter 5, Section 2.3</u>, page 116.

Exercise 9

End of the list:

- *air of mystery, sort of thing, tree of knowledge*: MWUs in which only the first component is inflected.
- *crow's nest, death's-head, hair's breadth*: MWUs in which only the last component is inflected.
- *boa constrictor, prince regent, secretary general*: MWUs in which only the first component is inflected.
- *altar rail, Anglo-Norman, bad boy*: MWUs in which only the last component is inflected.
- *after-life, eye-cup, sky-blue*: MWUs in which only the last component is inflected; word tokens can be written together.
- *three quarter, blood red, self glorification*: MWUs in which only the last component is inflected; hyphen can be added.
- *carry out, take off, wash out*: MWUs in which only the last component is inflected; hyphen can be added, space can be omitted.
- *attorney general, notary public*: MWUs in which either the first or the second component (but not both!) are inflected.

The correct solution is in <u>Chapter 5, Section 3</u>, page 118.

Exercise 10

Create a dictionary graph that recognizes all Roman numbers from 1 to 3 999. Use it to create a *Word List*.

The correct solution is in <u>Chapter 5, Section 5.1</u>, page 128.

4.4 Lexicon grammar exercise

Exercise 11

Before starting Lexicon exercise 10 Open the text:*MyUnitex**English**Corpus**UnitexGettingStarted**DombeyAndSon**DombeyAndSon.txt* Open the *particleVerbTable.xlsx* Lexicon-grammar in the folder: *MyUnitex**English**Graphs**UnitexGettingStarted**LexiconGrammar**ParticleVerbs*

Create a graph that recognizes all positive forms from the lexicon-grammar *particleVerbTable.xlsx*.

The correct solution is in <u>Chapter 6, Section 1</u>, page 137.

4.5 XAlign exercise

Exercise 12

Align the text:

with its translation:

*MyUnitex**English**Corpus**UnitexGettingStarted**DombeyAndSonFra**DombeyAndSonFra*.*txt* The correct solution is in <u>Chapter 6, Section 2</u>, page 141.

4.6 Script exercises

Exercise 13

Create a script to perform the cascade from Exercise 7, page 4.

The correct solution is in <u>Chapter 7, Section 2</u>, page 149.

Exercise 14

Complete the previous cascade to get an inventory of tag occurrences with the name of the file and the information inside *title* tags.

The correct solution is in <u>Chapter 7, Section 4</u>, page 158.

5 Acknowledgement

The authors sincerely thank all those who helped and advised them while preparing this work: Laurence Danlos, Bérengère David, Anne Dister, Nathalie Friburger, Cristian Martínez, Alexis Neme, Benoît Sagot and Duško Vitas. Authors are especially grateful to Éric Laporte and Anne-Lyse Minard who carefully read the entire manuscript.

1 Discovering Unitex

1.1 Unitex installation

The Unitex installation file can be found at the URL:

https://unitexgramlab.org

Unitex is a cross-platform software that can be installed on all widely used operating systems (Windows, Mac, Unix and so on). In the following text, we will describe the installation under Windows. Some details shown here may be slightly different on a non-Windows system. For other systems, refer to Section 1.3 of the Unitex user manual downloadable at the URL:¹³

https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf

1.1.1 Creating a private working folder

Before launching Unitex for the first time, we must create a private working folder in which we will work. For example, let us go to the *Documents* folder and create a new folder, called for example *MyUnitex* (No spaces and case sensitive):¹⁴

 $C: \ Ocuments \ MyUnitex$

1.1.2 Installation process

We start by launching the installation file:



¹³ There is also a French version of the Unitex user manual, downloadable at the URL: <u>https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-fr.pdf</u>

¹⁴ Unitex is a cross-platform software, so filenames should not contain spaces and diacritics, which only Windows accepts. Also, calls to these filenames may be case-sensitive.

We will click the *Next* button, and then twice the *I Agree* button.¹⁵ In the *Choose Components* window, we will get this screen and then we will follow the steps:

Choose Components Choose which features of Un	itex/GramLab you want to install.	
Check the components you w install. Click Next to continue	ant to install and uncheck the components you don't	want to
Select the type of install:	Standard	~
Or, select the optional components you wish to install:	Core Components (required) V Visual Integrated Environments Consequences Source Code V User Manual V Start Menu and Desktop Shortcuts	
Space required: 203.7MB	Description Position your mouse over a component to see its description.	s

- Click on the check box next to the Visual Integrated Environments line to uncheck it, then on the + sign placed on the left of this line to show possibilities and select the Unitex Java IDE line. Click on the – sign to the left of the Visual Integrated Environments line to hide the possibilities.
- Click on the check box next to the Languages Resources line to uncheck it, then on the + sign placed to the left of this line to show possibilities and select English and French.¹⁶ Click on the – sign to the left of the Languages Resources line to hide.
- Click on the check box next to the User Manual line to uncheck it, then on the + sign placed to the left of this line to show possibilities and select the line English.¹⁷ Click on the sign to the left of the User Manual line to hide.

Unitex programs are written in C language, while its interface is written in Java language. Java must therefore be installed on the computer, which is frequently the case. In fact, Unitex favors a certain version of Java, JRE, and tests if it is installed on the computer. However, if it is not, the following window appears:



We have three choices: install Java ourselves, let Unitex install JRE or continue without installing it. If we know that Java is installed on the computer and we do not want to

¹⁵ The two licenses (for the programs and for the resources) are LGPL licenses that authorize their free use including the use in commercial products.

¹⁶ In addition, some other languages, of course, if you want. We will use French in the chapter dealing with aligned texts (<u>Chapter 6, Section 2</u>, page 141).

¹⁷ You can select the *French* manual as well if you prefer or check only *French*.

additionally install the JRE version, we should choose the third option. Otherwise, we advise you to choose the second option and let Unitex try to install it during the setup process.

After selecting the second option click on the *Next* button. If you are the administrator of your computer, Unitex offers you an installation only for you or an installation for all users of the computer:



If you choose the *just for me* option, Unitex offers the installation in the User folder:

Install Location to to install Unitex/Gran	nLab.		Ū
e the folder in which to install Unitex/Gran	nLab.		
vill install the poplication in the following (
and select another.	folder. To install	in a different fo	older, click
that this folder will be the installation Unit	tex/GramLab fol	der. It must be	different from
nitex/GramLab workspace folder.			
nation folder where Unitex/GramLab will b	be installed		
\Users\maurel\AppData\Local\Unitex-Gra	mLab	Bro	wse
required: 42.6MB			
available: 168.7GB			
Paris-Est Marne-la-Vallée			

If you choose the *for anyone* option, Unitex offers the installation in the Program Files (x86) folder:¹⁹

C:\Program Files (x86)\Unitex-GramLab\App

¹⁸ userName is your name under the Windows system.

¹⁹ To choose this option, you must be an IT administrator.

Unitex Getting Started - Denis Maurel and Cvetana Krstev

Second Install Location					
hoose Install Location	-				
Choose the folder in which to install Unitex/Gram	LaD.			1	1
Setup will install the application in the following fo Browse and select another.	lder. To insta	ll in a diffe	rent fold	ler, dick	
Notice that this folder will be the installation Unite	ex/GramLab fo	older. It mu	ust be di	fferent fr	om
your Unitex/GramLab workspace folder.					
your Unitex/GramLab workspace folder.					
your Unitex/GramLab workspace folder.	e installed				
your Unitex/GramLab workspace folder. Destination folder where Unitex/GramLab will by C:\Program Files (x86)\Unitex-GramLab \	e installed		Brow	se	
your Unitex/GramLab workspace toider. Destination folder where Unitex/GramLab will be C:\Program Files (x86)\Unitex-GramLab\	e installed		Brow	se	
your Unitex/GramLab workspace folder. Destination folder where Unitex/GramLab will be C:\Program Files (x86)\Unitex-GramLab\ Space required: 42.6MB	e installed		Brow	se	
your Unitex/GramLab workspace folder. Destination folder where Unitex/GramLab will be C:\Program Files (x86)\Linitex-GramLab \ Space required: 42.6MB Space available: 409.3G8	e installed		Brown	se	
Destination folder where Unitex/GramLab will be C:\Program Files (x86)\Unitex-GramLab \ Space required: 42.6MB Space available: 409.3GB verské Paris-Est Marne-la-Valée	e installed		Brow	se	

Both options are fine. If you want to change this setting, don't install Unitex in the *MyUnitex* folder that you have created: this folder must remain separate from the installation folder. After choosing the folder for installation that is appropriate for your computer, click on the *Next* button, and after that on the *Finish* button:

SUnitex/GramLab 3.3 Rev. 8 (6	4-bit) Setup				
, с , с ,	Completi	ng Unitex	/GramLat	o Setup	
	Unitex/GramLal	b has been inst	alled on your cor	mputer.	
ัม ๆ เนื	Click Finish to d	lose Setup.			
Mi K (Launch the	Unitex Visual ID	E		
$\langle \rangle$					
	Visit the Unitex	/GramLab webs	ite for support i	nformation.	
		,			
		< Back	Finish	Cance	ł

Unitex starts automatically; if that does not happen, see <u>Section 1.2.1</u> that follows.

1.1.3 Initial choices

Unitex will then ask you to choose your private working directory (or working folder):



Select then the *MyUnitex* folder that you have already created (see <u>Section 1.1.1</u>, page 7):

C:\Documents\MyUnitex

And then choose the language (in this case, *English*) and click on the OK button:

Unitex		\times
?	User: maurel Choose the language you want to work on:	
	English	-
	OK Exit	

1.1.4 Check or modify your private working directory

If you want to check your private working directory, you can open the *Info/Preferences.../Directories* menu.

Preferences for English					
Morphologi	cal-mode diction	aries	SVN	Encoding]
General	Directories	Lang	luage	Presentati	on
Private Unite	ex directory (whe	re all us	ser's dat	a is to be sto	red):
U:\Docume	nts\MyUnitex			Set	
Graph repository: Set Produce log information in directory:					
C:\Users\Pi	ublic\Desktop			Set	
Clear all	logs				
			ОК	Cano	el

To modify your private working directory, click the *Set* button to choose a new one, and click on the *OK* button. Unitex must then be closed and reopened to take the changes into account.

×
Your private Unitex directory is now:
U:\Documents\MySecondUnitex
You must relaunch Unitex to take this change into account.
ОК

It is thus possible to have several private working directories, one per project.

The name of your private working directory is saved in a specific file which depends on where Unitex is installed.

If you chose the *just for me* option, Unitex save this name in the file:

C:\Users\userName\AppData\Local\Unitex-GramLab\Users\userName.cfg

Else, if you chose the *for anyone* option, Unitex save this name in the file:

C:\Users\userName\.unitex.cfg

To avoid opening and closing Unitex to change the private working directory, it is possible to directly modify this file. You can also create a shortcut to it on the desktop.

1.2 Using Unitex

1.2.1 Starting Unitex

It can happen that Unitex is not launched at the end of the setup or that you have accidentally closed Unitex. It is also possible that you are using a computer on which Unitex is already installed. In these three cases, you must launch Unitex yourself, as you will always do when

wanting to work with it. The easiest way is to use the shortcut placed on the desktop during installation and double click it: 20



If you do not have this shortcut or it does not work, it is possible to launch Unitex by going to the folder where the Unitex program has been installed:

C:\Program Files (x86)\Unitex-GramLab\App or C:\Users\userName\AppData\Local\Unitex-GramLab\App

Double-click on the file named *Unitex* having the type *Executable Jar File*. A permanent solution should be to send this file to the desktop as a shortcut by right clicking on this file and choosing the *Send to/Desktop* option. This will create a shortcut of the *Unitex executable jar file*.

1.2.2 The Unitex inteface

Unitex processes different types of files: texts, dictionaries, graphs and so on. In the Unitex interface, each menu is specific to one of them and therefore has, for instance, an *Open* submenu (or option).

Menus	types of files
Text	text to analyze
DELA	dictionary
FSGraph	graph
Lexicon-Grammar	lexicon-grammar
XAlign	parallel texts
File Edition	textual file to edit

In addition, there are three more menus: *Windows, Help* and *Info*. The Unitex user manual can be opened directly from Unitex, via the *Help* menu.

²⁰ Whenever you update Java, you will need to redo this shortcut by restarting the installation or changing its target.

1.2.3 **Opening a text**

Before starting this section

Go to the folder: MyUnitex\English\Corpus\UnitexGettingStarted Create a new folder named DombeyAndSon (without space). Go to the folder: $MyUnitex \ English \ Corpus \ Unitex \ GettingStarted \ DombeyAndSon^{21}$ Download the DombeyAndSon.txt file.

We will click the *Text/Open...* menu to open the file named:

MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon.txt

and answer No to the question Do you want to preprocess the text?



Unitex will create a new file: 22

MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon.snt



The top bar displays some numbers: 0 sentence delimiter, 803 028 (16 453 diff) tokens, 363 065 (16 434) simple forms, 0 (0) digits.²³ Their meaning is:

sentence delimiter: Unitex can insert the special symbol {S} at the end of a sentence during

the preprocessing (see Section 3.3.1, page 37). Since we skipped the preprocessing there are no such symbols.

²¹ This file is the novel *Dombey and Son*, by English author Charles Dickens. According to Wikipedia: "It follows the fortunes of a shipping firm owner, who is frustrated at the lack of a son to follow him in his footsteps; he initially rejects his daughter's love before eventually becoming reconciled with her before his death". ²² Unitex never works on an original text file.

²³ These numbers are recorded in the file named:

- *token*: it is either a string of letters²⁴ or any other single character (including spaces). There are 803 028 tokens in this novel, but only 16 453 different.
- *simple form*: it is a string of letters. There are 363 065 of them in this novel, but only 16 434 different.

At the bottom left is a reduced window, the *Token list*, that lists the tokens of the text and their number of appearances. By double-clicking on it, we obtain this list in decreasing order of frequency. You can also have it in the alphabetical order by clicking the *Char Order* button.²⁵

Toker	n list	r 🛛	\mathbf{X}
By F	requence	By Char Order	
353703			
37941	,		
15885	the		
13672			
13574	and		
9381	to		
9151	of		
7824	"		
7766	**		
7159	a		
6228	in		
5089	1		
5075	his		
5041	that		
4846	I		
4695	her		
4057	was		
3900	he		
3840	it		
3585	with		-
4			

We can notice that this *Token list* begins with space, punctuation characters and functional words, which is normal, because they are the most frequently used tokens.

1.2.4 Applying dictionaries

We will open the *Text/Apply Lexical Resources...* menu. There we will find available dictionaries, on the left those in the private working folder, on the right those that come with the Unitex distribution.²⁶ The default dictionaries are preselected (*dela-en-public.bin* and *Dnum.fst2*).

 $MyUnitex English Alphabet_sort.txt$

C:\Users\maurel\AppData\Local\Unitex-GramLab\English

or

and copy the three files *Alphabet.txt*, *Alphabet_sort.txt* and *Norm.txt*, and then paste them in the private folder *English*:

 $MyUnitex \ English$

²⁵ These lists are recorded in the three files:

²⁴ Letters of a language are defined in the file:

MyUnitex\English\Alphabet.txt

A sorting order, for arranging concordances and dictionaries, is specified in the file:

If these files are not present, Unitex will not work. In this case you have to open the folder where Unitex is installed:

C:\Program Files (x86)\Unitex-GramLab\English

 $[\]label{eq:myUnitex} wyUnitex \end{Son} on \ond{Son} snt\tokens.txt \\ MyUnitex\end{Son} on \ond{Son} on \on$

²⁶ We have not yet dictionaries in the private working folder, see <u>Chapter 5</u>, page 107.

Chapter 2: First steps

_			
🗒 Lexical Reso	urces		X
Select the diction using the arrows program before t	aries to be applie . Note that system he user ones.	d. You can sort the dictionaries are g	m one by one iven to the Dico
- User resources		System resource	es
		dela-en-public.b	in 🔺
		Dnum-ord.fst2	
	· ·	Dnum.fst2	
Right-click a diction	onary to get inform	ation about it :	
			×
Clear	Default	Set Default	Apply

By clicking the *Apply* button, we apply the dictionaries to the text and we obtain the *WordList*.

Word Lists in U:\Documents\MyUnitex\En	glish\Corpus\DombeyAndSon 🖬 🛛	X
DLF: 22604 simple-word lexical entries	ERR: 580 unknown simple words	
a, .DET+Dind:s	Filter unknown words with tags.ind	
aback, .ADV	abear	
abaft, ADV	admonitorial dolescens	
abandon, .N:s	adwentures	
abandon,.V:W:Pls:P2s:Plp:F	adwenturs	
abandoned, .A abandoned, abandon.V:K:Ils:	Aldgate	
abandoning,.N:s	Alfred	
	Alraschid	
DLC: 937 compound lexical entries	Amens	
able seaman, .N+XN+Hum+z1:s	Anne	
aching void, .N+XN+z1:s	Anno	
Act of Parliament, N+NPN+2	Antony	
after dark, .N+XN+z1:s	anyhows	
after-life, afterlife.N+XN+	appy	
all alone, A+DA+z1	ard	
all around, .A+DA+z1	ards	•

The WordList is displayed in three windows:²⁷ DLF (22 604 simple forms), DLC (937 compounds or multiword units) and ERR (580 words that are not in applied dictionaries).

1.2.5 The DLF window

We will first look at the format of the eleventh line of the *DLF* window (the first line not displayed in the above screenshot):

abandoning,abandon.V:G

At the beginning of the line is the word from the text, followed by its heading form (or lemma), after the comma and before the period. After the period comes a word's part of speech (*V* for verb), and it is followed, after the colon, by the inflection code (*G* for gerund). This word from the text is recognized as possibly being the gerund of the verb *to abandon*. If the lemma form

²⁷ These windows correspond to the following files:

MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon_snt\dlf MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon_snt\dlc MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon_snt\err

They come with three additional files: MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon_snt\dlf.n MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon_snt\dlc.n MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon_snt\err.n These contain just the number of entries in the corresponding files

These contain just the number of entries in the corresponding files.

does not appear, which is the case on the line just above the analyzed one where the period comes right after the comma, it means that the lemma is identical to the text word itself; in this case the lemma of the noun *abandoning* is simply *abandoning*.

It may seem strange to see two lines beginning with *abandoning*. We must keep in mind that words out of context are ambiguous, and *abandoning* can be the noun *abandoning* or the gerund of the verb *to abandon*.

It is possible to force one analysis rather than another by modifying the text. For instance, by replacing *after abandoning herself* with *after {abandoning, abandon.V:G} herself*, we explicitly assign the verb category to the word *abandoning*. This expression between curly braces is counted as a single token and is called a *lexical tag* (However, if you want to do that manually, you shall have to close the text with Unitex, edit the *.txt* file, save it and reopen it with Unitex).

A line can also contain one or more features for the entries, which are preceded by a plus sign. One example can be seen in the first line of the *DLF* window where the *Dind* features is used to indicate that the determiner is indefinite:

a,. DET+Dind:s

1.2.6 The DLC window

The *DLC* window has the same format as the *DLF* window. If the multiword form contains a character used as a delimiter (comma, period, plus, colon and slash),²⁸ it must be preceded by a backslash. Such a case does not occur in our *DLC* window.

1.2.7 The ERR window

The words that were not recognized by the applied dictionary are mostly proper names (*Aldgate*, *Alfred*, *Alraschid*...); but also foreign language words: *Anno* (Latin), Roman numbers, unusual orthography (*adwentures*), parts of contracted verbal forms (*didn*, *hasn*...) or simply words that were not recorded in the applied dictionaries (*nevyless*,²⁹ unbusiness, undauntable...).

2 A detailed example

Before starting this section

Go to the folder:MyUnitex\English\GraphsCreate a new folder named UnitexGettingStarted (without space).Go to the folder:MyUnitex\English\Graphs\UnitexGettingStartedCreate a new folder named Meet.

2.1 Concordances with one word query

Suppose, for example, that we want to study the use of the verb *to meet* in our novel. In order to do that we will open the *Text/Locate Pattern* menu, select the *Regular expression* option and type *meet* in the corresponding field.

²⁸ The slash enables comments, see <u>Chapter 5, Section 4.3.1</u>, page 125.

²⁹ The word *nevy* is an old form of *nephew*.

Chapter 2: First steps

🔲 Locate Pattern						
Locate configuration	Advanc	ed options				
Locate pattern in the fo	rm of:					
Regular expression	n:					
meet						
O Graph: Set						
Activate debug mode						
Index	Gramm	ar outputs				
 Shortest matches 	Are r	iot taken into account				
Longest matches	O Merg	e with input text				
 All matches 	Repl	ace recognized sequences				
Search limitation Stop after 200 Index all occurrence 	matches es in text	SEARCH				
Search algorithm: Paumier 2003, working on text (quicker) automaton intersection (higher precision)						

We will launch the search by clicking the *Search* button, which will yield 58 matches.

Result Info	\times
58 matches 58 recognized units (0.007% of the text is covered)	
ОК	

Clicking the *OK* button will open the *Located sequences* dialog box. In order to obtain concordances, we need to click the *Build concordance* button (situated at the bottom right of the *Located sequences* dialog box).

🗂 Located sequences 🛛						
Concordance Statistics						
Modify text						
Resulting .txt file:						
Set File	GO					
Extract units						
Set File:						
Extract matching units	Extract unmatching units					
Concordance presentation	Concordance presentation Use a web browser to view the concordance					
Show differences with	previous concordance					
Show ambig	uous outputs					
Show matching sequences in	context					
Context length: Stop at:	Sort according to:					
Left 40 chars {S}	Center, Left 🗾					
Right 55 chars (S)	Build concordance					

The produced list is a hypertext file in which each recognized sequence leads to the corresponding location in the text.³⁰ For instance, if we click on the first line, this opens the following window (right):



As the verb *to meet* has conjugated forms *meets, met* and *meeting,* we would have to repeat what we have just done three more times in order to obtain a concordance of these three forms as well.

i ma	ten												
10	for	the	Bride.	who	verv.	verv	celdom	moote	it.	Cousin	Feenix	ricoc.	147

📄 Concordance: U:Documents'MyUnitex/English\Corpus\DombeyAndSon\DombeyAndSon_snticoncord.html 📰 🖬 🛛	Concordance: U:\Documents\MyUnitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 🗖 🗹 🛛
46 matches	26 matches
head, the younger lady glancing at him, met his eyes. "You reside here, 🖆	eard, and Mr Sownds goes out. Mrs Miff, meeting Mr Dombey's eye as it is 📤
him more nearly in their mute address, met his, and did not know it! The	1 you, lady lass!" replied the Captain, meeting her eyes in confusion. "
of the Commander of the Cautious Clara met with no response to its impu	"Good morning, Ma'am," said the Major, meeting Miss Tox in Princess's P
e known her for his own child if he had <u>met</u> her in the streets, or would	k leave with all humility. But Withers, meeting him on the stairs, stood
1 the attention and distinction she had met with. Mr Dombey being hastil	own, after the nuptials, a man he knew, meeting him in the lobby of the l
hurt by the unmerited contumely she had met with from Louisa." But there	said Mrs MacStinger. "This is indeed a meeting! I bear no malice now, C
cry of distress. No, no. Mr Dombey had met with an accident in riding.	cide," he thought, "to give this fool a meeting, I am less likely to be
gain very near the place where they had met them first. After meeting the	his voice; their having been so near a <u>meeting</u> , face to face; he would l
s strange benches and candlesticks, and met a harp in a green greatcoat :	red to me that if I could bring about a meeting, unexpected on both side
s to come in. He followed directly, and met Harriet Carker on the stair.	o her to pat his horse's neck, Florence meeting his eyes, saw, rather the
ght when she went down to that room and met the never-to-be forgotten lo	ghtens my heart, even in such a strange meeting and parting, to hear this
lids droop; not least so, when the face met that now fronting it. Once,	and heir, Florence was prepared for the meeting: so, a favourable moment
out, and up towards her room, Florence met her immediately, coming down	nd alone, rather than incur the risk of meeting his displeasure. As the
es, but on her way to the door Florence met the visitor coming in. With	and said that he had had the honour of meeting Paul's father at a public
the ladies out. Then, and but then, he met her glance for an instant in	d the time since I had the happiness of meeting you? Have you walked at
back, bent her dark glance upon him. He met it with his blandest and mos	or reproach, and made a very hysterical meeting of it. Then converting t
nd lying dreadfully. It appears that he met everybody concerned in the l	and that her imperious glance droops in meeting his, and seeks the groun
, trembling, raised it to her lips, she met his glance. The look was col	ce where they had met them first. After meeting them thus, three or four
in no degree diminished by the look she met. "It is enough," said Edith,	it than on the first occasion of their meeting. She led Florence to a c
going to counterfeit surprise, but she <u>met</u> them, and stopped him, if su	y the childhood I have suffered, by our meeting in this dreary house at i
at the dead of night. For this, I have <u>met</u> you here, in a false name, a:	g, by a little stratagem, effected this <u>meeting</u> . I cannot say that I was
ered past outside; and God knows I have <u>met</u> with my reply. Oh mother, mo	ld me that in the having given you that meeting, and leaving you to be d
	<pre>i</pre>

If we look closely at the concordance of the word *meeting*, we can see that some lines represent a form of the verb to meet (Mrs Miff, meeting Mr Dombey's eye...), while others represent the noun meeting (This is indeed a meeting!). You will remember that in Section 1.2.5, page 15, we explained that some words can have several interpretations.

2.2 Concordances with a lemma

Instead of launching four different searches, we could have obtained the same result with one query in the form of the regular expression *meet+meets+meeting+met*. With this query, we

³⁰ These concordances are registered in the file:

*MyUnitex**English**Corpus**UnitexGettingStarted**DombeyAndSon**DombeyAndSon_snt**concord.html* With each new search and concordance produced, this file is replaced, but you can save it using a different name if you want to keep it. The saved concordance can be reopened by the *Text/Open Concordance* menu.

would obtain a list of 131 concordance lines, containing all the lines previously obtained in four subsequent runs (58+1+46+26). However, all these queries ignore the fact that Unitex uses dictionaries, containing among other entries, all forms of the verb *to meet*.³¹

meet,.A			
meet,.N	S		
meet,.V:	W:P1s:P2s:P1p:P2p:P3p		
meeting	N·s		
meeting	,meet.V:G		
meets,m	eet.N:p		
meets m	eet V·P3s		
met,mee	et.V:K:l1s:l2s:l3s:l1p:l2p:l3p		

Instead of launching four searches, or listing all forms of the verb, we can use one query with the lemma form of the verb, *<meet>*.³²

Cocate Pattern					
Locate configuration	Advanc	ed options			
Locate pattern in the fo	rm of:				
Regular expression:					
<meet></meet>					
O Graph: Set					
Activate debug mode					
Index Crammer outputs					
 Shortest matches 	Are n	not taken into account			
Longest matches	O Merg	ge with input text			
 All matches 	Repl	lace recognized sequences			
Search limitation					
Stop after 200	matches	SEARCH			
 Index all occurrence 	es in text				
Search algorithm: Paumier 2003, working on text (quicker) automation interconting (higher precision)					
· · · · · · · · · · · · · · · · · · ·					

With this query, we will obtain the same result, a list of 131 concordance lines.

2.3 Concordances with a graph query

Until now, only simple forms of the verb *to meet*, one of the four possible forms, appeared in the concordances. But we might like to highlight longer sequences in concordance lines: the sequences containing, beside forms of the verb *to meet*, also words that we see in the left context of concordance lines: *did meet*, *should meet*, *would meet*, *shall meet*, *will meet*, *would seldom meet*, *may ever meet*, *had met*, *have met*, *being met*, *had been met*, *have been met*, *had ever met*, *had never met* and *has met*. The reason to switch from regular expressions to using Unitex graphs will become obvious with these examples.³³

³¹ Note that the noun *meetings* (plural form of *meeting*) is not present in this text.

³² Note that this does not remove the ambiguity, because *<meet>* is also the lemma of a noun and an adjective.

³³ For the sake of simplicity, we will omit the modal form (*may ever meet*) and the emphatic past (*did meet*).

2.3.1 A first graph

Let us open the *FSGraph/New* menu. We will get a window with a starting point on the left and an ending point on the right.

E	🗒 Graph 🦾												P	Ø
		Ď	8	Î C	C	ß	7 X	••••	с Г	Q	₹>	() ()	<>	\$*
												 		▲
														=
		ò				\bigcirc								
						Ŭ								
							Ш							

We will now create in five steps a graph that recognizes four word forms *meet, meets, met* and *meeting*:

- 1. Click on the starting point (start box).
- 2. Position the cursor a little more to the right, click the right mouse button and select *Create box* from the displayed menu.



3. Type without any space *meet+meets+met+meeting*³⁴ and finish with clicking the *Enter* key.

³⁴ These words will appear in the formula bar, just below the task bar of the graph window. They will replace the *<E>* symbol, which means *empty*. The plus sign means *or*.



4. Click on the *meet+meets+met+meeting* box and then the ending point (the end box).



5. Click on the *Save* button (the leftmost in the task bar). Unitex will open the Graph folder to save it.³⁵

Go to the folder:

And save this graph in it using the name *meet*. Unitex will add the *.grf* extension.

³⁵ The whole process is shown in the video 2.2.3.1.meet.mp4 which is in the folder: $MyUnitex \setminus English \setminus Graphs \setminus UnitexGettingStarted \setminus Meet \setminus 2.2.3.1$

Notes

1. Clicking on a box, then on another one creates a path between them or, if it already exists, removes it.

2. To delete a box, you must delete its content (by using the *Delete* key, for example), then validate with clicking the *Enter* key.

2.3.2 Concordances

In order to produce a concordance with our first graph, we will open the *Text/Locate Pattern* menu and click the *Set* button next to the *Graph* field. The *Graphs* folder will open and we will move to the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Meet*

where we will find and select the meet.grf file.

🔲 Locate Pattern		X				
Locate configuration	Advance	ed options				
- Locate pattern in the fo	Locate pattern in the form of:					
O Regular expression:						
<meet></meet>						
Graph: MyUnitex\English\Graphs\UnitexGettingStartedWeet\meet.grf Set						
Activate debug mode						
Index O Shortest matches	Gramma Are n	ar outputs ot taken into account				
Longest matches	O Merg	e with input text				
 All matches 	Repl	ace recognized sequences				
Search limitation						
Stop after 200	matches	SEARCH				
 Index all occurrence 	es in text					
Search algorithm:						
Paumier 2003, wor	king on te	d (quicker)				
 automaton intersection 	tion (high	er precision)				

We will launch the search, as before, by clicking the *SEARCH* button. Finally, we will obtain the same list of 131 concordance lines.³⁶

As before, we can replace four word forms with their common lemma.

- 1. Select the meet+meets+met+meetingbox, type < meet> instead of meet+meets+met+meeting then press the *Enter* key.
- 2. Click the diskette icon, at the left end of the task bar, to save the graph.

³⁶ Note that *Locate Pattern* creates a new file, *meet.fst2*: it is a compiled version of the *meet.grf* graph. If we used the compiled version of a graph in *Locate Pattern* after modifying it, the modified version would not be used for search.

meet grf (U:Documents\MyUnitex\English\Graphs\UnitexGettingStartedMeet)	r 🛛 🛛
	() <> \$* {
	_
	~

Now, we will begin to highlight longer sequences in concordance lines by adding the future tense and the conditional, i.e. four examples in our concordance: *shall meet*, *will meet*, *should meet* and *would meet*.

2.3.3 Adding the future tense and the conditional

The future and the conditional use the forms of the auxiliaries, *will*, *would*, *shall* and *should*, which can be represented by two lemmas *<will>* and *<shall>*. We will add two boxes to our graph:

- 1. Click on the starting box, then move the cursor up and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 2. Type <will>+<shall>, then press the Enter key.
- 3. Click on the *<will>+<shall>* box, then move the cursor right, click the right mouse button and select *Create box* from the displayed menu.
- 4. Type *meet*, then press the *Enter* key.³⁷
- 5. Click on the *meet* box and then on the final box.
- 6. Rearrange the graph.
- 7. Click the Save button.

³⁷ Pay attention: *meet* without angle brackets.



The concordance produced with this new *meet.grf* graph contains again 131 lines and among them lines with highlighted *shall meet, should meet, will meet* and *would meet*. You will find these occurrences at the end of the concordance list.

Concordance: U:\Documents\MyUnitex\English\Corpus\DombeyAnd	Son\DombeyAndSon_snt\concord.html 🗾 🗖 🛛
131 matches	
a scarlet visage, explains that he has	met Miss Dombey by accident, and
ve the consent of my friend Gay-who has	met me in an exceedingly kind and
n he made allusion to him, and Diogenes	met that overture graciously, ha
it for a moment on both hands, her eyes	met those of Harriet. In a momen
he himself had entered. And their eyes	met. In the quick unsteadiness of
ime in silence. When their eyes at last	met, he began at once to pursue
and caresses to the last. Cousin Feenix	met her at the door, and took he
n the Pump-room; on the day after, they	met them again very near the place
an't. How long it was to be before they	met, or whose plan it was that the
er the good old creature was, whom they	met, and the cold return the dau
iciently to give him good day when they	met in the street. But as that w
red to love you, Florence. Good-bye. We	shall meet again soon. Good-bye!
errupted. four moments are precious. We	shall meet at dinner-time. In the
be present " "This is the last night we	shall meet for a long time-the li
e evening at the old time and place I	shall meet you as usual on your y
, lest, in passing out at the door, she	should meet him unexpectedly: be
me when there was no one else, that we	should meet like this?" Edith st
mbey a verbal invitation, whenever they	should meet, to come and cut his
e to offer a suggestion which he thinks	will meet the position in which
Diogenes was as ridiculous a dog as one	would meet with on a summer's day
t was far from being the last time they	would meet. "I hope so," said Mr

Let us move to the perfect tenses. They consist of the verb *to have* followed by the past participle, *met*. Three such forms can be spotted in our concordance, in the left context of the form *met*: *had met*, *have met* and *has met*.

2.3.4 Adding the perfect tenses

The perfect tenses use the auxiliary forms *have*, *has*, *had* and *having*, which can be represented by the lemma *<have>*. We will enhance our graph by using the copy-paste

technique. We will add four boxes to our graph, one for the *<have>* lemma, one for the *have* form, and two for the *met* form:

- 1. Click on the starting box, then move the cursor down and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 2. Type <have>, then press the Enter key.
- 3. Click on the *<have>* box, then move the cursor right, click the right mouse button and select *Create box* from the displayed menu.
- 4. Type *met*, then press the *Enter* key.
- 5. Click on the *met* box and then on the end box.
- 6. Select the *<have>* box and the *met* box by drawing with the mouse a rectangle that encloses both boxes, then click the *Copy* button (or use *CTRL-C* on the keyboard).
- 7. Click the Paste button (or use CTRL-V on the keyboard): the two boxes appear twice.
- 8. Move the copied boxes to the top of the graph. Click outside the graph to deselect them.
- 9. Click on the top <*have*> box and delete the angular brackets (*have* replaces <*have*>), then press the *Enter* key.
- 10. Click on the <will>+<shall> box and then on the have box.
- 11. Click on the *met* box and then on the end box.
- 12. Rearrange the graph.
- 13. Click the Save button.



The concordance produced with this graph contains the lines with five highlighted sequences *had met*, two sequences *has met* and three sequences *have met*.

Unitex Getting Started - Denis Maurel and Cvetana Krstev

Concordance: U:\Documents\MyUnitex\English\Corpus\DombeyAndS	Son\DombeyAndSon_snt\concord.html 📰 🛛
131 matches	
have known her for his own child if he	had met her in the streets, or w
r all the attention and distinction she	had met with. Mr Dombey being ha
ply hurt by the unmerited contumely she	had met with from Louisa." But t
d no cry of distress. No, no. Mr Dombey	had met with an accident in ridi
em again very near the place where they	had met them first. After meetin
with a scarlet visage, explains that he	has met Miss Dombey by accident,
I have the consent of my friend Gay-who	has met me in an exceedingly kind
you, at the dead of night. For this, I	<u>have met</u> you here, in a false nar
wandered past outside; and God knows I	have met with my reply. Oh mothe:
ned Carker. "As I am so fortunate as to	have met you, let me press my pe
ury. Rich colours, excellently blended,	meet the eye at every turn; in the
e this with her affection when they did	meet, she sought her in her own :
world. Now, when three men of the world	meet together, and are friends-a
wife dead, and one wife coming. Go and	meet her!" In spite of himself,
d had gathered up her flowers to go and	meet them, that they might know
t he heard the woman screaming, "Go and	meet her!" Preparations for a ch
had spoken of; and resolved to try and	meet her, then, upon the stairca
nose. "But I know this-that when people	meet with trials, they must bear
never on good terms, and wouldn't quite	<u>meet</u> without a struggle. She had
e to face for the last time. Wretch! We	meet tonight, and part tonight. 1
We had need to know each other when we	meet again!" CHAPTER XLI. New Vo:
full heart, "Oh, Mama, Mama! why do we	meet like this? Why were you ever
A him norr a smille con a han face with	have been and the second of th

Two important notes on how copy-paste functions in Unitex

1. If only one box is selected, the box is not copied, only its content is. In order to copy-paste a single box, you have to copy the box, create a new one and paste the content.

2. Copying and pasting multiple boxes does not use the usual clipboard. Therefore, there may be competition between pasting text and pasting boxes. If some operation pastes boxes when you wanted to paste text, just erase the unwanted boxes by clicking the *Delete* key, then the *Enter* key. If you have copied both boxes and text and if you want to paste the text into some box, there is a trick that ensures that what is pasted is the text and not the boxes: just type or erase some character in the box before pasting the text.

It remains to insert the passive forms to our graph. They consist of the verb *to be* followed by the past participle, *met*. Three such forms can be found in our concordance, if we look at the left context of *met*: *being met*, *had been met* and *have been met*.

2.3.5 Adding the passive forms

The passive forms use the forms of the auxiliary to be (am, are, was, were, be and being) which can be represented by their lemma <be>

We will add five boxes to our graph, one for the *<be>* lemma, one for the *be* form and one for the *met* form, two for the *been* form:

- 1. Select the whole graph by drawing with the mouse a rectangle that encloses all the boxes (or use *CTRL-A*³⁸ on the keyboard) and then move it down. Click somewhere outside the graph to deselect it.
- 2. Click on the starting box, then move the cursor down and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 3. Type *<be>*, then press the *Enter* key.
- 4. Click on the *<be>* box and then on the *met* box.
- 5. Click on the *<have>* box, then move the cursor up on the right, click the right mouse button and select *Create box* from the displayed menu.
- 6. Type been, then press the Enter key.
- 7. Click on the *been* box, then click on the *met* box.
- 8. Click on the *have* box, then move the cursor up and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 9. Click the *Paste* button (or use *CTRL-V* on the keyboard): *been* is pasted. Press the *Enter* key.
- 10. Click on the *been* box and then on the *met* box.
- 11. Click on the <*will>+*<*shall>* box, then move the cursor down and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 12. Type be, then press the Enter key.
- 13. Click on the *be* box, then move the cursor to the right, click the right mouse button and select *Create box* from the displayed menu.
- 14. Type *met*, then press the *Enter* key.
- 15. Click on the *met* box and then on the end box.
- 16. Click the Save button.



The concordance produced with this graph highlights sequences *being met*, *had been met* and twice *have been met*.

³⁸After the selection of two boxes by drawing with the mouse a rectangle that encloses both.

Unitex Getting Started - Denis Maurel and Cvetana Krstev

🔲 Concordance: U:\Documents\MyUnitex\English\Corpus\DombeyAndSon\DombeyAndSon_snticoncord.html	
131 matches	
patra trusted: community of feeling not	being met with every day: that i
er old house in Brook Street, where she	had been met by Mr Carker- Flore
have known her for his own child if he	had met her in the streets, or we
r all the attention and distinction she	had met with. Mr Dombey being ha
ply hurt by the unmerited contumely she	had met with from Louisa." But t
d no cry of distress. No, no. Mr Dombey	had met with an accident in ridi
em again very near the place where they	had met them first. After meetin
with a scarlet visage, explains that he	has met Miss Dombey by accident,
I have the consent of my friend Gay-who	has met me in an exceedingly kind
ut they have been met, Mrs Dombey; they	have been met, and turned back!"
ur affections in that channel, but they	have been met, Mrs Dombey; they l
you, at the dead of night. For this, I	<u>have met</u> you here, in a false nar
wandered past outside; and God knows I	have met with my reply. Oh mothe:
ned Carker. "As I am so fortunate as to	have met you, let me press my pe
ury. Rich colours, excellently blended,	meet the eye at every turn; in the
e this with her affection when they did	meet, she sought her in her own :
world. Now, when three men of the world	<u>meet</u> together, and are friends-a
wife dead, and one wife coming. Go and	meet her!" In spite of himself,
d had gathered up her flowers to go and	meet them, that they might know
t he heard the woman screaming, "Go and	meet her!" Preparations for a ch
had spoken of; and resolved to try and	meet her, then, upon the stairca
nose. "But I know this-that when people	meet with trials, they must bear
(the stand of the stand of the stand

If we look closer at our concordance, we will notice that sometimes an adverb is inserted between an auxiliary and a main verb, as in the sequences *would seldom meet*, *had ever met* and *had never met* which are therefore not recognized.

2.3.6 Adding adverbs

In order to retrieve these sequences as well, we can insert in our graph a box containing *seldom, ever* and *never*. However, we should try to be more general because almost any adverb can occur at this position. If we look closer at the *Word List* (see <u>Section 1.2.4</u>, page 14), we will see that its third line is:

aback,.ADV

We learn from it that the code for adverbs used in dictionaries is *ADV*. We will therefore add a box containing *ADV*> at appropriate positions in our graph.

- 1. Click on the *<be>* box, then move the cursor to the right, click the right mouse button and select *Create box* from the displayed menu.
- 2. Type <*ADV*>, then press the *Enter* key.
- 3. Click on the *<ADV>* box and on the *met* box.
- 4. Rearrange the graph.
- 5. Click on the *<have>* box, then click the *Copy* button (or use *CTRL-C* on the keyboard). Click somewhere in the blank space to deselect it.
- 6. Click on the start box, then move the cursor down and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 7. Click the *Paste* button (or use *CTRL-V* on the keyboard): <*have*> is pasted. Then press the *Enter* key.
- 8. Click on the previous $\langle have \rangle$ box and type $\langle E \rangle$.³⁹
- 9. Click on the new *<have>* box and on the empty box.
- 10. Click on the start box and on the empty box. $^{\rm 40}$
- 11. Click on the *<have>* box, then move the cursor up and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 12. Type <*ADV*>, then press the *Enter* key.
- 13. Click on this new <*ADV*> box and on the empty box.
- 14. Rearrange the graph.
- 15. Click on the <*will>+*<*shall>* box, then click the *Copy* button (or use *CTRL-C* on the keyboard). Click somewhere in the blank space to deselect it.
- 16. Click on the start box, then move the cursor up and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 17. Click the *Paste* button (or use *CTRL-V* on the keyboard): *<will>+<shall>* is typed. Then press the *Enter* key.
- 18. Click on the previous <will>+<shall> box and type <E>.
- 19. Click on the start box and on the new empty box.
- 20. Click on the *<will>+<shall>* box and on the new empty box.
- 21. Click on the <*will>+*<*shall>* box, then move the cursor up and to the right, click the right mouse button and select *Create box* from the displayed menu.
- 22. Type <*ADV*>, then press the *Enter* key.
- 23. Click on the new <*ADV*> box and on the new empty box.
- 24. Rearrange the graph.
- 25. Click the Save button.



³⁹ Remember that the *<E>* symbol means *empty*.

⁴⁰ This will remove the existing path.

Finally, the concordance produced with this graph highlights the sequences had ever met, had never met and would seldom meet.





2.4 Contexts

2.4.1 Left context

2.4.1.1 The verb to meet preceded by a pronoun

Suppose that instead of displaying all occurrences of the verb *to meet*, we only want to display occurrences in which the verb is preceded on the left by a pronoun (or a pronoun followed by an adverb). We will erase one box and add two boxes in our graph:

- 1. Open the *FSGraph/Save as...* menu to rename the graph to *meetLeft.grf*.
- Select the whole graph by drawing with the mouse a rectangle that encloses all the boxes, except the starting box, and then move it to the right. Click somewhere in the blank space to deselect it.⁴¹
- 3. Rearrange the graph to align vertically the *<will>+<shall>* box, the *<meet>* box, the *<have>* box and the *<be>* box.⁴²



4. Select the *<will>+<shall>* box, the *<have>* box and the *<be>* box, then click the *Reversed link between boxes* button in the tool bar, in the third group of options, fifth in the group. Click on the starting box (this will remove three paths).



Click the *Normal editing mode* button in the same third group of options, first in the group.
 Click on the starting box, then move the cursor to the right, click the right mouse button and select *Create box* from the displayed menu.

⁴¹ Or select any two boxes by drawing with the mouse a rectangle that encloses both, then select the whole graph by typing *CTRL-A* on the keyboard and then move it right. Click somewhere in the blank space to deselect it. Click on the starting box, then move it left.

⁴² To align boxes vertically (or horizontally), we can also select them and right-click to open the *Format/Alignment...* menu.

- 7. Click on the newly created box and type *I+you+he+she+it+we+they*, then press the *Enter* key.⁴³
- 8. Click on the *I+you+he+she+it+we+they* box, then move the cursor to the right, click the right mouse button and select *Create box* from the displayed menu. This box will remain empty.
- 9. Click on the *I+you+he+she+it+we+they* box, move the cursor up and to the right, then click the right mouse button and select *Create box* from the displayed menu.
- 10. Click on the newly created box and type *<ADV>*, then press the *Enter* key.
- 11. Click on the <*ADV*> box and on the empty box.
- 12. Select the *<will>+<shall>* box, the *<meet>* box, the *<have>* box and the *<be>* box.
- 13. Click the *Reversed link between boxes* button. Click on the empty box (this will create three paths).
- 14. Click the Normal editing mode button.
- 15. Click the Save button.



This graph will produce the concordance with forty matched sequences.⁴⁴

Concordance: U:\Documents\WyUnitex\English\Corpus\DombeyAndS	Son\DombeyAndSon_snt\concord.html 📃 🖬 🛛
40 matches	
at he was one of the most agreeable men	he had ever met, and that even M
e'd have known her for his own child if	he had met her in the streets, or
s, with a scarlet visage, explains that	he has met Miss Dombey by accider
and the ladies out. Then, and but then,	he met her glance for an instant
ng back, bent her dark glance upon him.	He met it with his blandest and r
d and lying dreadfully. It appears that	he met everybody concerned in the
th you, at the dead of night. For this,	<u>I have met</u> you here, in a false $i=$
as wandered past outside; and God knows	I have met with my reply. Oh moth
the evening, at the old time and place,	I shall meet you as usual on you:
to her old house in Brook Street, where	she had been met by Mr Carker- F.
n for all the attention and distinction	she had met with. Mr Dombey beind
"deeply hurt by the unmerited contumely	she had met with from Louisa." B
ence, trembling, raised it to her lips,	she met his glance. The look was
rm, in no degree diminished by the look	she met. "It is enough," said Ed
were going to counterfeit surprise, but	she met them, and stopped him, i:
tion, lest, in passing out at the door,	she should meet him unexpectedly
et them again very near the place where	they had met them first. After me
el, but they have been met, Mrs Dombey;	they have been met, and turned be
ed your affections in that channel, but	they have been met, Mrs Dombey;
ter in the Pump-room; on the day after,	they met them again very near the
; I can't. How long it was to be before	they met, or whose plan it was t
mother the good old creature was, whom	they met, and the cold return the
I I I I I I I I I I I I I I I I I I I	han mak in the start But as the

⁴³ If you wish to use the corpus *lvanhoe.txt*, distributed with Unitex, you must add to this list the second person singular, *thou*. However, *thou* never precedes a form of the verb *to meet* in the text we use.

⁴⁴ Do not forget to use the new *meetLeft.grf* graph in *Locate pattern*.
2.4.1.2 The verb to meet preceded by a pronoun and a left context

The concordance we obtained is not like the previous ones, because pronouns are part of the highlighted sequences. It is possible to avoid this if we use the so-called left context:

Select the <*will>+*<*shall>* box, the <*meet>* box, the <*have>* box and the <*be>* box. 1. 2. Click the *Reversed link between boxes* button in the tool bar, in the third group of options, forth in the group. 2 📳 🏶 text mark before box selection Click the first green button (\$*, *Inserts left context mark before box selection*) situated at the 3. right end of the graph editor toolbar. 4. Rearrange the graph. Click the Save button. 5. been) have met <ADV> <wi11> meet <shall> <ADV> Ъe met you he she <meet> it we <ADV> been they <have> met <ADV> <be>

We obtain nearly the same concordance: pronouns are not highlighted anymore, but occurrences are still displayed only when the pronoun is present.

Concordance: U:\Documents\MyUnitex\English\Corpus\DombeyAnd	Son\DombeyAndSon_snt\concord.html 📃 🖬 🛛
40 matches	
er old house in Brook Street, where she	had been met by Mr Carker- Flore
he was one of the most agreeable men he	had ever met, and that even Mr De
have known her for his own child if he	had met her in the streets, or we
r all the attention and distinction she	had met with. Mr Dombey being ha
ply hurt by the unmerited contumely she	had met with from Louisa." But t
em again very near the place where they	had met them first. After meeting
could have done so, and I would, if we	had never met, Florence." "I tru:=
with a scarlet visage, explains that he	has met Miss Dombey by accident,
ut they have been met, Mrs Dombey; they	have been met, and turned back!"
ur affections in that channel, but they	have been met, Mrs Dombey; they l
you, at the dead of night. For this, I	<u>have met</u> you here, in a false nar
wandered past outside; and God knows I	have met with my reply. Oh mothe:
e to face for the last time. Wretch! We	<pre>meet tonight, and part tonight.</pre>
We had need to know each other when we	meet again!" CHAPTER XLI. New Vo.
full heart, "Oh, Mama, Mama! why do we	meet like this? Why were you ever
n him now, a smile was on her face. "We	meet and part tonight," she said
ld me that in the having given you that	meeting, and leaving you to be d
the ladies out. Then, and but then, he	met her glance for an instant in
back, bent her dark glance upon him. He	met it with his blandest and most
nd lying dreadfully. It appears that he	met everybody concerned in the 1
, trembling, raised it to her lips, she	met his glance. The look was colo
in no degree diminished by the look she	<pre>met. "It is enough," said Edith,</pre>
A second se	web them and standard him if and

2.4.2 Right context

It is possible to do the same thing on the right. It is done in a slightly different way. We will start from the *meet.grf* graph and suppose that we want to highlight only the left context of the verb *to meet* (the auxiliaries) and not the verb itself.



The concordance obtained by applying this graph contains 28 matches (among 131 matches obtained by the *meet.grf* graph, 103 recognize only the main verb *to meet* in its various forms). In the recognized sequences, the verb *to meet* is not highlighted.

🔲 Concordance: U:Documents\MyUnitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📈 🗖 🛛
28 matches
patra trusted: community of feeling not being met with every day: that i
have known her for his own child if he <u>had</u> met her in the streets, or w
r all the attention and distinction she <u>had</u> met with.{S} Mr Dombey being
ply hurt by the unmerited contumely she <u>had</u> met with from Louisa."{S} Bu
ry of distress.{S} No, no.{S} Mr Dombey <u>had</u> met with an accident in ridi
em again very near the place where they <u>had</u> met them first.{S} After mee
er old house in Brook Street, where she <u>had been</u> met by Mr Carker-{S} Flo
he was one of the most agreeable men he <u>had ever</u> met, and that even Mr D $_{ m c}$
could have done so, and I would, if we <u>had never</u> met, Florence."{S} "I
with a scarlet visage, explains that he <u>has</u> met Miss Dombey by accident,
I have the consent of my friend Gay—who <u>has</u> met me in an exceedingly kind
u, at the dead of night.{S} For this, I <u>have</u> met you here, in a false nar
wandered past outside; and God knows I <u>have</u> met with my reply.{S} Oh mo
Carker.{S} "As I am so fortunate as to have met you, let me press my pe
ut they have been met, Mrs Dombey; they <u>have been</u> met, and turned back!"
ur affections in that channel, but they <u>have been</u> met, Mrs Dombey; they I
love you, Florence.{S} Good-bye.{S} We <u>shall</u> meet again soon.{S} Good-b
ed.{S} Your moments are precious.{S} We <u>shall</u> meet at dinner-time.{S} In-
} "Hope, my dear Wally.{S} Hope that we <u>shall</u> meet again, on this side of
present."{S} "This is the last night we <u>shall</u> meet for a long time-the la
e evening, at the old time and place, I <u>shall</u> meet you as usual on your t
, lest, in passing out at the door, she <u>should</u> meet him unexpectedly; be

2.4.3 Lexical masks

We have already introduced two types of lexical masks: for instance, *<meet>* matches all forms of the lemma *meet* (see <u>Section 2.2</u>, page 18) while *<ADV>* matches all adverbs (see <u>Section 2.3.6</u>, page 28). These masks rely on the dictionaries used.

There are other masks depending on the use of upper and lower case letters:⁴⁵ <*UPPER>* for a word in uppercase, <*FIRST>* for a word starting with uppercase and <*LOWER>* for a word in lowercase. Besides them, the mask <*WORD>* corresponds to a sequence of letters; the mask <*NB>* to a sequence of digits; the mask <*DIC>* to any dictionary entry; and the mask <*TOKEN>* to any token (see Section 1.2.3, page 13).⁴⁶

We will now transform the *meetLeft.grf* graph to recognize all forms of the verb to meet that are preceded by a word starting with uppercase. We will name this graph *meetLeftRight.grf*:

- 1. Open the *meetLeft.grf* graph.
- 2. Open the FSGraph/Save as... menu to rename the graph as meetLeftRight.grf.
- 3. Click on the *I+you+he+she+it+we+they* box and type <*FIRST*>.
- 4. Select the *ADV* box and the empty box by drawing the rectangle over these boxes, then type *Delete* and *Enter*.
- 5. Click on the *<FIRST>* box and the *\$** box.
- 6. Rearrange the graph.
- 7. Click the Save button.



Now, the concordance displays only twenty-one lines.

⁴⁵ The uppercase and lowercase letters are defined in the Alphabet.txt file, see <u>Section 3.3.3.1</u>, page 41.

⁴⁶ There are some other masks, see the Unitex user manual, Section 4.3.1. Beware that for the preprocessing, the topic of the next section, different masks are valid (see Section 2.5.2 in the Unitex user manual).

Concordance: U:\Documents\exMyUnitex\English\Corpus\DombeyAn	dSon\DombeyAndSon_snt\concord.html 🗾 🗖
21 matches	
d no cry of distress. No, no. Mr Dombey	had met with an accident in ridi
you, at the dead of night. For this, I	have met you here, in a false nar
wandered past outside; and God knows I	have met with my reply. Oh mothe:
e to face for the last time. Wretch! We	meet tonight, and part tonight. 1
n him now, a smile was on her face. "We	meet and part tonight," she said
tell you Misses Brown." "Ay, ay, ay! To	meet at an appointed place," chu
Florence, my sweet girl, farewell!" "To	meet again!" cried Florence. "Ne
inted time, Cousin Feenix and Mr Dombey	meet, and go down to Brighton, a
o her to pat his horse's neck, Florence	meeting his eyes, saw, rather the
ce where they had met them first. After	meeting them thus, three or four
hardly turn him from the subject." But	meeting the eyes of Florence, whi
of the Commander of the Cautious Clara	met with no response to its impul
out, and up towards her room, Florence	met her immediately, coming down
es, but on her way to the door Florence	met the visitor coming in. With
back, bent her dark glance upon him. He	met it with his blandest and most
. "Here he is again!" said Solomon. "We	met Mr Carker the Junior in the
n he made allusion to him, and Diogenes	met that overture graciously, ha.
and caresses to the last. Cousin Feenix	met her at the door, and took he
red to love you, Florence. Good-bye. we	shall meet again soon. Good-bye!
errupted. four moments are precious. we	shall meet you as yourd on your i
e evening, at the old time and place, i	SHALL MEEL YOU AS USUAL ON YOUR (
∥	

3 The .snt file

3.1 Document transformation

3.1.1 Standard normalization

Before splitting the text into tokens (see <u>Chapter 2, Section 1.2.3</u>, page 13), Unitex normalizes the text. This new text is saved with the same name as the original text, but with the *snt*. extension. It is constructed as follows:

- 1. Any sequence of spaces and tabulation characters is replaced by a single space (recognized by a path between two boxes).
- 2. Any series of line breaks (possibly with spaces and tabs) is replaced by a single line break (also recognized by a path between two boxes).⁴⁷
- 3. Braces are replaced by square brackets, unless they correspond to a dictionary entry (a *lexical tag*, see <u>Section 1.2.5</u>, page 15).

3.1.2 The Norm.txt file

The Norm.txt file in the Unitex distribution is important:

 $MyUnitex \ English \ Norm.txt$

We are going to open this file using the *File Edition/Open/Other files* menu. It only contains two lines.

\boxtimes
•

⁴⁷ The *No separator normalization* option, in the *Preprocessing* dialog box, only works for preprocessing graphs. For other graphs, a path on a graph between two boxes may pass not only through one space or one line break, but also through a series of spaces, tabs and line breaks.

These two lines define the rule for transforming braces into brackets. Each line consists of a character string (the text to be replaced), followed by the tabulation character, then another character string (the replacement text) and finally a line break (*Enter*).

Note that this file is strictly speaking not necessary, because this rule applies anyway, even if the *Norm.txt* file is not present. It is actually there to indicate how to add more lines.

3.1.2.1 The standardization of dashes and quotation marks

There are two types of dashes in our text: 2 034 *hyphens* (-) and 2 378 *em dashes* (-). Similarly, there are three types of quotation marks in our text: 7 766 *left double quotation marks* ("), 7 824 *right double quotation marks* (") and 2 *quotation marks* (").⁴⁸ If we would like to normalize them to hyphens (-) and quotation marks ("), we can add three lines to our *Norm.txt* file:⁴⁹

- 1. Open the *Norm.txt* file.
- 2. Type —, Tab, and Enter.⁵⁰
- 3. Type ", Tab, " and Enter.
- 4. Type ", Tab, " and Enter.
- 5. Click the Save button.



Now, close Unitex and reopen it. Open the text (the original *.txt* text, not the already processed *.snt* text) and skip all preprocessing to get 4 412 hyphens (-) and 15 592 quotation marks ("). However, if we look at the number of different tokens, we can notice that we have three unique tokens less than when processing the text with the old *Norm.txt* file: 16 450. Three unique characters were replaced with two existing unique characters.

3.1.2.2 The standardization of apostrophes

Looking closer, we see that the novel uses the *right single quotation mark* (') as an apostrophe and not the keyboard character, the *punctuation apostrophe* ('). The novel also uses the left single quotation mark ('). If we want to normalize these two single quotation marks to punctuation apostrophe, we can add two new lines to our *Norm.txt* file:

- 1. Open the Norm.txt file.
- 2. Type', Tab, ' and Enter.
- 3. Type ', Tab, ' and Enter.
- 4. Click the *Save* button.

⁴⁸ There are more than two types of dashes and three types of quotation marks according to the Unicode standard; however only those mentioned here appear in our text. If you do not know how to type these characters, the best thing is to locate them in the text and copy and paste them.

⁴⁹ In order to apply the modified *Norm.txt* file, it is necessary to close Unitex and then to reopen it.

⁵⁰ For example, by copying and pasting non-ASCII characters from the novel.

Chapter 2: First steps



Now, close Unitex and reopen it. Open the text (the original *.txt* text, not the already processed *.snt* text) and skip all preprocessing.

DombeyAndSon.snt (U:Documents/WyUnitex/English/Corpus/DombeyAndSon)	້ 🗹
0 sentence delimiter, 803028 (16448 diff) tokens, 363065 (16434) simple forms, 0 (0) digit	
Dombey and Son by Charles Dickens CTAPTER 1. Dombey and Son. Dombey sat in the corner of the darkened room in the great arm-chair by the bedside, and Son lay tucked up warm in a little basket bedstead, carefully disposed on a low settee immediately in front of the fire and close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. Dombey was about eight-and-forty years of age. Son about eight-and-forty minutes. Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be prepossesing. Son was very bald, and very red, and though (of course) an undeniably fine infant, somewhat crushed and spotty in his general effect, as yet. On the brow of Dombey, Time and his brother Care had set some marks, as on a tree that was to come down in good time-remorseless twins they are for striding through their human forests, notching as they go-while the countenance of Son was crossed with a thousand little creases, which the same deceiful Time would takk delight in smoothing out and wearing away with the flat part of his scythe, as a preparation of the surface for his deeper operations. Dombey, exulting in the long-locked-for event, jingled and jingled the heavy gold watch-chain that depended from below his trim blue coat, whereof the buttons sparkled phosphorescently in the feeble rays of the distant fire. Son, with his little fists curled up and clenched, seemed, in his feeble way, to be squaring at existence for having come upon him so	
"The House will once again, Mrs Dombey," said Mr Dombey, "be not only in name but in fact Dombey and Son;" and he added, in a tone of luxurious satisfaction, with his eyes half-closed as if he were reading the name in a device of flowers, and inhaling their fragrance at the same time; "Dom-bey and Son!" The words had such a softening influence, that he appended a term of endearment to Mrs Dombey's name (though not without some hesitation, as being a man but little used to that form of address): and said, "Mrs Dombey, my-my dear." A transient flush of faint surprise overspread the sick lady's face as she raised her eyes	•

If we compare the .snt text from <u>Section 1.2.3</u>, page 13, we can see that five tokens are missing (16 453 different tokens versus 16 448).

3.2 The preprocessing

Unitex offers two possibilities of preprocessing:

- Switch on one or two graphs, one working in Merge mode and one in Replace mode. These graphs can recognize a line break, which is represented by <^> in a box. After normalization, the line break will not be distinguished from the space and will be recognized by a path between two boxes.
- 2. Launch the default dictionaries (see <u>Section 1.2.4</u>, page 14).

By default, switching on *Merge* mode will use the graph:

 $MyUnitex \ English \ Graphs \ Preprocessing \ Sentence \ Sentence.grf$

while switching on *Replace* mode will use the graph:

MyUnitex\English\Graphs\Preprocessing\Replace\Replace.grf

3.3 Material (only for confident users)

3.3.1 The Sentence.grf graph

The *Sentence.grf* graph inserts the Unitex symbol {*S*} (*sentence delimiter*) at the end of the sentence.

We will open the file:

MyUnitex\English\Corpus\UnitexGettingStarted\UnitexGettingStarted\ DombeyAndSon\DombeyAndSon.txt by using the *Text/Open...* menu and then answer *Yes* to the question *Do you want to preprocess the text?* By default, two boxes are checked. If we check only the box next to *Apply graph in MERGE mode*



Preprocessing & Lexical parsing	×	
Preprocessing		
No separator normalization (allows preprocessing graphs to match multi-separators)		
Apply graph in MERGE mode: Jlish\Graphs\Preprocessing\Sente	ence\Sentence.grf S <u>e</u> t	
Apply graph in REPLACE m nglish\Graphs\Preprocessing\Re	place\Replace.grf <u>S</u> et	
Tokenizing The text is automatically tokenized. This operation is language-dependant, so that Unitex can handle languages with special spacing rules.		
Lexical Parsing		
Apply All default Dictionaries		
Analyse unknown words as free compound words (this option Cancel but token		
is available only for Dutch, German, Norwegian & Russian)		
Construct Text Automaton	Cancel and close text	

the displayed information for the file:

MyUnitex\English\Corpus\UnitexGettingStarted\UnitexGettingStarted\ DombeyAndSon\DombeyAndSon.snt

will change.



Unitex has inserted 16 515 sentence delimiters.⁵¹

A *sentence delimiter* can be placed and recognized in a box of a graph. For instance, create the *sentenceWithoutDot.grf* graph and record it in the same folder as the *Meet.grf* graph:

- 1. Open the FSGraph/New menu.
- 2. Click on the starting box, then move the cursor to the right, click the right mouse button and select *Create box* from the displayed menu.
- 3. Click on the newly created box and type ?+!, followed by Enter.
- 4. Click on the ?+! box, then move the cursor to the right, click the right mouse button and select *Create box* from the displayed menu.
- 5. Click on the newly created box and type {*S*}, followed by *Enter*.
- 6. Click on the {S} box and on the final box.
- 7. Click the *Save* button.

⁵¹ Since the last sentence is not followed by the *sentence delimiter*, there are therefore 16516 recognized sentences. The number of tokens is now 803 028+16515=819543, because {S} counts as one token. There is also one more different token.



In order to produce a concordance with this graph, we will open the *Text/Locate Pattern* menu and click the *Set* button next to the *Graph* field. The *Graphs* folder will open, and we will move to the folder

MyUnitex\English\Graphs\UnitexGettingStarted\Meet

where we will find and select *sentenceWithoutDot.grf*. The search with this graph has more than 200 matches; therefore, we must modify the *Search limitation* option by checking *Index all occurrences in text*. When we click the *SEARCH* button, we obtain 1316 sentence delimiters preceded by a question mark or an exclamation mark.



Result Info	×
1316 matches	
2632 recognized units	
(0.321% of the text is covered)	
OK	
OK	

If we select the *Regular expression* option and type ?+! in the search field, we obtain 4 469 question marks or exclamation marks in the text, difference due to the presence of direct speech.

Locate configuration	Advanc	ed options
Locate pattern in the fo	rm of:	
Regular expression	n:	
2+!		
Graph:		Set
Activate debug mod	te	
testes:		
Index	Gramm	ar outputs
Index Shortest matches	Gramm Are r	ar outputs not taken into account re with input text
Index Shortest matches Longest matches All matches	Gramm Are r Merg Rep	ar outputs not taken into account ge with input text lace recognized sequences
Index Shortest matches Longest matches All matches Search limitation	Gramm Are r Merg Rep	ar outputs not taken into account ge with inputted lace recognized sequences
Index Shortest matches I Longest matches All matches Search limitation Stop after 200	Gramm Are r Merg Rep matches	ar outputs not taken into account pe with input text lace recognized sequences SEARCH
Index Shortest matches Longest matches All matches Search limitation Stop after 200 Index all occurrence	Gramm Are r Merg Rep matches es in text	ar outputs not taken into account ye with input text lace recognized sequences SEARCH
Index Shortest matches Longest matches All matches Search limitation Stop after 200 e Index all occurrenc Search algorithm:	Gramm Are r Merg Rep matches es in text	ar outputs not taken into account ge with input text lace recognized sequences SEARCH
Index Shortest matches Longest matches All matches Search limitation Stop after 200 Index all occurrenc Search algorithm. Paumier 2003, wor	Gramm Gramm Merg Rep matches es in text king on te	ar outputs not taken into account e with input text lace recognized sequences SEARCH yt (quicker)

Result Info	×
4469 matches 4469 recognized units (0.545% of the text is covered)	
OK	

This may seem confusing, but if we look at the following sequence: *Dear Mama! what is...* (among the first matched sequence in the concordance), we can see that the *sentence delimiter* is not inserted after each question or exclamation mark by the applied *Sentence.grf*. The users may modify the sentence graph supplied by the Unitex distribution or write their own sentence graph.

3.3.2 The Replace.grf graph

The *Replace.grf* graph modifies the text by replacing the contracted verb forms with the corresponding full forms, for instance *I'm* is replaced by *I am*.

We will open the file

```
MyUnitex\English\Corpus\UnitexGettingStarted\UnitexGettingStarted\
DombeyAndSon\DombeyAndSon.txt
```

by using the *Text/Open...* menu and then answer *Yes* to the question *Do you want to preprocess the text?* If we check only the box next to *Apply graph in REPLACE mode*



the displayed information for the file:

MyUnitex\English\Corpus\UnitexGettingStarted\UnitexGettingStarted\ DombeyAndSon\DombeyAndSon.snt

does not change from the text without preprocessing (<u>Section 3.1.2.2</u>, page 36), because the number of tokens remains the same; however, frequencies of some tokens change. We can see, at the thirteenth paragraph, the first transformation: *couldn't* is replaced by *could not*.

DombeyAndSon.snt (U:Documents\MyUnitex\English\Corpus\DombeyAndSon)	ť 🛛
0 sentence delimiter, 803028 (16434 diff) tokens, 363065 (16420) simple forms, 0 (0) digit	
"Have you any fears for Walter's safety?" inquired Florence, from whose face the Captain (so enraptured he was with it) could not take his eyes: while she, in her turn, looked earnestly at him, to be assured of the shorerity of his reply. "No, Heart's-delight," said Captain Cuttle, "I am not afeard. Wal"r is a lad as'll go through a deal o' hard weather. Wal"r is a lad as'll bring as much success to that "ere brig as a lad is capable on. Waltr," said the Captain, his eyes glistening with the praise of his young friend, and his hook raised to announce a beautiful quotation, "is what you may call a out'ard and visible sign of an in'ard and spirited grasp, and when found make a note of." Florence, who did not quite understand this, though the Captain evidently thought it full of meaning, and highly satisfactory, mildly looked to him for something more. "I am not afeard, my Heart's-delight," resumed the Captain, "There is been most uncommon bad weather in them latitudes, there is no denyin', and they have drove and drove and been beat off, may be t'other side the world. But the ship's a good ship, and the lad's a good lad; and it ain't easy, thank the Lord," the Captain made a little bow, "to break up hearts of oak, whether they are in brigs or buzzums. Here we have 'em both ways, which is bringing it up with a round turn, and so I ain't a bit afeard as yet." "As yet?" repeated Florence.	
"Not a bit," returned the Captain, kissing his iron hand; "and afore I begin to be, my Hearts-delight, Wal"r will have wrote home from the island, or from some port or another, and made all taut and ship-shape." And with regard to old Sol Gills, here the Captain became solemn, "who I will stand by, and not desert until death do us part, and when the stormy winds do blow, do blow, do blow-overhaul the Catechism," said the Captain parenthetically, "and there you will find them expressions-if it would console Sol Gills to have the opinion of a seafaring man as has got a mind equal to any undertaking that he puts it alongside of, and as was all but smashed in his "prenticeship, and of which the name is Bunsby, that "ere man shall give him such an opinion in his own parlour as'll stun him. Ah!" said Captain Cuttle, vauntingly, "as much as if he would gone and knocked his head again a door!" "Let us take this gentleman to see him, and let us hear what he says," cried Florence. "Will	•

The both sequences have the same number of tokens, three; that is the reason that the number of tokens in the whole novel has not changed. On the other hand, *could* already existed in the text, so the number of different tokens decreases (16 448 to 16 434).

3.3.3 The two alphabet files

3.3.3.1 The Alphabet.txt file

Recall (see <u>Section 1.2.3</u>, page 13) that a token is either a string of letters or any other single character. What is considered a letter is defined by the *Alphabet.txt* file:

$MyUnitex \ English \ Alphabet.txt$

Since there are two cases in English, all letters are presented in pairs: uppercase and lowercase. We can open this file using the *File Edition/Open/Other Files* menu.

U:\Documents\WyUnitex\English\Alphabet.bt	r 🛛 🗆
File Edit	
Aa	-
Aá	
Ââ	
Aă	
CÇ	
çç	
Dd	
Le PA	
Eè	
Èè	
	_
Ff	
Gg	
Hh Tr	
II	
II .	
LI CONTRACTOR CONTRA TOR CONTRA TOR CONTRA TOR CONTRA TOR CONTRA TOR CONTRA C	
Mm	-

The pairs mean that a word in a box written in lowercase recognizes the same word with uppercase letters in a text. The reverse is not true; that is, if we place uppercase letters in a

box, only words with uppercase at that position are recognized. It is also possible to enforce case-sensitive matching using quotation marks.⁵²

This box	recognizes
	word, Word, WORD, but also wOrd, etc.
	<i>Word, WORD, WOrd,</i> but not <i>word, wORD,</i> etc.
	WORD and nothing else
	Word and nothing else

The same is true when a dictionary is consulted; the entry:

word,.N:s

matches words word, Word, WORD, but also wOrd, etc., while the imaginary entry

WORD,.N:s

would match only the word WORD.

3.3.3.2 The Alphabet_sort.txt file

The Alphabet_sort.txt file:

$MyUnitex \ English \ Alphabet_sort.txt$

defines the sort order of the letters of the alphabet. It is used to sort concordances and dictionaries, when presented in alphabetical order. The alphabetical sort implemented in Unitex is very fast, even on very large files. We will open this file by using the *File Edition/Open/Other Files* menu.



⁵² Or, as we have seen in <u>Section 2.4.3</u>, page 34, by using lexical masks.

In this file, letters are also represented in pairs, uppercase and lowercase. The sort order is defined in two levels. The first level is defined by lines, meaning that, for instance, each A is preceding each B, whether it is written in uppercase or not, whether it is accented or not. The second level is defined by each line. Consider the line:

EÉÈÊËeéèêë

It imposes that an *E* without an accent precedes an *É* with an acute accent, and also that an *É* with an acute accent precedes an *È* with a grave accent, and so on. For instance, in French, $P\hat{E}CHER$ or $p\hat{e}cher$ (to fish) and $P\hat{E}CHER$ or $p\hat{e}cher$ (to sin) are ranked as follows:

PÉCHER, PÊCHER, pécher, pêcher

However, the whole group of letters has its own order in relation to other groups of letters. Thus, the words *ÉTÉ* or *été* (*summer*) and *ÈRE* or *ère* (*era*) and *EUROPE* or *Europe* are ranked:

ÈRE, ère, ÉTÉ, été, EUROPE, Europe

because the *Rr* group precedes the *Tt* group and the *Tt* group precedes the *Uu* group.

Chapter 3: corpus annotation

 Before starting this chapter

 Go to the folder:
 MyUnitex\English\Graphs\UnitexGettingStarted

 Create three new folders named Will, Numbers and RomanNumerals.

1 First example: compound verbs using the verb will

We will start this chapter by opening the text *DombeyAndSon.txt* (*Text/Open* menu) in the folder:

and by answering Yes to the question *Do you want to preprocess the text*. A window will open in which we will remove the first two checkmarks, leaving just the last one *Apply All default Dictionaries*.



Preprocessing & Lexical parsing				
Preprocessing				
No separator normalization (allows preprocessing graphs to match multi-separators)				
Apply graph in <u>M</u> ERGE mode:	lish\Graphs\Preprocessing\Sente	ence\Sentence.grf	S <u>e</u> t	
Apply graph in REPLACE m	nglish\Graphs\Preprocessing\Replace\Replace.grf		<u>S</u> et	
Tokenizing				
The text is automatically tokenized. This operation is language-dependant, so that Unitex can handle languages with special spacing rules.				
Lexical Parsing				
Apply All default Dictionaries				
Analyse unknown words as free compound words (this option			nize text	
is available only for Dutch, German, Norwegian & Russian)				
Construct Text Automaton Cancel and close text			ose text	

1.1 First step: the simplest graph

Let us create a first graph that we will save, like all graphs in this chapter, in the folder:

MyUnitex\English\Graphs\UnitexGettingStarted\Will

We will name it *willV.grf*. This graph will recognize all forms of the verb *will* (forms *will* and *would*) followed by a verb in the infinitive (the future tense) and insert the XML tags <*verb>...</verb>* for the recognized forms. To insert an annotation into a box, add a trailing slash and type the desired annotation. This annotation will be placed just before the text recognized by the box. To place an annotation after the box, you must therefore add an empty box after it (with the annotation).

Generally, the lexical mask *«will»* represents all forms of the lexeme *will*. However, there are two lexemes *«will»*, a verb (*will*) and a noun (*a will, wills*). In order to avoid the recognition of nominal forms of *will*, we have to precise its part-of-speech with an appropriate code *«will.V»*.⁵³

⁵³ Pay attention to the fact that information inside a lexical mask is case sensitive.

In the English dictionary distributed with *Unitex*, the code for infinitive forms of verbs is *W*. The lexical mask that recognizes an infinitive form of any verb is *<V:W>*.



Start the search (*Text/Locate patterns*) using the *willV.grf* graph (*Set* button in the *Graph* line) with options *Merge with input text* and *Index all occurrences in text*. Click the *Search* button and then click the *OK* button. We get 1 146 matches. To make the concordance easier to read, we are going to slightly modify the display parameters by choosing to display 20 characters on the left and 255 characters on the right of the matched sequences. Finally, click on the *Build concordance* button.

🗌 Locate Pattern				I	3						
Locate configuration Advanced options				Located se	quences			X			
Locate pattern in the form of:				Concordance	Statistics						
O Regular expression:				Modify text				_			
							Resulting .txt f	ile:			
Graph: MyUnitex\E	nglish\Gra	aphs\UnitexGett	ingStarted\W	ill\willVK.grf Set	Ĩ		Se	t File	GC)	
Activate debug mod	e						Extract units				
					-		Set File:				
							Extract ma	tching units	Extract unmat	tching units	
							Concordance	presentation			
Index	Gramma	ar outputs					🔲 Use a web	browser to view	the concordance	Э	
 Shortest matches 	O Are n	ot taken into acc	count				Show	differences with	previous concore	dance	
Longest matches	Merge	e with input text						Show ambig	uous outputs		_
 All matches 	Replay	ace recognized	sequences				Show matchin	a sequences in	context		
Search limitation							Context leng	th: Stop at:	Sort according to	0:	
Stop after 200	matches		SEARC	н			Left 20	chars 🗌 (S)	Center, Left		-
Index all occurrence	s in text						Right 255	chars 🗌 (S)			
Search algorithm:									Build conc	ordance	
Paumier 2003, work	ting on tex	t (quicker)									
										_	
	Con	ncordance: U:\D	ocuments\M	/Unitex\English\Corp	us\DombeyAndSon\Don	nbeyAndSon_s	nt\concord.html		r 🛛 🖂	Ĵ.	
	1146 m	atches									
	nd,	that unle	ess you	<verb>will</verb>	accept	≥ this va	igue caut	ion from	one who		
	ully	indulged	d in, I	<verb>will</verb>	<u>add</u> af	fter the	gratific	ation of	ambitio		
	ss g	oodwill.	"All I	<pre><verb>will </verb></pre>	<u>add</u> is	, Heaver	1 forbid	that I sh	ould to		
	the sense	world who	tances	<pre><verb>will </verb></pre>	admit	of This	one rese is the	moies you last time	, Maram		
	rien	d Dombev.	, if he	<pre><verb>will</verb></pre>	allow	me to ca	ll him s	o." Cousi	n Feenia		
	ft a	lone, and	d if he	<verb>will</verb>	allow	me-not t	o take W	alter's p	lace, f		
	and y	would, Si	ir, you	<pre><verb>will</verb></pre>	allow	Bagstoc)	to add-	and would	l-give h		
	sine	ss matter	rs. You	<verb>will</verb>	allow	me to se	e you ag	ain, and	to consu		
	shin	g Took "A	And you	<verb>will</verb>	allow	me, perh	aps, to	apologise	, in hi		
	he s	econd, the	nat you	<pre><verb>will </verb></pre>	allow	me somet	imes, sa	y every M	londay m		
	rave	n images.	, and T	<pre><verb>will</verb></pre>	and mu	ist speak	'Uu, as y	bere anvb	ody the	l I	
	emen	ts, thus:	"Last	<verb>Will</verb>	and Te	estament	of Solom	on Gills.	" "Lette		
	ng r	esist his	s lofty	<verb>will</verb>	and pl	leasure!	He had r	esolved t	o conque		
	atio	n than th	ne good	- <u><verb>will</verb></u>	and fi	ixtures o	of a publ	ic-house;	and be:		
	d be	lieve tha	at they	<verb>will</verb>	arrive	 I made 	this li	ttle gift	for Pau		
	ther	Dombey a	and Son	<verb>will</verb>	ascend	-no 'ri	.se;' tha	t was Mrs	Toots'		
	s of	influence	ing, I	<pre><verb>will </verb></pre>	assist	you to	accompli	sh the de	sign, oi		
	pear	the shar	ne that	<verb>will</verb>	attach	to me-1	esolved	to know t	nat it a		

You will see that at the beginning of the concordance list there are some false recognitions. Namely, in a text segment *could I do more than render up my whole will and being to you, will* is a form of a noun and *and* is a conjunction. This is because, in the English dictionary distributed with *Unitex* a form *will* has two interpretations: a noun and a verb, while *and* has

has happened here, <u><verb>will be</verb></u> a warning to you, Towlinson." "Than sary. One of these, <u><verb>will be</verb></u> the engagement at Brighton (if it i

country during the total that total the total the

also two interpretations: a conjunction and a verb.⁵⁴ We will see in the section devoted to dictionaries how some cases of ambiguities can be avoided.

1.2 Second step: future passive and future perfect tense

When looking at the produced concordance we observe that in several cases *will be* and *will have* were recognized as a form of future tense, but the recognition was not complete. Now we will modify our graph so that it recognizes also passive tenses and forms of the future perfect tense. In the English dictionary distributed with *Unitex*, past participles are encoded with *K* and a lexical mask for retrieval of past participle word forms is *<V:K>*. In order to extend our graph, we will add a new path that exits from the *<will:V>* box. In this path, a form of a past participle must be preceded by *be* or *have*. Note that in this case, we use strings and not lexical masks, because these are exact forms that have to be recognized. We could have also used *<be.V:W>* and *<have.V:W>* but these masks correspond to exactly one form each.



Open the *Text/Locate Pattern…* menu and click the *Search* button. We will obtain again 1 146 matches. Instead of producing concordances, we will click the *Show differences with previous concordances* button in the *Locate sequences* dialog box. We will obtain the parallel list of matches obtained by the previous graph and matches obtained by the modified graph.

Concordance	Statistics			
Modify text				
Resulting .txt fil	e:			
Set	File	GO		
Extract units				
Set File:				
Extract mat	ching units	Extract unmatching units		
Concordance presentation				
Concordance p	resentation			
Concordance p	resentation prowser to view	the concordance		
Concordance p	resentation prowser to view lifferences with	r the concordance previous concordance		
Concordance p	resentation browser to view lifferences with Show ambig	r the concordance previous concordance wous outputs		
Concordance p Use a web I Show d	resentation prowser to view ifferences with Show ambig sequences in	r the concordance previous concordance uous outputs context		
Concordance p Use a web t Show d Show matching Context length	resentation prowser to view lifferences with Show ambig sequences in n: Stop at	r the concordance previous concordance uous outputs context Sort according to:		
Concordance p Use a web I Show d Show matching Context length Left 20 ch	resentation browser to view ifferences with Show ambig sequences in n: Stop at nars (S)	v the concordance previous concordance uous outputs context Sort according to: Center, Left		
Concordance p Use a web t Show d Show matching Context length Left 20 ct Right 255 ct	ifferences with Show ambig sequences in n: Stop at nars (\$) hars (\$) {\$}	I the concordance previous concordance uous outputs context Sort according to: Center, Left		

Concordance Diff	r' a' 🛛
Violet: identical sequences with differe Red: similar but different sequences Green: sequences that occur in only o Grey background=previous matches	ent outputs ine of the two concordances White background=new matches
es towards him. "He ≤	<pre><verb>will be</verb> christened Paul, m</pre>
es towards him. "He ≤	<pre><verb>will be christened</verb> Paul, m</pre>
no one in the world s	<pre><verb>would have</verb> received it wit</pre>
no one in the world <	<pre><verb>would have received</verb> it wit</pre>
nd books. Mr Dombey ≤	<pre><verb>would have</verb> reasoned: That</pre>
nd books. Mr Dombey s	<pre><verb>would have reasoned</verb>: That</pre>
wback. Yes. That he ≤	<pre><verb>would have</verb> allowed. With o</pre>
wback. Yes. That he ≤	<pre><verb>would have allowed</verb>. With o</pre>
tells us, Mr Dombey ≤	<pre><verb>would have</verb> added in a patr</pre>
tells us, Mr Dombey <	<pre><verb>would have added</verb> in a patr</pre>
pture, if examined, ≤	<pre><verb>would have</verb> been found to b</pre>
pture, if examined, <	<pre><verb>would have been</verb> found to b</pre>
sness of the result	<pre><verb>would have</verb> induced me to t</pre>
sness of the result s	<pre>(verb>would have induced me to t</pre>
me, Master Dombey,' s	<pre><verb>would have</verb> been much more</pre>
me, Master Dombey,' <	<pre><verb>would have been</verb> much more</pre>
sion about him. "It ≤	<pre>(verb>would have occurred to mos</pre>
sion about him. "It s	<pre><verb>would have occurred</verb> to mos</pre>
and-take couple. It <	<pre><verb>would have</verb> been, generally</pre>
and-take couple. It ≤	<pre><verb>would have been</verb>, generally</pre>
out of her name, it ≤	<pre><verb>would be</verb> considered in the</pre>

Open the *Text/Located sequences...* menu and click the *Build concordance* button. You will obtain the new concordance list.

⁵⁴ An intransitive verb with dialectal use, meaning: to breathe; whisper; devise; imagine (source: Wiktionary).

Unitex Getting Started - Denis Maurel and Cvetana Krstev

Concordance: U1Documents/MyUnitex/English\Corpus\DombeyAndSon\DombeyAndSon_snttconcord.html	Concordance: U:DocumentsWyUnitexEnglish\CorpusDombeyAndSon\DombeyAndSon_snftconcord.html
1146 matches	1146 matches
tant. "Dear me, six < <u>verb>will be changed</u> to sixteen, before we have	away from here, you <verb>will have concluded</verb> what is a mere matter o
es towards him. "He <u><verb>will be christened</verb></u> Paul, my-Mrs Dombey-of c	ction; of which you < <u>verb>will have formed</u> a general understanding free
artist himself. He <verb>will be delighted</verb> , I am sure, with Mrs Grane	and the comforts he <verb>will have</verb> about him, and the quiet way in w
nce of to-morrow. I <verb>will be exhibited</verb> to no one, as the refract	are what we are. I < <u>verb>will have</u> no youth and truth dragged down to
in, as your brother <verb>will be gone</verb> out at that hour. I don't ask	down to my level. I <verb>will have</verb> no guileless nature undermined, c
g you give Miss Tox <verb>will be hoarded</verb> and prized, I am sure, like	y right. In short I < <u>verb>will have</u> it. I consider it no unreasonable
en I die, its ghost <verb>will be laid</verb> . As it is clearly no inheritan	ses, perhaps; but I < <u><verb>will have</verb></u> submission first!-Carker," said M
back, that he never <verb>will be muzzled</verb> when Paul Dombey is in ques	arts-delight, Wal"r <verb>will have</verb> wrote home from the island, or from
his impertinence he <verb>will be punished</verb> terribly. You can pretend	o his presence. You <verb>will have</verb> to get the better of that objection
a promise that you <verb>will be reconciled</verb> to Walter-to my dearest 1	ule of my life. You <verb>will have</verb> the goodness to undertake this ch
ence home. Gay, you <verb>will be rewarded</verb> to-morrow." "Oh! thank you	towards myself. You < <u>verb>will have</u> the goodness to let Mrs Dombey know
of opinion that it <verb>will be shorn</verb> of some of its emoluments, and	and Walter; and you < <u>verb>will have</u> no secrets from Susan when she cor
nd I believe nobody <verb>will be surprised</verb> , either at its being requi	Mrs Dombey; and you < <u>verb>will have</u> the goodness, I am persuaded, to <
has happened here, <verb>will be</verb> a warning to you, Towlinson." "Than	nd me-oh, I see you <verb>will have</verb> it out of me, Misses Brown! But t
sary. One of these, <verb>will be</verb> the engagement at Brighton (if it i	an offer it, if you < <u>verb>will have</u> the goodness." "My dear Paul," re
, "may be dreadful, <verb>will be</verb> dreadful; but I feel that I should	ng brother. "If you <verb>will have</verb> the goodness." John Carker, with
se, in other words, <verb>will be</verb> able to hold its own, and maintain :	his greatness, "you <verb>will have</verb> the goodness not to omit or slur
, that reaches you, <verb>will be</verb> always most acceptable to me. That	ved; "I direct." "I < <u>verb>will hold</u> no place in your house to-morrow,
cle in his way, and <verb>will be</verb> happy to assent to any honourable c	wton. "If Mr Dombey < <u>verb>will honour</u> us, we shall be happy. Withers,
s he always was and <verb>will be</verb> !" "Thankee, Jemima," cried the simp	mber. "My love, you < <u>verb>will inform</u> Cornelia, if you please." "Assu
d it always was and <verb>will be</verb> so!" observes the housemaid. It is	e in the morning. I < <u>verb>will inquire</u> myself in the morning. Mama ha
holesome face, "and <u><verb>will be</verb></u> so pleased to see her dear Papa ton	as you like. No one <u><verb>will interfere</verb></u> with you, Florence, or wish
madine - and the second s	in the second combined in the second

You should note that, in the *Locate pattern* dialog box, we used the default option *Index longest matches*, which means that shortest matches at the same position will not be indexed or listed in concordances. In order to understand what it means, open the *Text/Locate pattern...* menu, change the default option to *Index all matches* and click the *Build concordance* button. Now, you will obtain 1 382 matches and among them are both shorter and longer matches, for instance *will be associated* and *will be*. You may wonder why two concordance lines containing these two matches are far apart in the concordance list. When building concordances, we used the default option for sorting concordance lines (*Located sequences/Sort according to*) which is *Center, Left*. This option means that concordances are first sorted according to matches (from left to right) and then according to the left context (from right to left). In this case, the occurrence *<verb>will be associated*.

Concordance: U:Documents/MyUnitex/English/Corpus/DombeyAndSon/DombeyAndSon_snttconcord.html 🛛 🗖 🖾	Concordance: U:Documents/MyUnitex/English/Corpus/DombeyAndSon/DombeyAndSon_snt/concord.html = 2 2 2
1382 matches	1382 matches
nd my daughter; who <verb>will be associated</verb> with the domestic life of	nd my daughter; who < <u>verb>will be</u> associated with the domestic life o
tant. "Dear me, six < <u>verb>will be changed</u> to sixteen, before we have 🗐	in, as your brother < <u>verb>will be</u> gone out at that hour. I don't ask
es towards him. "He < <u>verb>will be christened</u> Paul, my-Mrs Dombey-of c	rranged. The supper <verb>will be</verb> here this moment." These facts the
artist himself. He <verb>will be delighted</verb> , I am sure, with Mrs Gran	me to you, and ever < <u>verb>will be</u> . But what I do is not done for myse
nce of to-morrow. I <verb>will be exhibited</verb> to no one, as the refract	back, that he never <verb>will be</verb> muzzled when Paul Dombey is in ques
in, as your brother <verb>will be gone</verb> out at that hour. I don't ask	them old Skettleses < <u>verb>will be</u> a Godsend." "It is very good of Sir
g you give Miss Tox < <u>verb>will be hoarded</u> and prized, I am sure, like	my bridesmaid, that <verb>will be</verb> glad of your protection, Cap'en Cut
en I die, its ghost <verb>will be laid</verb> . As it is clearly no inheritan	rly attended to, it <verb>will be</verb> through her powerful intellect-Susa
back, that he never <verb>will be muzzled</verb> when Paul Dombey is in ques	ny other reason. It <verb>will be</verb> necessary that you should enable me
his impertinence he <verb>will be punished</verb> terribly. You can pretend	ourselves of it. It <verb>will be</verb> our own faults if we lose sight of
a promise that you <verb>will be reconciled</verb> to Walter-to my dearest i	to you, of which it < <u>verb>will be</u> her pleasure and privilege to garne:
ence home. Gay, you <verb>will be rewarded</verb> to-morrow." "Oh! thank you	t to go. I think it < <u>verb>will be</u> better." "Much better," interposed :
of opinion that it <verb>will be shorn</verb> of some of its emoluments, and	nfeignedly, that it <verb>will be</verb> a great comfort to me, and that what
nd I believe nobody <verb>will be surprised</verb> , either at its being requ	of opinion that it < <u><verb>will be</verb></u> shorn of some of its emoluments, and
has happened here, <verb>will be</verb> a warning to you, Towlinson." "Than	he is confident it <verb>will be</verb> ; for your continued show of affection
sary. One of these, <verb>will be</verb> the engagement at Brighton (if it i	nish the comfort it <verb>will be</verb> to him and me, when that time comes
, "may be dreadful, <verb>will be</verb> dreadful; but I feel that I should	oughtfully: "but it <verb>will be</verb> right to go. I think it will be bet
se, in other words, <verb>will be</verb> able to hold its own, and maintain	m so relieved." "It < <u>verb>will be</u> a greater relief and delight still,
, that reaches you, <verb>will be</verb> always most acceptable to me. That	, "if we live." "It < <u>verb>will be</u> anything but shrewd in you, my good
cle in his way, and <verb>will be</verb> happy to assent to any honourable c	," he returned, "it <verb>will be</verb> that I may hear tidings of you. Rel
s he always was and <verb>will be</verb> !" "Thankee, Jemima," cried the simp	him up quickly, "it < <u>verb>will be</u> of still less use, Gay, if you pers
d it always was and <verb>will be</verb> so!" observes the housemaid. It is	times." "The visit <verb>will be</verb> tedious to you, I am afraid." "Oh na
he locate from the sector of the sector of the sector beauting to th	Hot - and had starting the contract of the starting the s

1.3 Third step: recognizing the future perfect passive

We will modify our graph to recognize more compound verb forms with the lexeme *<will.V>*. Now we must split the *be+have* box because the forms *been* and *had* can follow *have* but not be.



Open the *Text/Locate Pattern…* menu, return to the original indexing option: *Longest matches* and click the *Search* button. By clicking the *Build concordances* button in the *Locate sequences* dialog box, you will obtain the new concordance. This list again contains 1 146 lines. Note that the added forms *been* and *had* are also past participles of verbs *to be* and *to have* respectively, so now some longer (complete) compound forms are recognized instead of shorter ones: for instance, would have been (before) is replaced by would have been added (now).⁵⁵

Concordance: U:\Documents\Mv	Unitex/English/Corpus/DombevAndSon/DombevAndSon_spt/concord.html
1146 matches	
of his second wife	<pre><verb>would have been added</verb> to his own-would [</pre>
pture, if examined,	<pre><verb>would have been found</verb> to be; that as for</pre>
er time the Captain	<pre><verb>would have been overbalanced</verb> by such a</pre>
joy. "I thought you	<pre><verb>would have been pleased</verb>," said Polly. "</pre>
of it from her. She	<pre><verb>would have been racked</verb> to death, and tu:</pre>
n as "breaking up,"	<pre><verb>would have been</verb> quite inapplicable to tl</pre>
woman like herself,	<pre><verb>would have been</verb> prostration, and that sl</pre>
d never seen again,	<pre><verb>would have been</verb> sufficient stain upon he</pre>
ews, were upon her,	<pre><verb>would have been</verb> sufficient embarrassed)</pre>
her with his coat,	<pre><verb>would have been</verb> worth a hundred state s:</pre>
dge of his mystery,	<pre><verb>would have been</verb> at all likely to unders</pre>
me, Master Dombey,'	<pre><verb>would have been</verb> much more congenial to r</pre>
but that he indeed	<pre><verb>would have been</verb> a very Davy-lamp of ref</pre>
closer to him, and	<pre><verb>would have been</verb> glad to touch his rugge</pre>
old Blimber's, and	<pre><verb>would have been</verb> now, if he'd have lived</pre>
r introduction; and	<pre><verb>would have been</verb> exactly like the first,</pre>
mpulsive of mankind	<pre><verb>would have been</verb> a milder enemy to encour</pre>
no doubt my advice	<pre><verb>would have been</verb> cavalierly received, and</pre>
orning's confidence	<pre><verb>would have been</verb> Mr Dombey's last. Yet e</pre>
yes. Happy Florence	<pre><verb>would have been</verb>, might she have venture</pre>
often what her life	<pre><verb>would have been</verb> if her father could have</pre>
for the militia, he	<pre><verb>would have been</verb> glad to do so, on libert</pre>
	•

1.4 Fourth step: recognizing continuous forms

In order to recognize continuous forms: future continuous and future perfect continuous, we will modify our graph again. Continuous forms are built using the present participle for which we use the lexical mask <*V*:*G*>. This mask will be used in two new boxes, once for the future continuous tense and once for the future perfect continuous tense. We must split the *been+had* box because the present participle can follow *will have been*, but not *will have had*.

⁵⁵ You can analyze all differences obtained by this new graph if you use the *Show differences with previous concordances* option in the *Located sequences* dialog box.



Now the concordance highlights four new sequences:⁵⁶ will be giving, would be dancing, would be enchanting and would be obliging.

🗇 Concordance: U1Documents/MyUnitex/English1Corpus/DombeyAndSon/DombeyAndSon_snftconcord.html 🗖 🖉 🔯	📋 Concordance: U:DocumentsMyUnitexEnglishlCorpusIDombeyAndSonIDombeyAndSon_snficoncord.html 🗖 🗖 🛛		
1146 matches	1146 matches		
it. My constitution <verb>will be giving</verb> way next. Besides, I had a v.	in the golden water <verb>would be dancing</verb> on the wall. He was visite		
in, as your brother <u><verb>will be gone</verb></u> out at that hour. I don't ask	ersation. Mr Dombey < <u>verb>would be delighted</u> to leave Florence in suc		
g you give Miss Tox <verb>will be hoarded</verb> and prized, I am sure, like	n!" "Any expedition <verb>would be enchanting</verb> in such society," return		
en I die, its ghost <verb>will be laid</verb> . As it is clearly no inheritan	understand that he <verb>would be expected</verb> to inform Doctor and Mrs 1		
back, that he never <verb>will be muzzled</verb> when Paul Dombey is in quest	the admiration that <verb>would be felt</verb> for her by all the boys, about		
his impertinence he <verb>will be punished</verb> terribly. You can pretend	y unlikely that she <verb>would be found</verb> in communion with the Establi		
a promise that you <verb>will be reconciled</verb> to Walter-to my dearest 1	g way, I believe it < <u>verb>would be found</u> out." "Ah, so sharp as that,		
ence home. Gay, you <verb>will be rewarded</verb> to-morrow." "Oh! thank you	en here any more he <verb>would be given</verb> into custody; and to keep to		
of opinion that it <verb>will be shorn</verb> of some of its emoluments, and	how soon her money <verb>would be gone</verb> -too much a child in worldly m		
nd I believe nobody <verb>will be surprised</verb> , either at its being requ	on, wheezing sounds <verb>would be heard</verb> ; and the Major's blue would <		
has happened here, <verb>will be</verb> a warning to you, Towlinson." "Than	y of Dombey and Son <verb>would be heightened</verb> and maintained, indeed,		
sary. One of these, < <u>verb>will be</u> the engagement at Brighton (if it i	rticle of furniture <verb>would be improved</verb> by the contents. The room		
, "may be dreadful, <verb>will be</verb> dreadful; but I feel that I should t	et corner where she <verb>would be left</verb> , until the clock struck three		
se, in other words, <verb>will be</verb> able to hold its own, and maintain	e that such conduct < <u>verb>would be looked</u> upon by Mr Dombey as presum		
, that reaches you, < <u>verb>will be</u> always most acceptable to me. That	e regretted that he <verb>would be obliged</verb> to return to London next me		
cle in his way, and <verb>will be</verb> happy to assent to any honourable c	op just now, and it < <u>verb>would be obliging</u> me if you would give one 1		
s he always was and <verb>will be</verb> !" "Thankee, Jemima," cried the simp	ering a remark that <verb>would be received</verb> with acquiescence, was gr		
d it always was and <verb>will be</verb> so!" observes the housemaid. It is	meone else. That it <verb>would be received</verb> as a great honour and dis		
holesome face, "and <verb>will be</verb> so pleased to see her dear Papa ton	notion of the ball, <verb>would be scouted</verb> as a malignant idiot. Mrs :		
-rending occurrence < <u>verb>will be</u> a warning to all of us, to accustom	belonging to them, < <u>verb>would be seen</u> gracefully draping a barricad		
and if my presence < <u>verb>will be</u> , in the least degree, unpleasant to	eir darling boy, he < <u>verb>would be touched</u> home, and would say, "Dear		
nd her children, he <verb>will be</verb> most repentant of his own part in the	"Oh! Good morning!" <verb>would be</verb> Mr Toots's first remark to the ser		
· · · · · · · · · · · · · · · · · · ·	The second secon		

1.5 Fifth step: insertion of an adverb after *will* or *would*

1.5.1 A single adverb

Now we will take into account that an adverb can be inserted after the verb *<will.V>*. Since the code used for adverbs in the English dictionary distributed with Unitex is *ADV*, we will use for them the lexical mask *<ADV>*. The tags around them in a text will be *<adverb>...</adverb>*.



Save the modified graph and use it in *Locate pattern*. You will obtain now 1 330 concordance lines. This time, they contain new matches, because occurrences with an adverb after *<will.V>* could not be retrieved with the previous version of the graph. Some of the occurrences can be seen in the following figure.

⁵⁶ The number of matches remains the same because the previous graph recognized *will be*.

Chapter 3: Corpus annotation

Concordance: U:\Documents\My	UnitexiEnglish\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🖬 🗹
1330 matches	
those words of mine	<pre><verb>will <adverb>always</adverb> be</verb> a balm and (*)</pre>
ng you together. He	<pre><verb>will <adverb>certainly</adverb> help</verb> us to </pre>
ely to be dead. You	<pre><verb>will <adverb>easily</adverb> forgive</verb> an old</pre>
nk or hope that you	<pre><verb>will <adverb>ever</adverb> forgive</verb> me. There</pre>
ubt very much if it	<pre><verb>will <adverb>ever</adverb> lead</verb> me to advan</pre>
pe that my poor boy	<pre><verb>will <adverb>ever</adverb> read</verb> these words</pre>
ing that the Doctor	<pre><verb>will <adverb>first</adverb> paint</verb> it, and p</pre>
and look here! "You	<pre><verb>will <adverb>further</adverb> please</verb>, Madam</pre>
g, to Mr Carker. He	<pre><verb>will <adverb>immediately</adverb> take</verb> care</pre>
espected parent. It	<pre><verb>will <adverb>naturally</adverb> be</verb> very pair</pre>
used to do; for he	<pre><verb>will <adverb>naturally</adverb> do</verb> that no r</pre>
dred thousand pound	<pre><verb>will <adverb>nearly</adverb> cover</verb> it. The v</pre>
hen you are old you	<pre><verb>will <adverb>never</adverb> depart</verb> from it,</pre>
hen you are old you	<pre><verb>will <adverb>never</adverb> depart</verb> from it.</pre>
for me, dear, and I	<pre><verb>will <adverb>never</adverb> forget</verb> how kind</pre>
ing that I hope you	<pre><verb>will <adverb>never</adverb> forget</verb> that you</pre>
eauty, and no grass	<pre><verb>will <adverb>never</adverb> grow</verb> upon the g:</pre>
stead, a homage you	<pre><verb>will <adverb>never</adverb> have</verb>." Although</pre>
te manner. That you	<pre><verb>will <adverb>never</adverb> speak</verb> of it to</pre>
, entreats that you	<pre><verb>will <adverb>never</adverb> talk</verb> to her any</pre>
ough paths that you	<pre><verb>will <adverb>never</adverb> tread</verb>, and my water </pre>
rnfully. "To you. I	<pre><verb>will <adverb>not</adverb> add</verb> to myself, as</pre>
Tan and and and Dambarr	

1.5.2 Several adverbs

However, it sometimes happens that an adverb is followed by another adverb or even more. We will add a loop between the two boxes concerned.



The new concordance shows 13 new matches. You can also use the *Show differences with previous concordance* option in the *Located sequences* menu to spot them.

Concordance: U:\Documents\My	/Unitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🖬 🛛
1343 matches	
alse and subtle men	<pre><verb>will <adverb>always</adverb> <adverb>secretly</adverb></verb></pre>
those words of mine	<pre><verb>will <adverb>always</adverb> be</verb> a balm and </pre>
ng you together. He	<pre><verb>will <adverb>certainly</adverb> help</verb> us to a</pre>
ely to be dead. You	<pre><verb>will <adverb>easily</adverb> forgive</verb> an old</pre>
nk or hope that you	<pre><verb>will <adverb>ever</adverb> forgive</verb> me. There</pre>
ubt very much if it	<pre><verb>will <adverb>ever</adverb> lead</verb> me to advance</pre>
pe that my poor boy	<pre><verb>will <adverb>ever</adverb> read</verb> these words</pre>
ing that the Doctor	<pre><verb>will <adverb>first</adverb> paint</verb> it, and p</pre>
and look here! "You	<pre><verb>will <adverb>further</adverb> please</verb>, Madam</pre>
g, to Mr Carker. He	<pre><verb>will <adverb>immediately</adverb> take</verb> care</pre>
espected parent. It	<pre><verb>will <adverb>naturally</adverb> be</verb> very pair</pre>
used to do; for he	<pre><verb>will <adverb>naturally</adverb> do</verb> that no r</pre>
dred thousand pound	<pre><verb>will <adverb>nearly</adverb> cover</verb> it. The v</pre>
hen you are old you	<pre><verb>will <adverb>never</adverb> depart</verb> from it,</pre>
hen you are old you	<pre><verb>will <adverb>never</adverb> depart</verb> from it.</pre>
for me, dear, and I	<pre><verb>will <adverb>never</adverb> forget</verb> how kind</pre>
ing that I hope you	<pre><verb>will <adverb>never</adverb> forget</verb> that you</pre>
eauty, and no grass	<pre><verb>will <adverb>never</adverb> grow</verb> upon the g:</pre>
stead, a homage you	<pre><verb>will <adverb>never</adverb> have</verb>." Although</pre>
te manner. That you	<pre><verb>will <adverb>never</adverb> speak</verb> of it to</pre>
, entreats that you	<pre><verb>will <adverb>never</adverb> talk</verb> to her any</pre>
ough paths that you	<pre><verb>will <adverb>never</adverb> tread</verb>, and my water </pre>
The second secon	Another and the second se

1.6 Sixth step: insertion of an adverb after *be*

The position after the verb *<will.V>* is not the only one where an adverb can occur. It can also be inserted between the verb *to be* and the present or past participle.

This should be added as optional after the *been* and *be* boxes in our previous graph. Therefore, we will modify our graph once again. However, rather than duplicating the boxes and loop of the previous section, we will create a subgraph to recognize adverbs.

1.6.1 A graph that recognizes and annotates adverbs

We observe that adverbs can appear at two positions in a verb group, so the same boxes for recognition and annotation would appear twice. In such situations, it is good to use subgraphs. First, the same problem of recognizing and annotating adverbs can appear in some other graph and then the subgraph can be reused.⁵⁷ Second, the graph that recognizes verb groups with *will/would* becomes less complex and easier to read and to develop further.

The easiest way to create a subgraph named *adverb.grf* from the existing graph is to use the *Export as new graph* option (select boxes that will belong to a subgraph and use the right mouse button). Here we select the *<ADV>/<adverb>* and *<E>/</adverb>* boxes, then we right-click to choose the *Export as new graph* option. We call this graph *adverb.grf*.

The boxes of the main graph, which recognize and annotate the adverbs, must then be replaced by a call to the subgraph (name preceded by a colon, here *:adverb*). In the main graph, boxes that call a subgraph have a gray background.



The concordance obviously remains the same.

1.6.2 The main graph

Let us modify our graph again to recognize occurrences of adverbs after the verb *to be* (in two places).

⁵⁷ See <u>Chapter 4, Section 4.3</u>, page 95 (only for confident users).



We obtain the same number of concordance lines, but longer sequences are recognized and highlighted.

Concordance: U:\Documents\Mv	المتاهيد المراقعة ال
1242 matchas	
1545 matches	
bear the shame that	<pre><verb>will attach</verb> to me-resolved to know that it attaches falsely-that you know it </pre>
fancy to, I'm sure	<pre><verb>will be <adverb>as</adverb> <adverb>readily</adverb> provided</verb> as if you were</pre>
holesome face, "and	<pre><verb>will be <adverb>so</adverb> pleased</verb> to see her dear Papa tonight." "Lork, Mrfg</pre>
us walk." "But you	<pre><verb>will be <adverb>so</adverb> tired</verb>, my love." "Oh no! I was very tired the fit</pre>
u think Miss Dombey	<pre><verb>will be <adverb>very</adverb> <adverb>much</adverb> affected</verb>, Captain Gills-</pre>
nd my daughter; who	<pre><verb>will be associated</verb> with the domestic life of our young Pilgrim to Parnassus 1</pre>
tant. "Dear me, six	<pre><verb>will be changed</verb> to sixteen, before we have time to look about us." "Ten year</pre>
es towards him. "He	<pre><verb>will be christened</verb> Paul, my-Mrs Dombey-of course." She feebly echoed, "Of course."</pre>
artist himself. He	<pre><verb>will be delighted</verb>, I am sure, with Mrs Granger's taste and skill." "Damme, S:</pre>
nce of to-morrow. I	<pre><verb>will be exhibited</verb> to no one, as the refractory slave you purchased, such a t:</pre>
in, as your brother	<pre><verb>will be gone</verb> out at that hour. I don't ask to speak to you. I merely ask to s</pre>
g you give Miss Tox	<pre><verb>will be hoarded</verb> and prized, I am sure, like a relic. But there is a way, my</pre>
en I die, its ghost	<pre><verb>will be laid</verb>. As it is clearly no inheritance for you then, I have thought it</pre>
back, that he never	<pre><verb>will be muzzled</verb> when Paul Dombey is in question. Now, damme, Sir," concludes</pre>
his impertinence he	<pre><verb>will be punished</verb> terribly. You can pretend to say that, if you please, Towlin</pre>
a promise that you	<pre><verb>will be reconciled</verb> to Walter-to my dearest husband-to the father of the litt:</pre>
ence home. Gay, you	<pre><verb>will be rewarded</verb> to-morrow." "Oh! thank you, Sir," said Walter. "You are very</pre>
of opinion that it	<pre><verb>will be shorn</verb> of some of its emoluments, and made uncomfortable by newly-dev:</pre>
nd I believe nobody	<pre><verb>will be surprised</verb>, either at its being required from you, or at your making :</pre>
has happened here,	<pre><verb>will be</verb> a warning to you, Towlinson." "Thank you, Miss, I'm sure," says Towl:</pre>
sary. One of these,	<pre><verb>will be</verb> the engagement at Brighton (if it is to be effected), of a very respect</pre>
, "may be dreadful,	<pre><verb>will be</verb> dreadful; but I feel that I should wish to hear them spoken. I feel</pre>
se, in other words,	<pre><verb>will be</verb> able to hold its own, and maintain its own, and hand down its own of</pre>
•	

1.6.3 A new subgraph

Before continuing to improve our graph, we will simplify it by creating a new subgraph. Since the next steps will enhance the beginning of the graph without changing anything in its second part, we will create a new subgraph from it. In order to make the main graph easy to understand we will leave the insertion of the end-tag </verb> in the main graph. We will give the name *infinitiveVerb.grf* to the new subgraph.



If you apply this modified graph to the text, you will again obtain the same concordances, because we have only reorganized the graph.

1.7 Seventh step: negative adverbs

1.7.1 The main graph

In the concordance, you can see two special adverbs, *never* and *not*, which mark negative forms.

Concordance: U:\Documents\My	Unitex English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🗖 🛛
1343 matches	
espected parent. It	<pre><verb>will <adverb>naturally</adverb> be</verb> very pair</pre>
used to do; for he	<pre><verb>will <adverb>naturally</adverb> do</verb> that no not not not not not not not not not</pre>
dred thousand pound	<pre><verb>will <adverb>nearly</adverb> cover</verb> it. The v</pre>
hen you are old you	<pre><verb>will <adverb>never</adverb> depart</verb> from it,</pre>
hen you are old you	<pre><verb>will <adverb>never</adverb> depart</verb> from it.</pre>
for me, dear, and I	<pre><verb>will <adverb>never</adverb> forget</verb> how kind</pre>
ing that I hope you	<pre><verb>will <adverb>never</adverb> forget</verb> that you</pre>
eauty, and no grass	<pre><verb>will <adverb>never</adverb> grow</verb> upon the g:</pre>
stead, a homage you	<pre><verb>will <adverb>never</adverb> have</verb>." Although</pre>
te manner. That you	<pre><verb>will <adverb>never</adverb> speak</verb> of it to</pre>
, entreats that you	<pre><verb>will <adverb>never</adverb> talk</verb> to her any</pre>
ough paths that you	<pre><verb>will <adverb>never</adverb> tread</verb>, and my way</pre>
rnfully. "To you. I	<pre><verb>will <adverb>not</adverb> add</verb> to myself, as</pre>
st my friend Dombey	<pre><verb>will <adverb>not</adverb> allow</verb> his mind to</pre>
" said Mrs Chick "I	<pre><verb>will <adverb>not</adverb> attempt</verb> to disgui;</pre>
omfort her; but she	<pre><verb>will <adverb>not</adverb> be comforted</verb> on at</pre>
between them. "You	<pre><verb>will <adverb>not</adverb> be surprised</verb> at m</pre>
eautiful Mama! "You	<pre><verb>will <adverb>not</adverb> be</verb> long dressing,</pre>
t earnestness. "You	<pre><verb>will <adverb>not</adverb> begin</verb> by hating me</pre>
d show of affection	<pre><verb>will <adverb>not</adverb> benefit</verb> its objec</pre>
said Florence, "you	<pre><verb>will <adverb>not</adverb> blame</verb> me, will you</pre>
ant composure, "you	<pre><verb>will <adverb>not</adverb> conciliate</verb> me, or</pre>
	Construction of the second sec

We will now treat them separately, as a negation. They usually follow another adverb, although sometimes they can precede it. We will create a subgraph for negation, but before doing that we will put a call to a future subgraph named *negation.grf* at the right place in the main graph. This graph does not exist yet, so this box has a red background.



1.7.2 The negation subgraph

We can easily create this new subgraph from the main graph: we need to click on the *:negation* box, click the right mouse button and select *Open subgraph* from the menu displayed, then we need to click the *Yes* button to create the new graph.

After creating and saving the negation graph, the background of the *:negation* box in the main graph becomes gray.



1.7.3 Negative right context

If we look closely at the concordance obtained, we may be surprised to see 93 more lines, a total of 1 436 lines.⁵⁸ These are lines in which *never* and *not* are annotated as *negation*; they appear in addition to the 93 lines annotated as *adverb*. The *Longest matches* option does not allow for choosing between the two annotations; they are therefore both displayed.





If we look at the list of differences between the two concordances (see <u>Section 1.2</u>, page 47), we will see that the new concordance includes all forms of adverbs *never* and *not* labeled also as negation.

Concordance Diff	
Violet. Identical sequences with different outputs Red : similar but different sequences Green: sequences that occur, in only one of the two concordances Grey background=previous matches White background=new matches	
last! If my brother <u><verb>will <negation>not</negation> make</verb></u>	
adjoining room, he <verb>would <negation>not</negation> have gone</verb>	
ing that I hope you <verb>will <negation>never</negation> forget</verb>	
nt at Tusculum, she <verb>would <negation>not</negation> have had</verb>	
st my friend Dombey (verb)will (negation>not allows/verb)	
of my friend Dombey (yerb)will (negation)not/(negation) require/(yerb)	
of my filend bombey <u>verbywill vnegations requires verby</u>	
n Fearing "that you (work) would (megation) not (megation) proceed (merk)	
in reality, that you <u>verb/would inegation/hegation/press/verb/</u>	
he have not it (non-based) a constitution (non-tion) he (non-based)	
, ne knows, now, it <u>verb/would (negation/never(/negation/ be(/verb/</u>	
ery dear to him. He <u><verb>will <negation>not</negation> grieve</verb></u>	
	H
st entreaty that he <u><verb>will <negation>not</negation> seek</verb></u>	
hen you are old you <u><verb>will <negation>never</negation> depart</verb></u>	-

This is normal, because *never* and *not* are both adverbs and negations, but that is not what we want. To force the negation tag, we will use a *negative right context* in the *adverb.grf* subgraph, which is not to be confused with the *right context* seen in <u>Chapter 2, Section 2.4.2</u>, page 32.



⁵⁸ There is the same number of matches as before, 1 343, but more concordance lines, 1 436.

Now, *never* and *not* are always annotated as negation. There are again 1 343 lines in the produced concordance.

Concordance: U:\Documents\My	'Unitex/English/Corpus\DombeyAndSon\DombeyAndSon_snttconcord.html 🗖 🗹
1347 matches	
dred thousand pound	<pre><verb>will <adverb>nearly</adverb> cover</verb> it. The the task</pre>
aid, "Gentlemen, we	<pre><verb>will <adverb>now</adverb> resume</verb> our studie</pre>
oor, "Gentlemen, we	<pre><verb>will <adverb>now</adverb> resume</verb> our studie;</pre>
hose mistakes which	<pre><verb>will <adverb>occasionally</adverb> arise</verb> in</pre>
one before him, and	<pre><verb>will <adverb>often</adverb> do</verb> again. Capta:</pre>
d, long undermined,	<pre><verb>will <adverb>often</adverb> fall</verb> down in a r</pre>
asiness which a cur	<pre><verb>will <adverb>often</adverb> manifest</verb> in a s;</pre>
er Antony Bagstock,	<pre><verb>will <adverb>often</adverb> speak</verb> of this,</pre>
d the Captain, "and	<pre><verb>will <adverb>only</adverb> confirm</verb> that the:</pre>
ever saw in you. I	<pre><verb>will <adverb>only</adverb> say</verb> to you, once</pre>
ss his opinions, he	<pre><verb>will <adverb>openly</adverb> say</verb>, that he define the definition of the say of the</pre>
r an Emergency that	<pre>will <adverb>sometimes</adverb> arise in the</pre>
y received all that	<pre><verb>will <adverb>soon</adverb> be left</verb> of him or</pre>
Florence, this lady	<pre><verb>will <adverb>soon</adverb> be</verb> your Mama." F</pre>
on, and Miss Dombey	<pre><verb>will <adverb>soon</adverb> be</verb> rising. You a:</pre>
I have a child who	<pre><verb>will <adverb>soon</adverb> call</verb> Walter by tl</pre>
s, on your child's.	<pre><verb>Will <adverb>that</adverb> <adverb>ever</adverb> be</verb></pre>
iendly remark, "you	<pre><verb>will <adverb>then</adverb> know</verb> that all gr:</pre>
hen you are old you	<pre><verb>will <negation>never</negation> depart</verb> from</pre>
hen you are old you	<pre><verb>will <negation>never</negation> depart</verb> from</pre>
for me, dear, and I	<pre><verb>will <negation>never</negation> forget</verb> how 1</pre>
ing that I hope you	<pre><verb>will <negation>never</negation> forget</verb> that</pre>
1	

1.8 Eighth step: adding the verb shall

The verb *shall* follows the same rules as the verb *will*, making it easy to add it in the main graph. Since this verb is not ambiguous with a noun or some other category, as *will* is, it is not necessary to add its category in a lexical mask, so we will use *<shall>*.⁵⁹



The new concordance contains 1 948 lines.

Concordance: U:\Documents\My	/Unitex/English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🗗 🛛
1948 matches	
to change so!" "We	<pre><verb>shall <adverb>all</adverb> change</verb>, mother,</pre>
hand fervently, "I	<pre><verb>shall <adverb>certainly</adverb> be</verb> present</pre>
rhaul the book-" "I	<pre><verb>shall <adverb>certainly</adverb> do</verb> so, Cap</pre>
en avoided. Nothing	<pre><verb>shall <adverb>ever</adverb> persuade</verb> me," ol</pre>
ish you well, and I	<pre><verb>shall <adverb>ever</adverb> wish</verb> you well. 1</pre>
de and insult me. I	<pre><verb>shall <adverb>merely</adverb> go</verb>. I shall no</pre>
etera, et cetera. I	<pre><verb>shall <adverb>now</adverb> beg</verb> to propose M</pre>
at all unkind, and	<pre><verb>shall <adverb>only</adverb> be <adverb>too</adverb></verb></pre>
d assure you that I	<pre><verb>shall <adverb>perfectly</adverb> understand</verb></pre>
"My dear Louisa, I	<pre><verb>shall <adverb>really</adverb> be</verb> proud, soon</pre>
right hand. My name	<pre><verb>shall <adverb>then</adverb> be</verb> at your serv:</pre>
, and that I fear I	<pre><verb>shall <negation>never</negation> be permitted</verb></pre>
gulated Families. I	<pre><verb>shall <negation>never</negation> cease</verb> to cease</pre>
me say it out, or I	<pre><verb>shall <negation>never</negation> say</verb> it. I</pre>
etter; and I hope I	<pre><verb>shall <negation>never</negation> see</verb> your f</pre>
shall merely go. I	<pre><verb>shall <negation>not</negation> be missed</verb>!" I</pre>
, or stand by me. I	<pre><verb>shall <negation>not</negation> be run</verb> in up</pre>
g over), at least I	<pre><verb>shall <negation>not</negation> be</verb> alone, within </pre>
ked together, but I	<pre><verb>shall <negation>not</negation> be</verb> so today.</pre>
r; and tell me this	<pre><verb>shall <negation>not</negation> be</verb> so any mo:</pre>
t, or to say, 'This	<pre><verb>shall <negation>not</negation> be</verb>!' a pride</pre>
y understand you. I	<pre><verb>shall <negation>not</negation> consider</verb> it </pre>
1	

⁵⁹ Keep in mind that the lexical mask *< shall.V* > would not be incorrect; these two masks produce the same results.

1.9 Continuation of this graph (only for confident users)

Before starting this section

In order to retrieve all results you should use the *Norm.txt* file introduced in <u>Chapter 2</u>, <u>Section 3.1.2</u>, page 35. If you have not used it when you preprocessed the text, download it, then close and reopen Unitex.

Open the original *DombeyAndSon.txt* file, process it by applying the default dictionaries (do not use *Sentence* or *Replace* graphs).

1.9.1 Addition of contracted forms

In the *willV* graph, we have omitted the recognition of contracted forms *won't*, *wouldn't*, *shan't* and *shouldn't*. We need a separate path for them because, in the case of contractions with *not*, additional negations are not allowed.



If you now apply the new *willV* graph to the text, you will get 2 069 lines, including 121 contracted forms, and among them all four possibilities.

🗖 Canada and Li Da su manha il hi la ila d'Easticht Canada Da mhaide d'Easticht d'Easti	
Concordance. C. Documents wyonitexenglismic or pusiblo mbey And Son Dombey And Son Sinicon conduitini	🗋 Concordance: U:IDocuments/MyUnitex/English/Corpus/DombeyAndSon/DombeyAndSon_snftconcord.html 🛛 🗖 🛛
2069 matches	2069 matches
expected." "Then I <verb>shall wish</verb> you good morning, Miss," said F	modesty. Now what I <verb>should wish</verb> to put to you," said the Capta*
hen the silent tomb <verb>shall yawn</verb> , Captain Gills, I shall be read	" said Mr Toots, "I <verb>should wish</verb> to be strictly honourable; but
heme, "to-morrow, I <verb>shan't be found</verb> in this here roadstead til	do in the world, I < <u>verb>shouldn't allow</u> of your coming here, to be
nto this state,' he < <u>verb>should <adverb>actually</adverb> fly</u> at me	"Suppose the wages <verb>shouldn't be paid</verb> !" The company sit for a
he room again, they <verb>should <adverb>all</adverb> make</verb> way for h	t thought of all, I < <u>verb>shouldn't be</u> here now. Wretchedness and ru
Concordance: L1:DocumentsWyl Initex/English/Corpus/DombevAndSon/DombevAndSon_spticoncord html	
	Concordance: U:\Documents\MyUnitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 🛛 🗖 🔯
2069 matches	Concordance: U/Documents/MyUnite//English/Corpus/DombeyAndSon/DombeyAndSon_snftconcord.html
2009 matches ies." "If Mr Dombey <u><verb>will walk</verb></u> upstairs," said Mrs Blimber, "I	□ Concordance UUDocumentsWMUUnitexEnglishiCorpusDombeyAndSonDombeyAndSon_sntconcord.html n Z Z 2069 matches □ G Instruction □ Gighteenth. That he < <u>verb>would write</u> to Mr Dombey, when he should h
2080 matches ies." "If Mr Dombey <u><verb>will walk</verb></u> upstairs," said Mrs Blimber, "I the sea." "And you <u><verb>will wear</verb></u> it still, Walter, for my old sak	□ Concordance: UUDocumentsWhjUnitexEnglishiCorpusDombeyAndSonDombeyAndSon_snitconcord.html n Z 2069 matches eighteenth. That he <u><verb>would write</verb></u> to Mr Dombey, when he should h≜ such an investment <u><verb>would yield</verb></u> a handsome premium. Solomon Gi
<pre>2069matches [ies." "If Mr Dombey <u>(verb)vill walk/verb)</u> upstairs," said Mrs Blimber, "I * the sea." "And you <u>(verb)vill wasc/verb)</u> is still, Walter, for my old sak back some clothes. <u>(verb)Wan't cadverb)that(/adverb) hec/verb)</u> best?" "I t</pre>	□ Concordance UUDcourentsWijUnitexEnglishiCopusDombeyAndSonDombeyAndSon_smiconcord.html n
<pre>2000 matches ies." "If Mr Dombey <u><verb>will walk</verb></u> upstairs," said Mrs Blimber, "I = the sea." "And you <u>verb>will wear</u> it still, Walter, for my old sak back some clothes. <u><verb>Won't allow</verb> hes best?" "I t ly charming, that I <u>verb>won't allow</u> you to disturb it, do you heat</u></pre>	□ Concordance UUDocumentsWMUnitexEnglishCorpusDombeyAndSonDombeyAndSon_sniconcord.html n' Z Z 2069 matches eighteenth. That he < <u>verb>would vrite</u> to Mr Dombey, when he should h≤ such an investment <u>vverb>would vield</u> to An Andsome premium. Solomon Gi think she was, but <u>vverb>wouldn't <adverb>quite</adverb>_uite_momt_on ogo terms, and <u>vverb>wuldn't <adverb>quite</adverb>_met_met_mitting </u></u>
<pre>2009 matches ies." "If Mr Dombey <u><verb>will walk</verb></u> upstairs," said Mrs Blimber, "I the sea." "And you <u><verb>will wal</verb></u> it still, Walter, for my old sak back some clothes. <u><verb>Won!t <advrb>that</advrb>that best?" "I t ly charming, that I <u><verb>won!t allow</verb></u> you to disturb it, do you hear me anything else. I <u><verb>won!t antwr</verb></u> my not disturb it. How later the same set the same set to </verb></u></pre>	□ Concordance UUDocumentsWMUnitexEnglishiCorpusDombeyAndSonDombeyAndSon_smitconcord.html n 2 2069 2069 matches eighteenth. That he weighteenth. verbb>would vite to Mr Dombey, when he should he such an investment weighteenth. such an investment guites/adverbb">weighteenth. Solomon Gi think she was, but guites/adverbb">weibbwould.tit.ca/dverbb>guites/adverbb New //weibbwould.tit.ca/dverbb>guites/adverbb on good terms, and guites/adverbb">weibbwould.tit.ca/dverbb>guites/adverbb New //weibbwould.tit.ca/dverbb zon, Miss Floy, and soc/adverbb">veibbwould.tit.ca/dverbb>soc/adverbb Ther
2009 matches ies." "If Mr Dombey <verb>will walk</verb> upstairs," said Mrs Blimber, "I	□ Concordance UtDocumentSWyUnitexEnglishCorpusDombeyAndSomDombeyAndSom_sontconcord html

1.9.2 Affirmative and negative constructions

We can now distinguish between affirmative and negative forms. We will add the attribute *construction* with possible values *affirmative/negative* to the XML start tag *<verb>*.



When you apply this graph to the text, you will obtain the same number of concordance lines, 2 069. The output is different since recognized verb groups are now annotated as affirmative or negative.

🗍 Concordance: U1Documents\exklyUnitexlEnglishlCorpus\DombeyAndSon\DombeyAndSon_snttconcord.html
069 matches
power to bestow, I <pre><pre>construction="negative">shall <negation>not</negation> scruple</pre></pre> <pre>to avail myself of his</pre>
neme, "to-morrow, I <u><verb construction="negative">shan't be found</verb></u> in this here roadstead till night-not till ar
Grinder. "'If you <verb construction="negative">should <negation>never</negation> hear</verb> of me, or see me more,
-I-I really think I <u><verb construction="negative">should <negation>never</negation> leave</verb></u> off wagging my tail.
se upon him that he <u><verb construction="negative">should <negation>never</negation> rest</verb></u> again, his drowsy sens

1.9.3 Interrogative constructions

In order to recognize interrogative constructions as well it is enough to create a subgraph for personal pronouns⁶⁰ and to duplicate several paths. The attribute type will now have two more possible values: *interro-negative* and *interrogative*.



⁶⁰ We add to the list of common pronouns the old form of the second person singular, *thou*.

If you apply this graph to the text you will obtain 2 224 concordance lines, some of which are represented below.

🛾 Concordance: U\Documents\exhtyUnitexEnglishCorpusiDombeyAndSon_enticoncord.html 🕫 🖬	\boxtimes
224 matches	
such an investment <verb construction="affirmative">would yield</verb> a handsome premium. Solomon Gills was at first stunned by the communication	
re felt it; and why <verb construction="interro-negative">should <pronoun>I</pronoun> <negation>not</negation> avow</verb> it in behalf of this unf	1
? If I feared you, <pre>construction="interro-negative">should <pronoun>I</pronoun> <negation>not</negation> have avoided you? If I feared</pre>	
etely at an end." " <pre></pre>	
shing river. "Why, < <u>verb construction="interro-negative">will <pronoun>it</pronoun> <negation>never</negation> stop</u> , Floy?" he would someti	-

2 Second example: numbers written with words

Our second exercise will deal with the annotation of numerals written using only words (not digits). We will proceed step-by-step using subgraphs in each of them. We will save this graph in the *Numbers* folder:

2.1 First step: from 2 to 9

We will start with numbers *two* to *nine*. We will skip the number *one* because it is ambiguous with the pronoun *one*. We will save this graph using the name *NB2-9.grf*.



When we apply this graph to our text *Dombey and son* in *Merge* mode and without the search limitation (*Index all occurrences in text*), we obtain 565 concordance lines, some of which are displayed in the screenshot.

Concordance: U:\Documents\My	'Unitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🖬 🗹
565 matches	
s. Here, there were	<pre><number>eight</number> young gentlemen in various stages</pre>
ry rail (there were	<pre><number>eight</number>-and-forty, as Miss Tox had often </pre>
our. At the hour of	<pre><number>eight</number>, a deep lowing, as of a nautical 1</pre>
ratio. Thus, taking	<pre><number>eight</number> as our standard and highest numbe;</pre>
wenty minutes after	<pre><number>eight</number> o'clock. I see him myself, with h</pre>
eir ride was six or	<pre><number>eight</number> miles long. When they drove throug</pre>
an conversation. At	<pre><number>eight</number> o'clock or so, the gong sounded a</pre>
s of age. Son about	<pre><number>eight</number>-and-forty minutes. Dombey was rat!</pre>
w. Dombey was about	<pre><number>eight</number>-and-forty years of age. Son about</pre>
a, at home, that by	<pre><number>eight</number> o'clock or so, they settle down in</pre>
a gentleman thirty-	- <u><number>eight</number></u> or forty years old, of a florid c
face. Four struck!	<pre><number>Five</number>! No Edith yet. But now there was s</pre>
me at a book-stall,	<number>five</number> lines of which utterly confounded 1
ious sorts of fowl,	<pre><number>five</number> thousand dishes." "Or try a crust</pre>
d to and fro again,	<number>five</number> hundred times, among the splendid 1
t, I do assure you,	<number>five</number> distinct times." "Is your wife so :
ope you'll keep so.	<number>Five</number> children. Youngest six weeks. The
a terrible extent.	<pre><number>Five</number> o'clock arriving, and with it Mr C</pre>
the dead of night;	<number>five</number> -and-twenty minutes past twelve when
eve my Martha's had	<pre><number>five</number> short days' good health," he answe</pre>
he hours of ten and	<pre><number>five</number>, slippers, pocket-books, sponges,</pre>
settle in one place	<number>five</number> minutes. He goes about, like a-jus
lin himself for the	annhand fine (annhand hundus dab time schethen the Canteria

2.2 Second step: from 2 to 99

2.2.1 A subgraph without output

We would like to use the previous graph as a subgraph. It is possible to do it for numbers from 2 to 9; for numbers from 21 to 29 it is also possible to use the previous graph as a subgraph

*MyUnitex**English**Graphs**UnitexGettingStarted**Numbers*

but the result would contain some unwanted nesting: for instance, *twenty-two* would be annotated as *<number>twenty-<number>two</number></number>.* In order to avoid this, we will create, based on *NB2-9* graph, an *NB1-9-subgraph* graph that produces no output and contains the number one. Note that *one* that follows other numbers written in words is no longer ambiguous. This graph annotates *twenty-two* as *<number>twenty-two</number>.*



2.2.2 The NB2-99.grf graph

Now we can produce a graph that recognizes numerals from 2 to 99. It will have one path for the recognition of numbers *two* to *nine*: this path will consist of a box that calls the *NB2-9.grf* subgraph, which means that the annotation will also be performed by the subgraph. Another path will contain the box that recognizes numbers *ten* to *nineteen*. The box that recognizes these numbers will insert the start tag *<number>*. The third path will recognize numbers *twenty* to *ninety*, which can be followed by numbers *one* to *nine*. For this recognition, the subgraph *NB1-9-subgraph.grf* will be used. A hyphen can be used between numbers written in words. The box that will recognize numbers *twenty* to *ninety* will insert the start tag *<number>*, while the end tag *</number>* will be produced by an empty box common to the second and third paths (but not the first one). We will save this graph under the name *NB2-99.grf*.



A few of the 709 concordance lines produced when applying this graph to the text are displayed in the screenshot.

Chapter 3: Corpus annotation

🔲 Concordance: U:\Documents\MyUnitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 🖉 🗹		
709 matches		
red. "May I ask for	<pre><number>five</number> minutes of your leisure?" After a *</pre>	
ace in this parlour	<pre><number>five</number> and forty years ago, I never could</pre>	
ometimes as many as	<pre><number>five</number>-and-twenty people attended service</pre>	
shall leave this at	<pre><number>five</number> o'clock, and ride there on horseba</pre>	
ns. It is half-past	<pre><number>five</number> o'clock, and an autumn afternoon, v</pre>	
lish brig, of about	<pre><number>five</number> hundred tons burden, together with</pre>	
l grew to be nearly	<pre><number>five</number> years old. He was a pretty little :</pre>	
a cloud for nearly	<pre><number>five</number> minutes-an extraordinary period fo:</pre>	
February, seventeen	<pre><number>forty-nine</number>!" cried Walter, with great a:</pre>	
ome fellow, Sir, of	<pre><number>forty-one</number>. He died, Sir, in the second</pre>	
ber one hundred and	<pre><number>forty-seven</number>, as, somehow, a part of its</pre>	
ber one hundred and	<pre><number>forty-seven</number>, under the awful shadow of I</pre>	
is one hundred and	<pre><number>forty-seven</number>. Louisa, you can tell her."</pre>	
." "One hundred and	<pre><number>forty-seven</number>, " said Mrs Chick "The dress</pre>	
s a man about town,	<pre><number>forty</number> years ago; but he is still so juve</pre>	
taining any injury.	<pre><number>Forty</number> years at least had elapsed since</pre>	
is parlour five and	<pre><number>forty</number> years ago, I never could have been</pre>	
forgotten 'em these	<pre><number>forty</number> year. "But I never wanted two or</pre>	
man thirty-eight or	<pre><number>forty</number> years old, of a florid complexion</pre>	
lleged to thirty or	<pre><number>forty</number> friends that she was bored to dea</pre>	
e repeated at least	<pre><number>forty</number> times during tea; polishing his rail</pre>	
emost into five-and-	- <u><number>forty</number></u> feet of water, Mrs Richards, but	
l'anna anna ai mha an a'		

2.3 Third step: from 2 to 999

2.3.1 The NB1-99-subgraph.grf graph

For the same reason as before, we will produce a graph that recognizes numbers *two* to *ninety-nine* but does not insert any tag. Namely, we want to use it as a subgraph of the graph that will recognize numbers from *two* to *nine hundred ninety-nine*. We will base this graph on the already developed *NB2-99* graph and name it *NB2-99-subgraph.grf*.



2.3.2 The NB2-999.grf graph

In this graph we will use as a subgraph *NB2-99.grf* to recognize and annotate all numbers in the interval from 2 to 99. For numbers in the interval from 100 to 199, the word *hundred* must be present and can be preceded by an indefinite article *a* or by a number *one*. For numbers in the interval 200 to 999, the word *hundred* has to be preceded by a number from 1 to 9 (possibly followed by a hyphen); for that, we will use the appropriate subgraph that does not insert tags. The previously produced *NB1-99-subgraph.grf* subgraph recognizes numbers from 1 to 99, without annotating them. Note that numbers from 1 to 99 that follow hundreds can be preceded with the conjunction *and*.



The application of the main graph to the text produces 719 concordance lines.

Concordance: U:\Documents\M	'Unitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🗖 🛛
719 matches	
airy tales, who, at	<pre><number>a hundred and fifty</number> or two hundred year</pre>
nothing. There were	<pre><number>a hundred</number> thousand shapes and substance;</pre>
e would freely give	<pre><number>a hundred</number> thousand pounds (if he had it</pre>
if he were to live	<pre><number>a hundred</number> years, he never could forget :</pre>
at the distance of	<pre><number>a hundred</number> yards or so, and watched it, </pre>
ning indignation of	<pre><number>a hundred</number> women, strong in passion and :</pre>
e,-ah! maybe out of	<pre><number>a hundred</number>, pretty,-has been saved by the</pre>
uld have been worth	<pre><number>a hundred</number> state sights. "And now," said</pre>
d one-not less than	<pre><number>a hundred</number> thousand pound. Mr Perch don'</pre>
e's a Tartar within	<pre><number>a hundred</number> miles of where we're now in ce</pre>
ply of learning for	<pre><number>a hundred</number>, on the lowest estimate; and :</pre>
think himself that	<pre><number>a hundred</number> thousand pound will nearly cor</pre>
y friend Gills many	<pre><number>a hundred</number> pound already. More particula;</pre>
s. Here, there were	<pre><number>eight</number> young gentlemen in various stages</pre>
ry rail (there were	<pre><number>eight</number>-and-forty, as Miss Tox had often </pre>
our. At the hour of	<pre><number>eight</number>, a deep lowing, as of a nautical 1</pre>
ratio. Thus, taking	<pre><number>eight</number> as our standard and highest numbe;</pre>
wenty minutes after	<pre><number>eight</number> o'clock. I see him myself, with h:</pre>
eir ride was six or	<pre><number>eight</number> miles long. When they drove throug</pre>
an conversation. At	<pre><number>eight</number> o'clock or so, the gong sounded a</pre>
s of age. Son about	<pre><number>eight</number>-and-forty minutes. Dombey was rat</pre>
w. Dombey was about	<pre><number>eight</number>-and-forty years of age. Son about</pre>
A and have about her	<pre>countration in the claume in the claume is the claum</pre>

2.4 Fourth step: from 2 to 999 999

2.4.1 The NB100-999-subgraph.grf graph

We will produce a graph named *NB100-999-subgraph.grf* that recognizes numbers from 100 to 999 without annotating them. It presents the modification of the previously developed *NB2-999.grf* graph, but the path recognizing numbers from 2 to 99 has been deleted. The reason why this path was deleted will become clear when observing the next graph, which recognizes numbers from 2 to 999 999.



2.4.2 The NB2-999999.grf graph

This graph will look similar to the *NB2-999.grf* graph, except that here the word *thousand* has to be used, optionally preceded by *a* or a number between 1 to 99 (*NB1-99-subgraph.grf*, see <u>Section 2.3.1</u>, page 61) or between 100 to 999 (*NB100-999-subgraph.grf*). The word *thousand* can be followed by a number between 1 to 99 (subgraph *NB1-99-subgraph.grf*), which can be preceded by a conjunction *and*. The word thousand can also be followed by a number between 100 to 999 (*NB100-999-subgraph.grf*), but in that case, the conjunction *and* is not

used and that is the reason that we separated the cases of numbers between 1 to 99 and between 100 to 999.



If we apply this graph to the text, we will obtain 736 concordance lines for numbers from 2 to 999 999.

36 matches	736 matches
arry tales, who, at <u><number>a hundred and fifty</number></u> or two hundred year	y friend, by Number <u><number>nine</number></u> Brig Place?" inquired the Captain.
nothing. There were <u><number>a hundred thousand</number></u> shapes and substance	Monday morning, at <u><number>nine</number></u> o'clock-habit again-I must be busine
would freely give < <u>number>a hundred thousand</u> pounds (if he had it	At about half-past < <u>number>nine</u> o'clock the odour of a warm sweet-b:
d one-not less than <u><number>a hundred thousand</number></u> pound. Mr Perch don'	when they were only < <u>number>nine</u> in family. Mr Towlinson has a singi
think himself that < <u>number>a hundred thousand</u> pound will nearly co	nters Heaven, among < <u>number>ninety-nine</u> just men. It was not dimmed (
if he were to live < <u>number>a hundred</u> years, he never could forget	ewter badge, number < <u><number>one hundred and forty-seven</number></u> , as, somehow
at the distance of < <u>number>a hundred</u> yards or so, and watched it, {	leave to see number < <u>number>one hundred and forty-seven</u> , under the au
ning indignation of < <u>number>a hundred</u> women, strong in passion and	a hackney-coach, is <u><number>one hundred and forty-seven</number></u> . Louisa, you
a, -ah! maybe out of <u><number>a hundred</number></u> , pretty, -has been saved by the	you can tell her." "< <u>number>One hundred and forty-seven</u> ," said Mrs C
ald have been worth < <u>number>a hundred</u> state sights. "And now," said	s a few. Annuity of < <u>number>one hundred</u> pound premium also ready to
e's a Tartar within <u><number>a hundred</number></u> miles of where we're now in c	pful Company did in < <u>number>one thousand six hundred and ninety-four</u>
oly of learning for < <u>number>a hundred</u> , on the lowest estimate; and	s are not the same. < <u>number>Seven</u> -eighths of my stock is old-fashion
y friend Gills many < <u>number>a hundred</u> pound already. More particula	from the Latin, and <u><number>seven</number></u> from the Greek, which moved the you
as they pass along, < <u>number>a thousand</u> heads are turned to look at	ive. But one of the < <u>number>seven</u> mild men unexpectedly leaped into
and yet entertained < <u>number>a thousand</u> indistinct and visionary fand	York to six of the < <u>number>seven</u> mild men (the ambitious one was ut
b look at them, and <u><number>a thousand</number></u> sober moralists revenge them	divers times, with < <u>number>seven</u> sixpences, in token of satisfaction
if she lives to be < <u>number>a thousand</u> years old." Miss Tox elevated	n again, slept till <u><number>seven</number></u> o'clock. Nor was he free from his
of his cloak, made < <u>number>a thousand</u> indistinct shapes, answering	ner-hour is a sharp <number>seven</number> , Mr Carker." With that, the Major,
day long, and have < <u>number>a thousand</u> and one appointments with al	etters of the other <u><number>seven</number></u> lines, deserting him at that point
g, in the midst of <number>a thousand</number> extravagances, to look back	sume our studies at < <u>number>seven</u> to-morrow;" and then, for the firs
n was crossed with < <u>number>a thousand</u> little creases, which the sa	sume our studies at <number>seven</number> tomorrow," the pupils bowed again,
as unconscious, for < <u>number>a thousand</u> pounds. "My Heart's Delight!	hour was half-past < <u>number>seven</u> o'clock, and when the object was q
nine Ma Dember for Aparts the second of four board a side and the shares i	the second secon

3 Third example: annotating Roman numerals

We will now continue with Roman numerals which we want to tag as *romanNumeral*. We want to describe them, as number written in words in the previous section, with graphs. However, we encounter a problem if we use graphs in a similar way to before. Namely, with three successive boxes containing the letter *I*, we can recognize three successive words *I I I* and not one word *III*. In order to solve this problem, we use the morphological mode that allows for matching a word, that is, a sequence of letters.

3.1 First step: from 2 to 9

We will start with the graph that recognizes Roman numerals with values 2 to 9. As before, we will skip the numeral *I* because it is ambiguous with the pronoun *I*. We will name this graph *RN2-9.grf*. We will build this graph bearing in mind the rules for constructing Roman numerals. Once we have done that we will select all boxes, except the start box and the end box. Then we will click the *Surround box selection with morphological mode tags* button (you will find this button at the right end of the toolbar represented as violet angular brackets). The graph is now surrounded with two boxes that contain codes \$ and \$. Inside this *morphological part*, the content of boxes is read as concatenated.

In order to obtain a tidy graph, we will align the left and right parts of the graph. This can be done by opening the *FSGraph/Format/Alignment/Horizontal/Center* menu option. This menu can be accessed by the menu bar or by a click with the right mouse button.



When applying this graph to the text we obtain a concordance with the numbers of the first eight chapters of our novel *Dombey and Son* (except the first one, which we omitted on purpose).

Concordance: U:	Documents\M	/Unitex/English/Corpus/DombeyAndSon/DombeyAndSon_snt\concord.html 📃 🗖 🛛
8 matches		
the world.	CHAPTER	<pre><romannumeral>II</romannumeral>. In which Timely Provisi</pre>
accustomed.	CHAPTER	<pre><romannumeral>III</romannumeral>. In which Mr Dombey, as</pre>
hstanding!"	CHAPTER	<pre><romannumeral>IV</romannumeral>. In which some more First</pre>
, far away.	CHAPTER	<pre><romannumeral>IX</romannumeral>. In which the Wooden Mide</pre>
Daughter!"	CHAPTER	<pre><romannumeral>V</romannumeral>. Paul's Progress and Chrice</pre>
ering down.	CHAPTER	<pre><romannumeral>VI</romannumeral>. Paul's Second Deprivation</pre>
s about it.	CHAPTER	<pre><romannumeral>VII</romannumeral>. A Bird's-eye Glimpse of</pre>
y or a Son.	CHAPTER	<pre><romannumeral>VIII</romannumeral>. Paul's Further Progret</pre>
•		

3.2 Second step: from 2 to 99

3.2.1 The subgraph from 1 to 9

As before, we will use the previously created *RN2-9.grf* graph to create a subgraph that does not insert tags. We will also cancel the morphological mode (by deleting the boxes that contain \$ and \$), because it will have been introduced by the invoking graph, and nesting of morphological modes is not allowed. We will add a path that recognizes the roman numeral *I*, because in the context it will be used. We will name this graph *RN1-9-subgraph.grf*.



3.2.2 The graph from 2 to 99

Roman numerals are constructed very regularly. Therefore, in order to recognize numbers greater than 10 and up to 99, we can use the preceding *RN2-9.grf* graph, and replace in it *I* with *X*, *V* with *L* and *X* with *C*. Then we have to add a path for *X*, a call to the *RN2-9.grf* subgraph outside the morphological mode (this to recognize numerals from 2 to 9), and a call to the *RN1-9-subgraph.grf* (inside the morphological mode) for numerals from 1 to 9 when they are added to numerals greater or equal to ten. We will name this graph *RN2-99.grf*.

Chapter 3: Corpus annotation



When we apply this graph to our text, we obtain 63 concordance lines. Note that this text (novel) has 64 chapters designated with Roman numerals. With the *RN2-99.grf* graph we recognize 63 of them (we do not recognize the first chapter "I").

Concordance: U:\	Documents\My	رUnitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🗖 🛛
63 matches		
the world.	CHAPTER	<pre><romannumeral>II</romannumeral>. In which Timely Provisi</pre>
accustomed.	CHAPTER	<pre><romannumeral>III</romannumeral>. In which Mr Dombey, as</pre>
hstanding!"	CHAPTER	<pre><romannumeral>IV</romannumeral>. In which some more Firs</pre>
, far away.	CHAPTER	<pre><romannumeral>IX</romannumeral>. In which the Wooden Mid:</pre>
. "Anything	about a	<pre><romannumeral>L</romannumeral>," said the Captain, "or a</pre>
f victuals.	CHAPTER	<pre><romannumeral>L</romannumeral>. Mr Toots's Complaint The:</pre>
he chips ou	at P-A-U-	- <romannumeral>L</romannumeral> in the marble slab before
r, jintly!"	CHAPTER	<pre><romannumeral>LI</romannumeral>. Mr Dombey and the World</pre>
e together.	CHAPTER	<pre><romannumeral>LII</romannumeral>. Secret Intelligence God</pre>
him of it.	CHAPTER	<pre><romannumeral>LIII</romannumeral>. More Intelligence The:</pre>
was empty.	CHAPTER	<pre><romannumeral>LIV</romannumeral>. The Fugitives Tea-time</pre>
intry wind.	CHAPTER	<pre><romannumeral>LIX</romannumeral>. Retribution Changes have</pre>
re shining.	CHAPTER	<pre><romannumeral>LV</romannumeral>. Rob the Grinder loses h:</pre>
n of ashes.	CHAPTER	<pre><romannumeral>LVI</romannumeral>. Several People delight(</pre>
dear love.	CHAPTER	<pre><romannumeral>LVII</romannumeral>. Another Wedding Mr Sov</pre>
y far away!	CHAPTER	<pre><romannumeral>LVIII</romannumeral>. After a Lapse The set</pre>
isposed of.	CHAPTER	<pre><romannumeral>LX</romannumeral>. Chiefly Matrimonial The</pre>
too much!"	CHAPTER	<pre><romannumeral>LXI</romannumeral>. Relenting Florence had</pre>
f a dream."	CHAPTER	<pre><romannumeral>LXII</romannumeral>. Final A bottle that have</pre>
Daughter!"	CHAPTER	<pre><romannumeral>V</romannumeral>. Paul's Progress and Chri:</pre>
ering down.	CHAPTER	<pre><romannumeral>VI</romannumeral>. Paul's Second Deprivation</pre>
s about it.	CHAPTER	<pre><romannumeral>VII</romannumeral>. A Bird's-eye Glimpse o:</pre>
	OUL DEED	<pre></pre>

3.3 Third step: from 2 to 999

3.3.1 The subgraph from 1 to 99

We will repeat the same procedure to produce the *RN2-9-subgraph.grf*, namely, we will cancel the morphological mode and the insertion of tags. We should not forget to replace the call to the *RN2-9.grf* subgraph with *RN1-9-subgraph.grf* because the later subgraph does not insert tags, does not impose morphological mode and recognizes *I* (this subgraph is never used independently).



3.3.2 The graph from 2 to 999

In order to recognize Roman numerals between 2 and 999, we will again use the already developed *RN2-99.grf* graph in which we will replace *X* with *C*, *L* with *D* and *C* with *M*. Again, we will replace the call to the *RN2-9.grf* subgraph with *RN2-99.grf* outside the morphological mode (for numerals between 2 and 99), and a call to the subgraph with *RN1-99-subgraph.grf* (inside the morphological mode) for numerals between 1 to 99, when they are added to hundreds. We will name this graph *RN2-999.grf*.



When we apply this graph to our text, we will obtain 76 concordance lines. But there are no Roman numerals greater than 62 (the last chapter) in this novel, so these thirteen new lines are due to false recognitions: The Roman numerals for 100 and 500 are confused:

- With the letters C and D.⁶¹
- With abbreviations, i.e. K.C.B.: Knight Commander of the Bath (see the Wikipedia).⁶²

⁶¹ He acquitted himself very well, nevertheless; and Miss Blimber, commending him as giving promise of getting on fast, immediately provided him with subject B; from which he passed to C, and even D before dinner.

⁶² If that had been Joe's character, Joe might have been, by this time, Lieutenant-General Sir Joseph Bagstock, K.C.B., and might have received you in very different quarters.

Chapter 3: Corpus annotation

🔲 Concordance: U:\Documents\MyUnitexEnglish\Corpus\DombeyAndSon\DombeyAndSon_sntlconcord.html 🗖 🗹
76 matches
which he passed to < <u>romanNumeral>C</u> , and even D before dinner
I should think, Mr < <u>romanNumeral>C</u> .," retorted his helpmate,
Joseph Bagstock, K. <u><romannumeral>C</romannumeral></u> .B., and might have receive
ference to them. A. science had no concern with Anno
ell? There you are. <pre><rreater north<="" pre="" the=""></rreater></pre>
I brings up short. < <u><romannumeral>D</romannumeral>'</u> ye make me out pretty cle
ssed to C, and even < <u>rromanNumeral>D</u> before dinner. It was har
n went so far as to < <u>romanNumeral>D</u> - Mr Baps to Lady Skettles
mind. "I say! Miss < <u><romannumeral>D</romannumeral></u> .O.M. you know!" The Capta:
hus they both spelt <pre><rreathing< pre=""></rreathing<></pre>
," said Mrs Brown. " <u><romannumeral>D</romannumeral>'</u> ye understand what I say?
ould have excused: " <u><romannumeral>D</romannumeral></u> d bad," "Jack Bunsby," w
ly, on the table. "` <u><romannumeral>D</romannumeral></u> ,'" the old woman read alo
the world. CHAPTER commonstants.com the world. The world. The world. The world.
accustomed. CHAPTER < <u>romanNumeral>III</u> . In which Mr Dombey, as
hstanding!" CHAPTER < <u>rromanNumeral>IV</u> . In which some more Firs
, far away. CHAPTER < <u>rromanNumeral>IX</u> . In which the Wooden Mid
. "Anything about a < <u>romanNumeral>L</u> ," said the Captain, "or a
f victuals. CHAPTER < <u>rromanNumeral>L</u> . Mr Toots's Complaint The
he chips out P-A-U- <u><romannumeral>L</romannumeral></u> in the marble slab before
r, jintly!" CHAPTER < <u><romannumeral>LI</romannumeral></u> . Mr Dombey and the World
e together. CHAPTER <u><romannumeral>LII</romannumeral></u> . Secret Intelligence Go

Unfortunately, without linguistic analysis, the ambiguity of the language prevents us from distinguishing a letter from a Roman numeral. However, one could assume that in general (except at the end of a sentence!), a Roman numeral is not followed by a period or an apostrophe (as *D'ye want* for *Do you want*). We could add this rule to our graph, with a negative right context (see Section 1.7.3, page 55). In fact, we were lucky so far, because we could have had in the concordances, from the first graph, a confusion either with the letters *V* or *X*, possibly used in an abbreviation. Let us finish our complete recognition of all the Roman numerals and we will see later what we can do about ambiguities (Section 3.5, page 68).

3.4 Fourth step: from 2 to 3 999

The numbers greater than 3 999 are generally not written as Roman numerals because there are no symbols for 5 000 and 10 000.

3.4.1 The subgraph from 1 to 999

We will repeat the same procedure for the third time. Starting with the *RN2-999.grf* graph we will modify it by canceling the morphological mode and the insertion of tags. We should not forget to replace the call to the *RN2-99.grf* subgraph with *RN1-99-subgraph.grf* because the later subgraph does not insert tags, does not impose morphological mode and recognizes *I*.



3.4.2 The graph from 2 to 3 999

In order to recognize thousands (one, two and three thousand) it is enough to take the graph *RN2-999.grf*, to replace in it, *C* with *M*, and to delete all other paths. As before, we will replace
a call to the subgraph with *RN2-999.grf* outside the morphological mode (for numerals between 2 to 999), and a call to the subgraph with *RN1-999-subgraph.grf* (inside the morphological mode) for numerals between 1 to 999 when the are added to thousands. We will name this graph *RN2-3999.grf*.



The application of this graph to our text yields 81 matches. Five new false recognitions are added to those previously detected by the occurrence of "M" with abbreviations, i.e. *M.A..*: *Master of Arts* (see the Wikipedia).⁶³

Concordance: U:\Documents\My	Unitex/English/Corpus/DombeyAndSon/DombeyAndSon_snt/concord.html 🗖 🗹 🛛
81 matches	
too much!" CHAPTER	<pre><romannumeral>LXI</romannumeral>. Relenting Florence had</pre>
f a dream." CHAPTER	<pre><romannumeral>LXII</romannumeral>. Final A bottle that have</pre>
rend Alfred Feeder,	<pre><romannumeral>M</romannumeral>.A., who was to perform the</pre>
, Mr Alfred Feeder,	<pre><romannumeral>M</romannumeral>.A., had already gone on,</pre>
. Mr Alfred Feeder,	<pre><romannumeral>M</romannumeral>.A., was afterwards very have</pre>
d Mr Alfred Feeder,	<pre><romannumeral>M</romannumeral>.A., on the way home, that</pre>
d. "I say! Miss D.O.	. <pre><romannumeral>M</romannumeral>. you know!" The Captain, </pre>
Daughter!" CHAPTER	<pre><romannumeral>V</romannumeral>. Paul's Progress and Chrice</pre>
ering down. CHAPTER	<pre><romannumeral>VI</romannumeral>. Paul's Second Deprivatic-</pre>
s about it. CHAPTER	<pre><romannumeral>VII</romannumeral>. A Bird's-eye Glimpse o:</pre>
y or a Son. CHAPTER	<pre><romannumeral>VIII</romannumeral>. Paul's Further Progres</pre>
o the last. CHAPTER	<pre><romannumeral>X</romannumeral>. Containing the Sequel of =</pre>
you'd do!" CHAPTER	<pre><romannumeral>XI</romannumeral>. Paul's Introduction to </pre>
ver coming. CHAPTER	<pre><romannumeral>XII</romannumeral>. Paul's Education After</pre>
oared away! CHAPTER	<pre><romannumeral>XIII</romannumeral>. Shipping Intelligence</pre>
onfinement? CHAPTER	<pre><romannumeral>XIV</romannumeral>. Paul grows more and mo;</pre>
his enemy. CHAPTER	<pre><romannumeral>XIX</romannumeral>. Walter goes away The weak and the second second</pre>
the street. CHAPTER	<pre><romannumeral>XL</romannumeral>. Domestic Relations It water</pre>
eet again!" CHAPTER	<pre><romannumeral>XLI</romannumeral>. New Voices in the Wave:</pre>
ife withal. CHAPTER	<pre><romannumeral>XLII</romannumeral>. Confidential and Acci</pre>
er husband. CHAPTER	<pre><romannumeral>XLIII</romannumeral>. The Watches of the N:</pre>
ut in you!" CHAPTER	<pre><romannumeral>XLIV</romannumeral>. A Separation With the</pre>
a li stunke d aus parts	Variable and the state of the s

The continuation of this exercise (Section 3.5) is only for confident users.

3.5 Ambiguities of Roman numerals (only for confident users)

Before starting this section

If you have not modified the *Norm.txt* file as explained in <u>Chapter 2, Section 3.1.2</u>, page 35, do so, or download it, then close and reopen Unitex.

As we said in <u>Section 3.3.2</u>, page 66, Roman numerals are ambiguous with letters or abbreviations; and it is almost impossible to disambiguate them. We can reduce these ambiguities by assuming that a Roman numeral is not followed by a full stop or an apostrophe. However, false annotations cannot be completely avoided, because a Roman numeral can occur at the end of a sentence and thus be followed by a period. Let us also add a list of seven common words that are ambiguous with a number written in Roman numerals: *CD* (compact

⁶³ the Reverend Alfred Feeder, M.A., who was to perform the ceremony.

disc), *CI* (criminal investigation), *CM* (College of Medicine), *CV* (curriculum vitae), *DC* (District of Columbia), *MI* (Michigan) and *MM* (abbreviation of month or of millimeter).⁶⁴

3.5.1 The subgraph of ambiguous Roman numerals

Let us create a graph based on the above remarks. This graph, named *ambiguousRN.grf* does not insert tags.



3.5.2 The use of weights

If we create a graph with two subgraphs, *RN2-3999* and *ambiguousRN*, and if we use it to recognize, for instance, the sequence *CD*, we do not know if the result will be <romanNumeral>CD</romanNumeral> (*:RN2-3999* path) or *CD* (*:ambiguousRN* path). We would not like to annotate *CD* as a Roman numeral. The *Longest matches* option cannot help here, because both annotations are equally long. In order to get what we want, we have to use weights. We will save this graph as *nonAmbiguousRN.grf*.

One can assign a weight to a box by inserting $\{weight\}\$ in its output where *weight* is a positive integer or zero (*weight* ≥ 0). The path with the highest last weight has the priority. The use of weights is interesting for graphs that produce output because it enables choosing an output when there are several possibilities (as in our case). One must keep in mind that weights apply only to cases when two (or more) graph paths recognize exactly the same sequence of tokens; otherwise, the *Longest matches* option reduces the range of possibilities to those paths that recognize the longest sequence. If one path has several weights, only the last one is applied. A path with a weight always has priority over a path without any weight.



If you apply the *nonAmbiguousRN* graph to the text, you will obtain 81 concordance lines, but only 65 with the *romanNumeral* tags. Of these, only the seven capital letters are errors while

⁶⁴ This list is language dependent. For example, for French, DIX is the number ten, and so on.

the other tags correspond to 58 chapters (the first is not counted, but also the fifth, tenth and fiftieth, due to the period after the Roman numeral).

🔲 Concordance: U:Documents/MyUnitex/English/Corpus/DombeyAndSon/DombeyAndSon_snticoncord.html 📃 🖆 🗹
81 matches
which he passed to < <u>romanNumeral>C</u> , and even D before dinner
ssed to C, and even < <u>romanNumeral>D</u> before dinner. It was har
n went so far as to < <u>romanNumeral>D</u> - Mr Baps to Lady Skettles
ould have excused: " <u><romannumeral>D</romannumeral></u> d bad," "Jack Bunsby," wl
ly, on the table. "` <u><romannumeral>D</romannumeral></u> ,'" the old woman read alo
the world. CHAPTER < <u>rromanNumeral>II</u> . In which Timely Provisi
accustomed. CHAPTER < <u>romanNumeral>III</u> . In which Mr Dombey, as
hstanding!" CHAPTER < <u>romanNumeral>IV</u> . In which some more Firs
, far away. CHAPTER < <u>rromanNumeral>IX</u> . In which the Wooden Mid:
. "Anything about a < <u>romanNumeral>L</u> ," said the Captain, "or a
he chips out P-A-U- <u><romannumeral>L</romannumeral></u> in the marble slab before
r, jintly!" CHAPTER <u><romannumeral>LI</romannumeral></u> . Mr Dombey and the World
e together. CHAPTER <u><romannumeral>LII</romannumeral></u> . Secret Intelligence Go
him of it. CHAPTER < <u>romanNumeral>LIII</u> . More Intelligence The:
was empty. CHAPTER < <u>romanNumeral>LIV</u> . The Fugitives Tea-time
intry wind. CHAPTER <u><romannumeral>LIX</romannumeral></u> . Retribution Changes ha
re shining. CHAPTER <u><romannumeral>LV</romannumeral></u> . Rob the Grinder loses h
n of ashes. CHAPTER <u><romannumeral>LVI</romannumeral></u> . Several People delight(
dear love. CHAPTER <u><romannumeral>LVII</romannumeral></u> . Another Wedding Mr Sov
y far away! CHAPTER ">comanNumeral> . After a Lapse The sea
isposed of. CHAPTER < <u><romannumeral>LX</romannumeral></u> . Chiefly Matrimonial The
too much!" CHAPTER < <u><romannumeral>LXI</romannumeral></u> . Relenting Florence had
A dream W GUADMUD (parallement) TVTT (arment) ment) Dinel & bettle thet 1

Chapter 4: CasSys

Before starting this chapter **1.** Go to the folder: MyUnitex \English \Corpus \UnitexGettingStarted \UnitexGettingStarted Create three new folders named (without space): JubileePlainText and JubileeXmlText. Go to the folder: MyUnitex\English\Corpus\UnitexGettingStarted\UnitexGettingStarted\JubileePlainText Download the jubileePlainText.txt file. Go to the folder: *MyUnitex**English**Corpus**UnitexGettingStarted**JubileeXmlText* Download the JubileeXmlText.xml file.65 2. Go to the folder: *MvUnitex**English**Graphs**UnitexGettingStarted* Create two new folders named Entities and Measures. Go to the folder: *MyUnitex**English**Graphs**UnitexGettingStarted**Entities* Create two new folders named Analysis and Synthesis. Go to the folder: *MyUnitex**English**Graphs**UnitexGettingStarted**Measures* Create two new folders named Analysis and Synthesis. 3. Go to the folder: MyUnitex\English\CasSys Create a new folder named UnitexGettingStarted (without space). Go to the folder: *MyUnitex**English**CasSys**UnitexGettingStarted* Create three new folders named PlainTextEntities, XmlTextEntities and Measures.

1 First example: named entity recognition in a plain text

We will introduce cascades that analyze texts with aim to annotate named entities in a text (dates and places). First, we will use the text *jubileePlainText.txt*.

This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on June 2, 2022 and end on June 5. The Queen ascended the British throne on February 6, 1952. Our small town of Nearlondon will also celebrate this event. Nearlondon has always been a loyal

subject of Her Majesty. Although the city of London is dear to us, on Thursday June 2 we will change the name of London Street to Queen Elizabeth Street. Everyone will fondly remember 2022!

God Save the Queen!

We will open this text in Unitex and answer Yes to the question *Do you want to preprocess the text*. A window will open in which we will remove the first two checkmarks, leaving just the last one, *Apply All default Dictionaries*.

1.1 Analysis cascade

All graphs in the Section 1.1 will be saved in the folder:

⁶⁵ These two texts were prepared specifically to illustrate this chapter.

1.1.1 The absoluteDate.grf graph

Let us create a first graph, which will recognize dates where the year is specified; we will call such dates *absolute*. In order to simplify this graph, we will not describe the days of months and years in detail but will instead use the lexical mask *<NB>*, which recognizes strings consisting of digits. We will put the recognized date between symbols { and *,.AbsoluteDate*}. By doing so, we are producing a lexical tag.⁶⁶ Assuming that a date in a text is *June 23, 2022*, the graph we are building will recognize it and replace it with a lexical tag {June 23}, 2022,.AbsoluteDate}. This text segment does not consist of eight tokens anymore, but just one, *June 23, 2022* and this token is classified as *AbsoluteDate*; in other words, the part-of-speech field in the lexical tag contains AbsoluteDate, in the same way as *June* was classified before as *N* (noun) by the application of dictionaries.⁶⁷

In this graph, we will duplicate the box containing the names of months in order to avoid opening braces twice in the same path. We will add a tracing feature to the recognized sequence: the name of the graph itself +grf=absoluteDate, which can be useful in the debugging process. We will save this graph as absoluteDate.grf. In order to use the graph in a cascade (see Section 1.1.3, page 73), the saved graph must be compiled by clicking the Compile graph button at the left corner of the graph processing toolbar, next to the Save graph button.



We will create a concordance with this graph. After clicking the *SEARCH* and *OK* buttons, the *Located sequences* window appears. At the bottom right of this window, we will open the *Sort according to* drop-down menu and choose the *Text order* option, instead of *Center, Left*. We will click on the *Build concordance* button and the concordance lines, in this case two lines, will be displayed in the order in which matched sequences appear in the text.

⁶⁶ See <u>Chapter 2, Section 1.2.5</u>, page 15. The common way to insert a { character is by the box that recognizes the beginning of the phrase you want to tag. If the { character is inserted just before, by a separate <E> box, the inserted { character can be followed by a space in the output, which is unsatisfactory because the braces are supposed to delimit a precise phrase.

⁶⁷ Remember that each alphabetic string is one token (*June* in this case), a punctuation one token (, in this case), and each digit is a separate token (2, 3, 2, 0, 2, 2 in this case).



1.1.2 The relativeDate.grf Graph

We will now create the graph that recognizes relative, or incomplete, dates, when the year is not mentioned. We will use the previously constructed *absoluteDate.grf* graph, remove in it the boxes for year and separating comma and save it under the new name *relativeDate.grf* (we will use the *FSGraph/Save as...* menu). Do not forget to modify the output as well: *...AbsoluteDate+grf=absoluteDate}* should be replaced by *...RelativeDate+grf=relativeDate}*. After compiling the graph, we will apply it to the text.



Now we will obtain four concordance lines in which all dates will be annotated as relative dates, including two recognized by the previous graph (note that in these cases, the years are outside the scope of the lexical tags). In order to correct this, we will apply these two graphs one after the other, using what is known as a cascade of graphs.

Concordance: U:\Docum	ents\MyUnitex\English\Corpus\UubileePlainText\jubileePlainText_snt\concord.html 📃 🗖 🛛
4 matches	
party will start	on <u>{June 2,.RelativeDate+grf=relativeDate}</u> , 2022 and end on
2, 2022 and end	on <u>{June 5,.RelativeDate+grf=relativeDate}</u> . The Queen ascen
e British throne	on <u>{February 6,.RelativeDate+grf=relativeDate}</u> , 1952. Our s
n is dear to us,	on {Thursday June 2,.RelativeDate+grf=relativeDate} we will

1.1.3 The analysis.csc cascade

Now we will build a small cascade containing both graphs. We will open the *Text/Apply CasSys Cascade...* menu and click the *New* button. The Cascade configuration frame will open. Then we must find the location of the graphs that we want to put into the cascade, in this case:

We will notice that the left window shows only files with the *.fst2* extension that is only graphs that have been compiled. It means that we can put into a cascade only already compiled graphs.⁶⁸

With the mouse, we can drag and drop the *absoluteDate.fst2* graph into the right window; and we will repeat the same with the *relativeDate.fst2* graph. Alternatively, we can select a graph and click the *Add* button when we want to add it to a cascade. We will save this cascade as *analysis.csc* in the folder:

MyUnitex\English\CasSys\UnitexGettingStarted\PlainTextEntities

by using the *Save* button.

CasSys Transducer Configuration : (unsaved)	¤ [≮] ∅
Look [n: Analysis AbsoluteDate fst2 relativeDate fst2 File Name: relativeDate fst2 Files of Type: 1fst2	# Disabled Name Merge Replace Until Fix Point 1 absoluteDate fst2 Image: Comparison of the point of the

We will close this frame by clicking the *Close* button.

Now we will select our cascade and click the *Launch* button, which will initiate the analysis. When analysis is over, we can build concordances. We will obtain four concordance lines, but now the dates will be correctly annotated. Why is it so? Because the first graph in the cascade, *absoluteDate.grf*, has annotated all absolute dates and produced lexical tags. As we explained before, a lexical tag is treated as one token, and the subsequently applied graph *relativeDate.grf* cannot analyze inside its content.

Concordance: U:\Documents\MyUnitex\English\Cor	rpus\JubileePlainText\jubileePlainText_snt\concord.html 📃 🗗 🗹
4 matches	
e British throne on {February	6 1952,.AbsoluteDate+grf=absoluteDate}. Our
party will start on {June 2	2022, AbsoluteDate+grf=absoluteDate} and end or
2, 2022 and end on {June 5,.R	RelativeDate+grf=relativeDate}. The Queen ascen
n is dear to us, on <u>{Thursday</u>	June 2,.RelativeDate+grf=relativeDate} we will

1.1.4 The town.grf graph

We will now prepare the graph that will recognize names of towns and cities and annotate them as *Town*.⁶⁹ We will presume that any sequence of letters of the English alphabet

⁶⁸ If we use subgraphs, only the main graph needs to be compiled, not the subgraphs.

⁶⁹ One has to be careful when choosing the annotation tags in order to avoid ambiguity with dictionary lemmas. For instance, if we had chosen *town* instead of *Town* as a tag for inhabited places, then the lexical mask *<town>* would correspond to recognized names of towns, but also the lemma *town*, that is, its forms *town* and *towns*. On the contrary, the lexical mask *<Town>* corresponds only to the annotated names of towns (remember that lexical masks are case sensitive).

(<LETTER>) starting with an uppercase letter (<UPPER>) and following the sequences *city of* or *town of* is the name we want to annotate. In order to recognize such a sequence, we must use the morphological mode in order to describe the structure of a single word.⁷⁰ We will save this graph as *town.grf* and compile it.



We will drag and drop this graph to the last position in the cascade. As a reminder: open the *Text/Apply CasSys Cascade...* menu, then, find the folder:

MyUnitex\English\Graphs\UnitexGettingStarted\Entities\Analysis

select the *analysis.csc* cascade, click the *Edit* button, move to the folder containing the graphs, then drag and drop the *town.fst2* graph to the list of graphs. Save the cascade.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		absoluteDate.fst2	~			
2		relativeDate.fst2	~			
3		town.fst2	V			

After launching this cascade two new concordance lines will be produced by the *town.grf* graph with *Nearlondon* and *London* annotated as *Town*.

Concordance: U1DocumentsMyUnitexIEnglish1CorpusUublieePlainTextjubileePlainText_snttconcord.html 🗖 🗹
6 matches
e British throne on <u>{February 6 1952,.AbsoluteDate+grf=absoluteDate}</u> . Our 🖆
party will start on <u>{June 2 2022,.AbsoluteDate+grf=absoluteDate}</u> and end or
2, 2022 and end on {June 5,.RelativeDate+grf=relativeDate}. The Queen ascen
n is dear to us, on <u>{Thursday June 2,.RelativeDate+grf=relativeDate}</u> we will
jesty. Although the <u>city of {London,.Town+grf=town}</u> is dear to us, on Thursd
6, 1952. Our small <u>town of {Nearlondon,.Town+grf=town}</u> will also celebrate

1.1.5 Generalization graphs

If you look at the short text we are annotating, you will notice that *Nearlondon* in the sequence *Nearlondon has always been...* was not annotated as *Town* because it was preceded neither with *town of* nor *city of.* Moreover, *Nearlondon* is not even in the dictionary of the text, it is an unknown word.

In order to solve this problem we will use a generalization graph which will annotate a word (or a sequence of words) even when it does not appear in a required context, if the same word appeared somewhere else in the same text in the required context, and was annotated by some graph in a cascade. We have such a situation with the word *Nearlondon* in our text: it was annotated once as *Town* by the *town.grf* graph, because it occurred in the required context, so its second occurrence will be annotated as *Town* by the *townGeneralization.grf* graph.



In order to create a simple generalization graph it is enough to introduce a box with the output consisting of the annotation that has to be generalized (in our case *,.Town*), select this box and

⁷⁰ See <u>Chapter 3, Section 3.1</u>, page 63. In CasSys cascades, the *<FIRST>* or *<UPPER>* lexical mask does not always detect if a word has an uppercase initial letter, because some words or expressions may have been replaced by lexical tags, as in the preceding examples. This is why we show another way of detecting words with an uppercase initial.

click to the *Insert generic graph mark before the selected box* button at the far right end of the toolbar.⁷¹



After compiling this graph, we will add it as the last one in the list of graphs in the *analysis.csc* cascade (remember that it has to be in the list after the *town.grf* graph) in the same way as we have done before.⁷² We must do one more thing: the configuration panel offers options for the modes of graph application. By default, this option is *Merge* mode, which suited us so far for all the graphs in our list. However, for this last graph we have also to check the *Generaliz...* box in order to signal that this is a generalization graph.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		absoluteDate.fst2	~			
2		relativeDate.fst2	2			
3		town.fst2	~			
4		townGeneralization.fst2	~			2

After launching our cascade, we will obtain a concordance with two new lines, one of them with *Nearlondon* annotated although it was preceded neither with *city of* nor *town of*. We can see that the second occurrence of *London* was annotated as well, although we did not want that, since the context is *change the name of London Street to Queen Elizabeth Street*.

Concordance: U:\Documents\MyUnitex\English\CorpusUubileePlainText\jubileePlainText_snt\concord.html
8 matches
party will start on <u>{June 2 2022,.AbsoluteDate+grf=absoluteDate}</u> and end of
2, 2022 and end on {June 5,.RelativeDate+grf=relativeDate}. The Queen ascen
e British throne on <u>{February 6 1952,.AbsoluteDate+grf=absoluteDate}</u> . Our
6, 1952. Our small town of {Nearlondon,.Town+grf=town} will also celebrate
lebrate this event. <u>{Nearlondon,.Town+grf=townGeneralization}</u> has always been
jesty. Although the city of {London,.Town+grf=town} is dear to us, on Thursda
n is dear to us, on <u>{Thursday June 2,.RelativeDate+grf=relativeDate}</u> we will
change the name of <u>{London,.Town+grf=townGeneralization}</u> Street to Queen El:

1.1.6 The street.grf graph

We will now create a graph that recognizes names of streets with the *Address* category. We will recognize two patterns before the words *street* and *avenue*: (1) a date or a town name already recognized by the cascade and tagged as *<AbsoluteDate>*, *<RelativeDate>* or *<Town>*;⁷³ (2) one or more upper-case words. In order to make the loop for uppercase words, we can duplicate two boxes (the big angular bracket and the lexical mask *<UPPER>*), but without the opening braces as output. Alternatively, we can add the new *<UPPER>* box, surround it with morphological mode angular brackets by clicking the *Surround box selection with morphological mode tags* button, and then delete the unnecessary closing bracket and connect the box with *<LETTER>*.

⁷¹ Generalization graphs for annotation can be more sophisticated; see the <u>Section 4.5.2</u>, page 104.

⁷² The *townGeneralization.grf* file must be compiled before we place it in the cascade, but the program uses also the *.grf* file which must therefore also be present in the folder, whereas this is not obligatory for other graphs.

⁷³ This enables recognition of *May 8, 1945 Avenue*.



After saving the graph with the name *street.grf* we will compile it and add it as the last graph in our *analysis.csc* cascade to be used in the *Merge* mode. After launching the cascade and producing a concordance, we will obtain nine concordance lines. One new line contains the match *Queen Elizabeth Street* recognized by the second rule in the *street.grf* graph (two upper-case words). The first rule is applied to *London Street* since *London* was already tagged as *Town*. If we look at that line in the concordance list (it is the second-to-last line in the *Text Order* option), we will see that there we have nested tagging, that is a lexical tag inside another lexical tag.⁷⁴

8		
Concordance: U:\Documents\MyUnitex\Eng	glish\CorpusUubileePlainTexttjubileePlainText_snttconcord.html 🗖 🗖	\boxtimes
9 matches		
party will start on {June	2 2022,.AbsoluteDate+grf=absoluteDate} and end on June 5. The Queen ascen	
2, 2022 and end on {June	5,.RelativeDate+grf=relativeDate}. The Queen ascended the British throne or	2
e British throne on {Febr	<pre>uary 6 1952,.AbsoluteDate+grf=absoluteDate}. Our small town of Nearlondon</pre>	
6, 1952. Our small <u>town</u>	<u>of {Nearlondon,.Town+grf=town}</u> will also celebrate this event. Nearlondon ha	
lebrate this event. <u>{Near</u>	<pre>london,.Town+grf=townGeneralization} has always been a loyal subject of Her</pre>	
jesty. Although the city	of {London,.Town+grf=town} is dear to us, on Thursday June 2 we will change	
n is dear to us, on <u>{Thur</u>	<u>sday June 2,.RelativeDate+grf=relativeDate}</u> we will change the name of Londo	
change the name of ${\{ \ Lo \ }$	ndon\.Town\+grf=townGeneralization\} Street,.Address+grf=street} to Queen	
of London Street to <u>{Quee</u>	n Elizabeth Street,.Address+grf=street}. God Save the Queen!	
		۲.

1.1.7 Created files

The role of the Cassys program is not only to produce concordances; it produces as output numerous files. If we open the folder:

*MyUnitex**English**Corpus**UnitexGettingStarted**jubileePlainText*

we will see in it, besides the original *jubileePlainText.txt* file, also *jubileePlainText.snt* and the *jubileePlainText_snt* folder, as for all texts processed by Unitex. Besides this, we will see four more files with the suffix _csc and one new folder.

Nom	Modifié le 🎽	Туре	Taille
jubileePlainText_csc.raw	22/03/2023 09:53	Fichier RAW	1 Ko
📔 jubileePlainText_csc.txt	22/03/2023 09:53	Fichier TXT	2 Ko
📔 jubileePlainText_csc_raw_offsets.txt	22/03/2023 09:53	Fichier TXT	1 Ko
📔 jubileePlainText_csc_txt_offsets.txt	22/03/2023 09:53	Fichier TXT	1 Ko
📔 jubileePlainText.snt	20/03/2023 16:12	Fichier SNT	1 Ko
📔 jubileePlainText.txt	26/10/2022 09:19	Fichier TXT	1 Ko
📙 jubileePlainText_snt	22/03/2023 09:53	Dossier de fichiers	
jubileePlainText_csc	22/03/2023 09:53	Dossier de fichiers	

1.1.7.1 The _csc files

The *jubileePlainText_csc.raw* file contains the text modified by the cascade. Namely, all sequences recognized by cascade graphs are replaced by lexical tags enclosed in braces.

⁷⁴ In order to do that the delimiting characters of the internal lexical tag $({.+})$ are protected by the backslash character.

This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on {June 2\, 2022, AbsoluteDate+grf=absoluteDate} and end on {June 5, RelativeDate+grf=relativeDate}. The Queen ascended the British throne on {February 6\, 1952, AbsoluteDate+grf=absoluteDate}. Our small town of {Nearlondon,.Town+grf=town} will also celebrate this event. {Nearlondon,.Town+grf=townGeneralization} has always been a loyal subject of Her Majesty. Although the city of {London,.Town+grf=town} is dear to us, on {Thursday June 2,.RelativeDate+grf=relativeDate} we will change the name of {\{London\,.Town\+grf=townGeneralization\} Street,.Address+grf=street} to {Queen Elizabeth Street,.Address+grf=street}. Everyone will fondly remember 2022! God Save the Queen!

The *jubileePlainText_csc.txt* file is a quasi XML version of the same output in which the braces are replaced by XML tags *<csc></csc>*, the recognized text is surrounded by tags *<form></form>*, while the code in the POS field and its additional features appear in tags *<code></code>.*⁷⁵

This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on <csc><form>June 2. 2022</form><code>AbsoluteDate</code><code>grf=absoluteDate</code></csc> and end on <csc><form>June 5</form><code>RelativeDate</code><code>grf=relativeDate</code></csc>. The Queen ascended the British throne on <csc><form>February 6. 1952</form><code>AbsoluteDate</code><code>grf=absoluteDate</code></csc>. Our small town of <csc><form>Nearlondon</form><code>Town</code><code>grf=town</code></csc> will also celebrate this event. <csc><form>Nearlondon</form><code>Town</code><code>grf=townGeneralization</code></cs c> has always been a loyal subject of Her Majesty. Although the city of <csc><form>London</form><code>Town</code><code>grf=town</code></csc> is dear to us, on <csc><form>Thursday June 2</form><code>RelativeDate</code>code>grf=relativeDate</code></csc> we will change the name of <csc><form><csc><form>London</form><code>Town</code><code>grf=townGeneralization</c ode></csc> Street</form><code>Address</code>code>grf=street</code></csc> to <csc><form>Queen Elizabeth Street</form><code>Address</code><code>grf=street</code></csc>. Everyone will fondly remember 2022! God Save the Queen!

The remaining two files contain offsets corresponding to two previously described files.⁷⁶

1.1.7.2 The _csc folder

The folder with the name ending with _*csc* contains all intermediary files, which are very useful in debugging cascades.

⁷⁵ <*lemma*></*lemma*> is one more pair of XML tag that can appear in the content of <*csc*></*csc*>. It does not appear in our case because in the description of all lexical tags the lemma field (between the comma and the period) is empty. If it were not the case, this field would be copied here enclosed in <*lemma*></*lemma*> tags.

⁷⁶ See the Unitex user manual, Section 15.13.10.

Nom	Modifié le	Туре	Taille
📔 jubileePlainText_5_0.snt	22/03/2023 09:53	Fichier SNT	1 Ko
📔 jubileePlainText_0_0.snt	22/03/2023 09:53	Fichier SNT	1 Ko
jubileePlainText_1_0.snt	22/03/2023 09:53	Fichier SNT	1 Ko
📔 jubileePlainText_2_0.snt	22/03/2023 09:53	Fichier SNT	1 Ko
🔐 jubileePlainText_3_0.snt	22/03/2023 09:53	Fichier SNT	1 Ko
🔐 jubileePlainText_4_0.snt	22/03/2023 09:53	Fichier SNT	1 Ko
jubileePlainText_0_0_snt	22/03/2023 09:53	Dossier de fichiers	
jubileePlainText_1_0_snt	22/03/2023 09:53	Dossier de fichiers	
jubileePlainText_2_0_snt	22/03/2023 09:53	Dossier de fichiers	
jubileePlainText_3_0_snt	22/03/2023 09:53	Dossier de fichiers	
jubileePlainText_4_0_snt	22/03/2023 09:53	Dossier de fichiers	
jubileePlainText_5_0_snt	22/03/2023 09:53	Dossier de fichiers	

The original text is copied in this folder and renamed as *jubileePlainText_0_0.txt*. The result of the first graph in the cascade is given the name *jubileePlainText_1_0.txt*, and so on. If some graph in the cascade is used in the *until fix point* mode (this will be explained in the following sections) it means that it will be repeatedly used and the second number in a file's name tells which iteration produced the file (in our present example it is always 0).

1.2 Synthesis cascade

We have seen in the previous section that there are two outputs of an analysis cascade.

1. The *_csc.raw*

In the *_csc.raw* version of the text the recognized matches are transformed into lexical tags that can be subsequently used in Unitex for further text exploration. It can be used in Unitex for searching in the usual way. We can open the text *jubileePlainText_csc.raw*, without preprocessing, and then launch with *Locate pattern* a search using classical as well as new categories.⁷⁷ For instance, a pattern <*Address>* retrieves two matches:

Concordance: U:\Documents\MyUnite	x\English\Corpus\UubileePlainText\jubileePlainText_csc_snt\concord.html	r 🛛 🖂
2 matches		
change the name of $\{ \setminus \}$	{London\.Town\+grf=townGeneralization\} Street,.Address+grf=street} to	{Queen E
ress+grf=street} to {Qu	<pre>leen Elizabeth Street,.Address+grf=street}. Everyone will fondly remembe:</pre>	r {\{2022
	J	•

2. The _csc.txt

The *_csc.txt* version of the text is the resulting text where the lexical tags are replaced by XML tags. This format is used to transform the output of an analysis cascade into whatever other format needed.

We will now open (with the *Text/Open* menu) the file:

*MyUnitex**MyUnitex**English**Corpus**UnitexGettingStarted*\ *jubileePlainText**jubileePlainText_csc.txt*

and reply No to the question Do you want to preprocess the text.

We will delete tags <*csc*>, <*form*> and so on, and replace them with tags that we have chosen, namely, <*AbsoluteDate*></*AbsoluteDate*>, <*RelativeDate*></*RelativeDate*>, <*Town*></*Town*> and <*Address*></*Address*>.

1.2.1 The tag.grf graph

Let us create the first graph, which we will save in the folder:

MyUnitex\English\Graphs\UnitexGettingStarted\Entities\Synthesis

⁷⁷ That is to say <*N*>, <*V*> or <*ADV*>, but also, <*AbsoluteDate*>, <*Town*> or <*Address*>.

In this folder we will save the rest of the graphs for synthesis that we will introduce in this section. The aim of this graph that we will use in *Replace* mode is:

- 1. delete tags csc, form and code;
- 2. save the matched text in a variable *\$text\$* and its category in a variable *\$tag\$*;
- 3. replace all that with the tagged text: <*\$tag\$>\$text\$</\$tag\$>*; that is, everything recognized is replaced (thus the graph must be used in the *Replace* mode) by the recognized entity (*\$text\$*) enclosed in appropriate tags (<*\$tag\$></\$tag\$>*).



An input variable in Unitex stores the part of a text recognized by boxes inside (big red) parentheses. These parentheses are placed in the same way as context and morphological mode brackets by clicking on the *Surround box selection with an input variable* button (red parentheses at the right-hand side of the toolbar).

The user can freely choose the name of the variable. If the value of a variable is used in output, then it is surrounded by a pair of \$ character.

1.2.1.1 A first attempt

In order to collect the text we will use a loop that recognizes a continuous sequence of tokens (lexical mask: *<TOKEN>*) except that it will exclude *</form>* (the tag at the end of the recognized text).⁷⁸ After the loop, the graph will match one of the used categories, followed by everything until the end of the coded sequence that we are analyzing (*</csc>*).

A very important note: in Unitex queries, angle brackets have the special meaning, so if we want to recognize an angle bracket in a text, we have to *protect* it by preceding it with a backslash.⁷⁹ For instance, if we want to recognize *<csc>* in a text we have to write *\<csc\>* (otherwise it would be interpreted as a lexical mask recognizing all forms of the lemma *csc*). Similarly, since a slash also has a special meaning in Unitex graph boxes (it marks the beginning of an output) if we want to recognize it in a text, it also has to be preceded by a backslash. For instance, in order to recognize *</csc>* we must write *\<\/csc\>*.

In order to make this complex graph easier to understand, we will provide an example inside a comment box (a box starting with a slash and not connected with any other boxes). Text in such a box will appear in green.

⁷⁸ The negative right contexts (see <u>Chapter 3, Section 1.7.3</u>, page 55) are interpreted like this: *<TOKEN>* is recognized if it is not the < character, followed by */form>* (first negative right context), respectively */csc>* (second negative right context).

⁷⁹ See the Unitex user manual, Section 5.2.7. We do not have to take care about this in the output text - no protection is needed there.



<csc><form>June 2, 2022</form><code>AbsoluteDate</code><code>grf=absoluteDate</code></csc>

We will now save the graph as *tag.grf*; compile it; create the new cascade in which we will add in the first position this graph; and choose the *Replace* mode by checking the appropriate box. We will call this cascade *synthesis.csc* and we will save it in the same folder as the *analysis.csc* cascade.



After launching this cascade, we will obtain a new text file:⁸⁰

MyUnitex\MyUnitex\English\Corpus\UnitexGettingStarted\ jubileePlainText\jubileePlainText_csc_csc.txt

This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on <AbsoluteDate>June 2, 2022</AbsoluteDate> and end on <RelativeDate>June 5</RelativeDate>. The Queen ascended the British throne on <AbsoluteDate>February 6, 1952</AbsoluteDate>. Our small town of <Town>Nearlondon</Town> will also celebrate this event. <Town>Nearlondon</Town> has always been a loyal subject of Her Majesty. Although the city of <Town>London</Town> is dear to us, on <RelativeDate>Thursday June 2</RelativeDate> we will change the name of <Town><csc><form>London</Town> Street</form><code>Address</code><code>grf=street</code></csc> to <Address>Queen Elizabeth Street</Address>. Everyone will fondly remember 2022! God Save the Queen!

We can observe in the resulting text that the transformation of annotations of non-nested elements was done successfully. For instance, the occurrence of the first date in the text became <*AbsoluteDate>June 2, 2022*</*AbsoluteDate>*. However, for nested elements, the result is surprising: <*Town><csc>form>London</Town> Street</form><code>Address* </*code><code>grf=street</code></csc>*! What happened is that the graph generated the opening tag using the internal markup <*Town>*.

⁸⁰ Because the graph in this cascade is used in the *Replace* mode, it is difficult to visualize the output as a concordance. We will therefore from now on display the file obtained as a result, which is in this case *jubileePlainText_csc_csc.txt*.

1.2.1.2 A second attempt

To avoid this, we will add *<csc>* (the beginning of a nested coded sequence) in the first negative context box in order to avoid deletion and replacement in cases when an entity has nested entities in its content.



<csc><form>June 2, 2022</form><code>AbsoluteDate</code><code>grf=absoluteDate</code></csc>

<csc><form><csc><form>London</form><code>Town</code><code>gf=townGeneralization</code></csc> Street</form> <code>Address</code><code>gf=street</code></csc> to <csc><form>Queen Elizabeth Street</form><code>Address</code><code>grf=street</code></csc>

We will compile the graph and relaunch the cascade on *jubileePlainText_csc.snt*.

This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on <AbsoluteDate>June 2, 2022</AbsoluteDate> and end on <RelativeDate>June 5</RelativeDate>. The Queen ascended the British throne on <AbsoluteDate>February 6, 1952</AbsoluteDate>. Our small town of <Town>Nearlondon</Town> will also celebrate this event. <Town>Nearlondon</Town> has always been a loyal subject of Her Majesty. Although the city of <Town>London</Town> is dear to us, on <RelativeDate>Thursday June 2</RelativeDate> we will change the name of <csc> form><Town>London</Town> Street</form><code>Address</code>code>grf=street</code></csc> to <Address>Queen Elizabeth Street</Address>. Everyone will fondly remember 2022! God Save the Queen!

We can now observe in the resulting text that, this time, the nested element was well transformed: ccsc><form><Town>London</Town> Street</form><code>Address</code></code>

1.2.1.3 The Until fix point option

In order to transform the annotations of the external element it would be enough to launch the same graph for the second time (add it to the cascade). However, if we had a three-level nesting, then we would have to launch (and add to the cascade) the graph a third time. In order to avoid this, we will check the *Until Fix Point* box that will launch the *tag.fst2* graph as long as it makes some changes to the text.

	1					
#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		tag fst2		V	V	

After launching again this cascade (applying it to the text *jubileePlainText_csc.txt*) the resulting text *jubileePlainText_csc_csc.txt* becomes correct, and all annotations have been correctly transformed: <*Address*><*Town>London*</*Town> Street*</*Address*>.

This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on <AbsoluteDate>June 2, 2022</AbsoluteDate> and end on <RelativeDate>June 5</RelativeDate>. The Queen ascended the British throne on <AbsoluteDate>February 6, 1952</AbsoluteDate>. Our small town of <Town>Nearlondon</Town> will also celebrate this event. <Town>Nearlondon</Town> has always been a loyal subject of Her Majesty. Although the city of <Town>London</Town> is dear to us, on <RelativeDate>Thursday June 2</RelativeDate> we will change the name of <Address><Town>London</Town> Street</Address> to <Address>Queen Elizabeth Street</Address>. Everyone will fondly remember 2022! God Save the Queen!

If we open the *JubileePlainText_csc_csc* folder, we will find in it four files (and four corresponding folders): the original text *jubileePlainText_csc_0_0.snt*, the result of the first run of the graph (resolves outer-level and one-level annotations), *jubileePlainText_csc_1_0.snt*, the result of the second run of the graph (resolves second-level annotations), *jubileePlainText_csc_1_0.snt*, the result of the result of the third run of the graph, *jubileePlainText_csc_1_2.snt*, that establishes that the text was not modified (the last two files are identical).

Nom	Modifié le 🗡	Туре	Taille
jubileePlainText_csc_1_0.snt	24/03/2023 13:12	Fichier SNT	1 Ko
jubileePlainText_csc_1_1.snt	24/03/2023 13:12	Fichier SNT	1 Ko
jubileePlainText_csc_1_2.snt	24/03/2023 13:12	Fichier SNT	1 Ko
jubileePlainText_csc_0_0.snt	24/03/2023 13:12	Fichier SNT	2 Ko
jubileePlainText_csc_1_0_snt	24/03/2023 13:12	Dossier de fichiers	
jubileePlainText_csc_1_1_snt	24/03/2023 13:12	Dossier de fichiers	
jubileePlainText_csc_1_2_snt	24/03/2023 13:12	Dossier de fichiers	
jubileePlainText_csc_0_0_snt	24/03/2023 13:12	Dossier de fichiers	

1.2.2 The internalDeletion.grf graph

In some cases, we do not want nested annotations. For instance, annotating a date or a city occurring in an address is not obligatorily relevant. While recognizing dates and inhabited places has been essential for recognizing addresses, we would like to remove the date and town tags now that we have used them.

The *internalDeletion.grf* graph that will remove nested tags in the *<Address>* element begins with the recognition of the start tag *<Address>*, which is followed by the recognition of a date tag or a city tag; the content of this second element (start tag and end tag excluded) becomes the value of the variable *\$inside\$*. The end tag *</Address>* is then searched for and everything found between the end tag of the nested annotation and the end tag *</Address>* is remembered in the variable *\$outside\$*. The output of this graph, which works in the *Replace* mode, are tags *<Address></Address>* surrounding the content of the nested element (*\$inside\$*) followed by the content of the *<Address>* element itself (*\$outside\$*).



After saving and compiling the graph, we will add it as the second graph in the *synthesis.csc* cascade (do not forget to check the *Replace* box).

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		tag.fst2		~	~	
2		internalDeletion.fst2		V		

After launching this cascade, we will get the desired result.

This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on <AbsoluteDate>June 2, 2022</AbsoluteDate> and end on <RelativeDate>June 5</RelativeDate>. The Queen ascended the British throne on <AbsoluteDate>February 6, 1952</AbsoluteDate>. Our small town of <Town>Nearlondon</Town> will also celebrate this event. <Town>Nearlondon</Town> has always been a loyal subject of Her Majesty. Although the city of <Town>London</Town> is dear to us, on <RelativeDate>Thursday June 2</RelativeDate> we will change the name of <Address>London Street</Address> to <Address>Queen Elizabeth Street</Address>. Everyone will fondly remember 2022! God Save the Queen!

2 Second example: named entity recognition in an XML text

We will work now with the same text but prepared in the XML format, *jubileeXmlText.xml*. This text is in the folder:

Please note that despite being an XML text it should not have the *.xml* extension because Unitex automatically removes in preprocessing all XML tags from a text having an*.xml* extension. In order to avoid this, we will rename our text to *jubileeXmlText.xml.txt*.

<xml> <title>The municipal newspaper of the city of Nearlondon</title></xml>
<pre><date>Posted on Wednesday June 1, 2022</date></pre>
<pre><body></body></pre>
This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign!
The party will start on June 2, 2022 and end on June 5. The Queen ascended the British throne on
February 6, 1952.
Our small town of Nearlondon will also celebrate this event. Nearlondon has always been a
loyal subject of Her Majesty. Although the city of London is dear to us, on Thursday June 2 we will
change the name of London Street to Queen Elizabeth Street.
Everyone will fondly remember 2022!
God Save the Queen!

We will open the text *jubileeXmlText.xml.txt* in Unitex (*Text/Open*) and answer *Yes* to the question *Do you want to preprocess the text*. A window will open in which we will remove the first two checkmarks, leaving just the last one *Apply All default Dictionaries*. We will launch the *analysis.csc* cascade (*Text/Apply CasSys Cascade*), then open the resulting text, *jubileeXmlText.xml_csc.txt*, and launch the *synthesis.csc* cascade. We will get as a result the *jubileeXmlText.xml_csc_txt* file. Let us look at it:

<xml> <title>The municipal newspaper of the city of <town>Nearlondon</town></title> <date>Posted on <absolutedate>Wednesday June 1, 2022</absolutedate></date> <body></body></xml>
This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign!
The party will start on <absolutedate>June 2, 2022</absolutedate> and end on
<relativedate>June 5</relativedate> . The Queen ascended the British throne on
<absolutedate>February 6, 1952</absolutedate> .
Our small town of <town>Nearlondon</town> will also celebrate this event.
<town>Nearlondon</town> has always been a loyal subject of Her Majesty. Although the city of
<town>London</town> is dear to us, on <relativedate>Thursday June 2</relativedate> we will
change the name of <address>London Street</address> to <address>Queen Elizabeth</address>
Street.
Everyone will fondly remember 2022!
God Save the Queen!

The result of annotating the element *<body>* was the same as before. However, we see that *Nearlondon* and *Wednesday June 1, 2022* appearing in the header were annotated as well although they are not part of the text. We would like to avoid this, because we do not want to annotate the header.

We will use the cascades developed before to which we will add some new graphs. We will copy two cascades (*analysis.csc* and *synthesis.csc*) from the folder:

*MyUnitex**English**CasSys**UnitexGettingStarted**PlainTextEntities*

and paste them in the folder:

2.1 Analysis cascade

We will open the text *jubileeXmlText.xml.txt* in Unitex (*Text/Open*) and answer *No* to the question *Do you want to preprocess the text*.

We will build two new graphs, saved in the folder:

MyUnitex\English\Graphs\UnitexGettingStarted\Entities\Analysis

2.1.1 The toolXml.grf graph

We will assume that the name in an XML tag is a sequence of letters that can be recognized in Unitex with the lexical mask *WORD>*.⁸¹ The *toolXml.grf* graph recognizes XML tags, transforms them into brace-enclosed lexical tags, with *Xml* in the part-of-speech (POS) field and with as features the name of the recognized tag and the indication whether it was an opening or closing tag.⁸²



2.1.2 The toolHidden.grf graph

The role of the *toolHidden.grf* graph is to hide and protect the XML elements that can be found in the text header.⁸³ By doing this, we disable the annotation of entities in the content of hidden elements. This graph transforms the recognized content, together with the opening and closing tags, into a lexical tag categorized as *Hidden* in the POS field.



We will put these two new graphs at the beginning of the *analysis.csc* cascade, which will now consist of seven graphs. These two graphs will work in the *Merge* mode like the rest of the analysis graphs. Do not forget to apply the *analysis.csc* cascade to the original *jubileeXmlText.xml.snt* file.

⁸¹ Of course, XML names can be more complex, but we want to keep our example simple. It would be possible to call in this box a subgraph that describes possible names more precisely.

⁸² This graph is designed for the simple XML document that we are using in our examples. A general graph that would recognize all possible XML tags, including their attributes, would be more complex.

⁸³ This graph only works if the text to hide is not too long. Otherwise, see <u>Section 4.4.1</u>, page 98 (for confident users).

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		toolXml.fst2	V			
2		toolHidden.fst2	~			
3		absoluteDate.fst2	~			
4		relativeDate.fst2	~			
5		town.fst2	~			
6		townGeneralization.fst2	~			~
7		etraat fet?	V			

The result of the first graph in the cascade (*toolXml.grf*), *jubileeXmlText.xml_1_0.snt* file, is recorded in the folder:

We can see that all XML tags in the text have been transformed into lexical tags with *Xml* in the POS field (only the first line is displayed here).

{<xml>,.Xml+name=xml+openTag+grf=toolXml}

The result of the second graph is recorded in the *jubileeXmlText.xml_2_0.snt* file. We can see that two tag-enclosed XML elements, *<title>/</title>* and *<date>/</date>*, which appear in the header, have been transformed into lexical tags *Hidden* (only the second and third lines are displayed here). Note that in order to preserve the literal meaning of the special characters (period, comma, parentheses and so on), they are protected by an escape character backslash when they appear inside the first field of a lexical tag. These escape characters are inserted before the special characters only if the application of the graph has been launched as an item in a cascade.

{\{<title>\,\.Xml\+name=title\+openTag\+grf=toolXml\}The municipal newspaper of the city of Nearlondon\{</title>\,\.Xml\+name=title\+closeTag\+grf=toolXml\}.Hidden+grf=toolHidden} {\{<date>\,\.Xml\+name=date\+openTag\+grf=toolXml\}Posted on Wednesday June 1\, 2022\{</date>\,\.Xml\+name=date\+closeTag\+grf=toolXml\}.Hidden+grf=toolHidden}

2.2 Synthesis cascade

Now, we will open the text *jubileeXmlText.xml_csc.txt* in Unitex (*Text/Open*) and answer *No* to the question *Do you want to preprocess the text*.

We will open (*Graphs/Open* menu or *Graphs/Open Recent* menu) the existing *tag.grf* graph that lists all annotation categories that must remain (the output is *<\$tag\$>\$text\$</\$tag>*). We will modify this graph by adding a path where we list the annotation categories to be removed (*Xml* and *Hidden*); the output is *\$text\$*.



csc><form><xml></form><code>Xml</code><code>name=xml</code><code>openTag</code><code>gf=toolXml</code></csc>

We will save the *tag.grf* graph, compile it and launch the *synthesis.csc* cascade. The resulting *jubileeXmlText.xml* csc csc.txt file will contain exactly what we expected.

<xml>

<title>The municipal newspaper of the city of Nearlondon</title> <date>Posted on Wednesday June 1, 2022</date> <body> This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on <AbsoluteDate>June 2, 2022</AbsoluteDate> and end on <RelativeDate>June 5</RelativeDate>. The Queen ascended the British throne on <AbsoluteDate>February 6, 1952</AbsoluteDate>. Our small town of <Town>Nearlondon</Town> will also celebrate this event. <Town>Nearlondon</Town> has always been a loyal subject of Her Majesty. Although the city of <Town>London</Town> is dear to us, on <RelativeDate>Thursday June 2</RelativeDate> we will change the name of <Address>London Street</Address> to <Address>Queen Elizabeth Street</Address>. Everyone will fondly remember 2022! God Save the Queen! </body> </xml>

3 Third example: measure recognition

We will now go back to the text DombeyAndSon in the folder:

with the idea of recognizing measures in it. We will do that in two phases: first, we will recognize numbers and then units. In order to keep things simple, we will produce graphs that recognize only units that occur in the novel: currencies (pence, sixpence, shilling and pound), duration (minute, hour, day, week, month and year), lengths (mile, foot and yard), speed (mile an hour), temperature (degree) and weight (ton).

Sixpence is the name of an old coin. We will check its presence in the Unitex distribution dictionary. We will open the *Dela/Lookup...* menu, select the *dela-en-public.bin* dictionary and type *sixpence* in the *Word* field. Unitex displays:

sixpence,.N:s:p

Unitex considers *sixpence* as both singular and plural, but Charles Dickens uses a plural form with an *s*. Sure enough, if we use the *Dela/Lookup...* menu again for the form *sixpences*, we will not find it.⁸⁴

Therefore, in the graph of <u>Section 3.1.3</u>, page 90, we cannot use the *<sixpence>* mask, but *sixpence+sixpences*.⁸⁵

3.1 Analysis cascade

Most probably, we will be able to use the *Text/OpenRecent* menu and open directly *DombeyAndSon.snt*. If this file is not in the offered list, we will use the *Text/Open* menu, find the *DombeyAndSon* folder, and choose the *All Files* option in the *Files of Type* drop-down menu. If *DombeyAndSon.snt* exists in it, we will open it because it means that it has already been preprocessed. Otherwise, we will open *DombeyAndSon.txt* and answer *Yes* to the question *Do you want to preprocess the text*. A window will open in which we will remove the first two checkmarks, leaving just the last one *Apply All default Dictionaries*.

3.1.1 The number.grf graph

We will copy the NB2-999999.grf graph and all the subgraphs (*-subgraph.grf) from the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Numbers*

and paste them in the folder:

Then we will rename *NB2-999999.grf* as *number.grf*. We should remember that this graph recognizes all numbers between 2 and 999 999.⁸⁶

We must make a final modification to the *number.grf* graph.

- 1. We replace the call to the *NB2-999* graph by a call to the two subgraphs, *NB1-99-subgraph* and *NB100-999-subgraph* (these two subgraphs produce no output and the first one recognizes number one).
- 2. We modify the two paths leading to and from this box (because the subgraphs don't produce output).
- 3. We remove the box :NB100-999-subgraph after the box thousand and replace the subgraph NB1-99-subgraph by the subgraph NB100-999-subgraph.
- 4. Since we are going to use it in a cascade, we will change its output to produce lexical tags belonging to the category *Number*: the opening brace will replace the opening

⁸⁴ The Merriam-Webster Dictionary records that this noun has two forms of plural: *sixpence* and *sixpences* (<u>https://www.merriam-webster.com/dictionary/sixpence</u>). In the Unitex distribution dictionary only the first possibility is recorded.

⁸⁵ The *<sixpence>* mask only matches the *sixpence* form present in the dictionary.

⁸⁶ See <u>Chapter 3, Section 2.4.2</u>, page 62.

XML tag *<number>* and the rest of the lexical tag including the closing brace will replace the closing XML tag *</number>*.



This graph recognizes all numbers between 1 and 999 999. We will compile it.

3.1.2 Cascade

We will now produce a new analysis cascade that will initially contain only one graph, *number.grf*. Remember that in order to add it to the cascade this graph must be compiled. We will save the cascade in the folder

$MyUnitex \ English \ CasSys \ UnitexGettingStarted \ Measures$

and name it analysis.csc. The graph will work in the Merge mode.

When we launch this cascade, we will obtain 1 652 lines: the same 736 matches that we obtained with the *NB2-999999.grf* graph (the only difference is the output, the matches are in lexical tags now) and 916 matches for the ambiguous word *one*, which we previously omitted.

3.1.3 The measure.grf graph

We will now build a graph that recognizes and classifies measurement entities, that is numbers followed by measurement units. As explained before, it will recognize only those units that occur in the novel. The graph classifies the recognized entities as *Measure* and sub-classifies them as *Currencies, Durations, Lengths, Speed, Temperature* and *Weight*. Note that we have added the Part-Of-Speech code N in the lexical masks for the units when corresponding lemmas are ambiguous with verbs or adjectives.



We will now compile this graph and add it as the second graph in the analysis cascade. This graph will also be applied in the *Merge* mode.



After launching this cascade, the result will be again a list of 1 652 concordance lines, some of which contain measurement entities and the other just numbers or the word *one* (tagged as a number as well).

Concordance: D:\Documents\WvL	JnitexEnglishiCorpus\DombevAndSon\DombevAndSon_sntconcord.html
1652 matches	••••••••••••••••••••••••••••••••••••••
d one-not less than	{\/ia hundred thousand\.Number\+grf=number\} pound,.Measure+Currencie+grf=measure}. Mr Perch don't think 🔺
think himself that	<pre>//(a hundred thousand\.Number\+grf=number\} pound,.Measure+Currencie+grf=measure) will nearly cover it.</pre>
e would freely give	<pre>//(a hundred thousand\.Number\+grf=number\) pounds,.Measure+Currencie+grf=measure} (if he had it) for We</pre>
e's a Tartar within	<pre>{\{a hundred\.Number\+grf=number\} miles,.Measure+Length+grf=measure} of where we're now in conversation</pre>
y friend Gills many	<pre>{\{a hundred\.Number\+grf=number\} pound,.Measure+Currencie+grf=measure} already. More particulars in pi</pre>
at the distance of	<pre>{\{a hundred\.Number\+grf=number\} yards,.Measure+Length+grf=measure} or so, and watched it, during all</pre>
if he were to live	<pre>{\{a hundred\.Number\+grf=number\} years,.Measure+Duration+grf=measure}, he never could forget it. Somev</pre>
as unconscious, for	<pre>{\{a thousand\.Number\+grf=number\} pounds,.Measure+Currencie+grf=measure}. "My Heart's Delight!" said</pre>
oxing Mr Dombey for	<u>{\{a thousand\.Number\+grf=number\} pounds,.Measure+Currencie+grf=measure}</u> a side and the championship
if she lives to be	<pre>{\{a thousand\.Number\+grf=number\} years,.Measure+Duration+grf=measure} old." Miss Tox elevated her eye</pre>
eir ride was six or	<pre>(\{eight\.Number\+grf=number\) miles,.Measure+Length+grf=measure} long. When they drove through certain</pre>
e were worth a good	<pre>{\{eighty\.Number\+grf=number\} pounds,.Measure+Currencie+grf=measure} a-year to me-and I can't afford t</pre>
e glass again, "for	<pre>{\{fifty thousand\.Number\+grf=number\} pounds,.Measure+Currencie+grf=measure}!" The Major couldn't for</pre>
from twenty-five to	<pre>{\{fifty\.Number\+grf=number\} miles an hour,.Measure+Speed+grf=measure}, or he was sleeping after his 1</pre>
om ten shillings to	{\{fifty\.Number\+grf=number\} poundsMeasure+Currencie+grf=measure}, and exposed them in the window to
n, and make him buy	<pre>{\ffifty\.Number\+grf=number\) pounds,.Measure+Currencie+grf=measure1 worth of instruments for ready me</pre>
in, "any time these	<u>{\{fifty\.Number\+grf=number\} yearMeasure+Duration+grf=measure}</u> . When you see Ned Cuttle bite his nat
in a barouche, some	<u>{\{fifty\.Number\+grf=number\} years,.Measure+Duration+grf=measure}</u> before, by a then fashionable artist
lish brig, of about	<u>{\{five hundred\.Number\+grf=number\} tons,.Measure+Weight+grf=measure}</u> burden, together with a portion
the house cost full	<pre>(\{five thousand\.Number\+grf=number\} pound,.Measure+Currencie+grf=measure} if they cost a penny; and N</pre>
settle in one place	<pre>(\five\.Number\+grf=number\) minutes,.Measure+Duration+grf=measure). He goes about, like a-just like a</pre>
ided in again. Some	<pre>(\five\.Number\+grf=number\) minutes,.Measure+Duration+grf=measure} elapsed before Captain Cuttle could </pre>
I had not non-	

3.2 Synthesis cascade

Our next step is to produce the cascade that will transform lexical masks into XML tags. We will open the text *DombeyAndSon_csc.txt* from the folder:

without preprocessing.

3.2.1 The tag.grf graph

We will create a *tag.grf* graph that will be similar to the graph with the same name that we produced for the named entity recognition problem. So, we can copy this graph from the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Entities**Synthesis*

and paste it in the folder:

We will delete the upper part of the previous *tag.grf* (the part used to recognize *Hidden* and *Xml* codes) and modify the rest. We will replace the old names of the tags (*Measure* and *Number* will replace *AbsoluteDate*, *RelativeDate*, *Town* and *Street*) and add the recognition of subclasses that will be values of the attribute *type*. Note that subclasses are optional (class *Measure* have them, while class *Number* does not) and that these codes are also inside tags *<code>* and *</code>* produced by the *analysis.csc* cascade. Two paths that lead to the ending node after the recognition of codes for classes (*Measure* and *Number*) need to have weights, otherwise the recognition may just randomly skip everything up to the closing tag of a matched *</csc>*.⁸⁷ That is why the path for the recognition of subclasses has the higher priority *\$2\$*.



(3) <Measure type="Length"><Number>a hundred</Number> miles</Measure>

3.2.2 Cascade

We will now create a new cascade containing only the *tag.grf* graph that we have just built and compiled. This graph will work in the *Replace* mode and it will iterate until it comes to a fixed point (no change made). Therefore, we must check these two options. We will save this cascade in the folder:

*MyUnitex**English**CasSys**UnitexGettingStarted**Measures*

⁸⁷ See the <u>Chapter 3, Section 3.5.2</u>, page 69.

giving it the name synthesis.csc.

After launching this cascade, we will get the resulting *DombeyAndSon_csc_csc.txt* file in the folder:

Here is an excerpt from this file in which we see that numbers are tagged both by themselves and inside *<Measure></Measure>* tags. In addition, *Measure* tags have the attribute *type* with a value indicating the type of measurement unit.

We will take the following sentence as an example:

Dombey was about <csc><form>eight</form><code>Number</code>code>grf=number</code> </csc>-and-<csc><form><csc><form>forty</form><code>Number</code> <code>grf=number</code></csc> years</form><code>Measure</code> <code>Duration</code><code>grf=measure</code></csc> of age.

After applying the cascade, this sentence becomes:

Dombey was about <Number>eight</Number>-and-<Measure type="Duration"><Number>forty </Number> years</Measure> of age.

3.2.3 The numberDeletion.grf graph

We would like to annotate only measurement entities and not numbers unless they are part of them. We will thus create a *numberDeletion.grf* graph that removes *Number* tags found outside *Measure* tags.



We will compile this graph and add it as the second in the *synthesis.csc* cascade to be used in the *Replace* mode.

This cascade applied to the text *DombeyAndSon_csc.txt* gives us the desired result. Now, the annotation of the first number is deleted.

Dombey was about eight-and-<Measure type="Duration"><Number>forty</Number> years </Measure> of age.

However, we see that a *number-and-number* sequence actually represents one number, while only the second number has been included in the Measure entity. We need to correct this.

3.2.4 The numberAndNumber.grf graph

We will just add a new graph, *numberAndNumber.grf*, between the two previous graphs of the *synthesis* cascade, to include the whole sequence in *Measure* tags. We have to take care to recognize the type of the *Measure* entity.



The previous example is now correctly annotated. The text obtained as a result of the *tag.fst2* graph:

Dombey was about <Number>eight</Number>-and-<Measure type="Date"><Number>forty</Number> years</Measure> of age.

becomes now:

Dombey was about <Measure type="Duration"><Number>eight-and-forty</Number> years </Measure> of age.

4 Additional possibilities (only for confident users)

4.1 How does a generalization graph work?

The generalization graph developed in <u>Section 1.1.5</u>, page 75, enables the automatic creation of a graph which will replace it in the cascade and which is constructed on the basis of occurrences found in a text. This graph, having the same name *townGeneralization.grf*, is in the folder (this graph is the fourth graph in the *analysis* cascade):

*MyUnitex**English**Corpus**UnitexGettingStarted*\ *JubileePlainText**jubileePlainText_csc**jubileePlainText_4_0_snt*

After rearrangement, this graph looks like this:



As compared to the generalization graph, the automatically generated graph contains an additional box with the list of all words that were previously tagged as *Town* (*London* and *Nearlondon*). The G box is replaced by a negative right context in order to avoid tagging again as *Town* words that were already tagged *Town*.

4.2 Testing the variables

In the folder:

we can observe in the *numberDeletion.grf* graph, <u>Section 3.2.3</u>, page 93, that some part is written twice (recognition of the element tags and the content). In order to avoid this, we could create a subgraph. However, we will instead explore another solution, the possibility to test variables.

We will create a new graph named *numberDeletionWithTest.grf* from the previous *numberDeletion.grf* graph. In this graph we recognize the element only once, but we decide what output will be produced based on the variable *\$tag\$* (value of the open tag of a measurement entity with its attribute). In this graph the *\$tag.SET\$* box is successful only if the *\$tag\$* variable is not empty, that is, the *Number* tag is inside a *Measure* tag. On the other hand, the path through the *\$tag.UNSET\$* box is successful only if the *\$tag\$* variable is empty, that is, the *Number* tag.



If we compile this graph and replace the second graph in the *synthesis.csc* cascade with it, the cascade will produce the same result as before when applied to the text *DombeyAndSon_csc.txt*. To do the substitution, we can use the *Disabled* column that enables switching on and off some graphs in a cascade.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		tag.fst2		V	2	
2		numberAndNumber.fst2		V		
	1	numberDeletion.fst2		V		
3		numberDeletionWithTest.fst2		V		

4.3 Reuse the same graph

Redears can recall that we have constructed the *number.grf* graph from the *NB2-999999.grf* graph (see <u>Section 3.1.1</u>, page 89). We will now propose another solution, a single graph that can be used to annotate both with XML tags and with lexical tags. We will copy the *number.grf* graph and paste it in the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Numbers*

We will name it *numberXmlOrLt.grf* and we will remove the insertion of an opening tag <*E*>/{ and replace it with an empty box (<*E*>); and similarly, we will remove the insertion of a closing tag <*E*>/,.Number+grf=number} and replace it by an empty box (<*E*>).



We will now describe how to use this graph for both types of annotation: We will use output variables.

4.3.1 Output variable

We will create a new graph, *numberTest.grf*, and place it in the same folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Numbers*

In this graph, the call to the *numberXmlOrLt.grf* graph follows a test on an output variable *\$test\$*.



A new graph named *numberXml.grf* will replace the *NB2-999999.grf* graph created in <u>Chapter 3, Section 2.4.2</u>, page 62. In this graph, the *<E>/XML* empty box is surrounded by special blue output parentheses, which you can insert by clicking the *Surround box selection with an output variable* button. The value of the output variable *\$test\$* is the output of the path enclosed in the blue parentheses, here *XML*.⁸⁸ To tag only numbers greater than 1, as we saw in <u>Chapter 3, Section 3.4.2</u>, page 67, we will use weights.



If we apply this graph to the novel text, we will obtain 1658 concordance lines, but only 736 lines with <Number></Number> tags, as in the precedent chapter.

⁸⁸ See Unitex user manual, Section 6.8.

🛅 Concordance: D:\Documents\MyUnitex\English\Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 📃 🖬 🛛
1658 matches
ses may draw out my <u><number>two</number></u> front double teeth, Mrs Richards, k
d be strange if any <u><number>two</number></u> persons could be together for so ma
out of window every < <u>Number>two</u> or three minutes, in impatient remo
me into the shop." " <u><number>Two</number></u> , Uncle, don't you recollect? There
observed the man. " <u><number>Two</number></u> gentlemen here, Sir, but they're wa
nt his children-one- <u><number>two</number></u> -stop, and go back, and begin again 📃
nning, cunning man! <u>One</u> shouldn't tell these things; your sex, my dear Dombe
so happy now, dear! <u>One</u> more good-bye, my precious, one more!" How Susan doe
your good husband!" One time when Walter was in his room, he beckoned him to
"for Mrs Richards," one of Susan's fellow-servants ominously whispered, as s
1, Sir, thank you." One evening Mr Toots was sitting at his desk, oppressed
ook at Paul in bed, one night when she took up Mrs Wickam's supper. "Ah!" si $_{ar{\bullet}}$

In the same way, a new graph named *numberLt.grf* will replace the *number.grf* graph, created in <u>Section 3.1.1</u>, page 89.



4.3.2 Graph repository

This solution can be improved by using a graph repository. For this, let us create a new folder, *Repository*, inside the folder:⁸⁹

*MyUnitex**English**Graphs**UnitexGettingStarted*

In order to make this folder our *graph repository*, we have to open the *Info/Preferences/Directories* menu, click the *Set* button next to the *Graph repository* option in order to choose this specific folder, and finish the action by clicking the *OK* button.⁹⁰

Preference	ces for English					X
Morphologie	al-mode diction	aries	SVN	Enco	ding	
General	Directories	Lang	uage	Pres	entatio	n
Private Unite	x directory (whe	re all us	ser's data	a is to t	be stor	ed):
U:\Documer	nts\MyUnitex				Set.	
Graph repository: English/Graphs/UnitexCettingStarted/Repository Set Produce log information in directory.						
C:\Users\Pu	iblic\Desktop				Set.	
Clear all logs						
			OK		Cance	el

We will copy the *numberXmlOrLt.grf* graph and its three subgraphs⁹¹ from the folder:

 $MyUnitex \ English \ Graphs \ UnitexGettingStarted \ Numbers$

and paste them in the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Repository*

⁸⁹ This folder is not necessarily located in the *MyUnitex* folder. It can be common to several projects, each corresponding to a particular private working directory. See <u>Chapter 2, Section 1.1.4</u>, page 11.

⁹⁰ This action adds a line to the configuration file located in the folder: *MyUnitex**English*

⁹¹ NB1-9-subgraph.grf, NB1-99-subgraph.grf and NB100-999-subgraph.grf.

A call to a subgraph that is in the repository is obtained by preceding its name with two colons; the box will have yellow background. It doesn't matter in which folder the main graph is, since it calls a subgraph that is in the repository.

We will create two new graphs in the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Numbers*

The first graph, named *numberXmlRepository.grf*, annotates numbers with <*Number></Number>* tags, while the second graph named *numberLtRepository.grf* transforms numbers in *lexical tags*.



4.4 Hiding or deleting part of a text

Before starting this section

Go to the folder:MyUnitex\English\Corpus\UnitexGettingStartedCreate a new folder named (without space): JubileeLoremIpsum.Go to the folder:MyUnitex\English\Corpus\UnitexGettingStarted\JubileeLoremIpsumDownload the JubileeLoremIpsum.xml file.

4.4.1 Hiding part of a text

As mentioned in <u>Section 2.1.2</u>, page 86, the *toolHidden.grf* graph causes an error if the text to be hidden is too long. Actually, a Unitex graph analyzes a text by traversing all the possible paths (between the initial box and the final box) before deciding which path to choose. However, the maximum length of this run is set at 1,000 tokens. This value can be increased by using a script (see <u>Chapter 7</u>, page 149), but there is still a limit.

We will use a text that we prepared specifically for this exercise, *JubileeLoremIpsum.xml*. You will find it in the folder:

This file contains 596 sequences of letters and 750 other tokens between tags <title></title>.

As we explained before, we will rename this file as *JubileeLoremIpsum.xml.txt*, open it in Unitex and answer *Yes* to the question *Do you want to preprocess the text*. A window will open in which we will remove the first two checkmarks, leaving just the last one *Apply All default Dictionaries*.

Now we will open the *Text/Apply CasSys Cascade...* menu and select the *analysis.csc* file in the folder:

When we click the *Launch* button, we obtain an error message.

Done						
Messages with a colored backgroun	lessages with a colored background are generated by the interface, not by the external programs.					
Construct cascade concord						
Concord File U:\Documents\MyUnite	ex\English\Corpus\JubileeLoremIpsu	mUubileeLoremIpsum.xml_snt\d				
Building xml concord						
Concord "U:\Documents\MyUnitex\E	nglish\Corpus\JubileeLoremIpsum\J	ubileeLoremlpsum.xml_snt\con				
Merging outputs with text						
Done.						
Done.						
Construct cascade concord						
Concord File U:\Documents\MyUnite	ex\English\Corpus\JubileeLoremIpsu	mUubileeLoremIpsum.xml_snt\c				
Concord "U:\Documents\MyUnitex\E	nglish\Corpus\JubileeLoremIpsum\J	lubileeLoremlpsum.xml_snt\con				
Merging outputs with text						
Done.						
Done.		-				
Maximal stack size reached!						
(There may be longer matches not recognized!) with grammar U:\Documents\MyUnitex\English\Graphs\Unite						
22{,.Xml+name=date+closeTag+grf=toolXml} < <here>>{<title>,.Xml+name=title+openTag+grf=tool</title></here>						
ОК	Cancel	Copy error message				

We have to modify the *toolHidden.grf* graph that caused the error. We will open it (*Graphs/Open* menu) in the folder:

MyUnitex\English\Graphs\UnitexGettingStarted\Entities\Analysis

and save it under the new name toolHiddenDateTag.grf (we will use the FSGraph/Save as... menu). We should not forget to modify the tracing feature as well: *..Hidden+grf=toolHidden* should be replaced by *..Hidden+grf=toolHiddenDateTag*. We will delete the first path, leaving only the path that hides the content between the *<date></date>* tags. We will compile this graph.



We will create another new graph, named *toolHiddenTitleTag.grf*, to hide the content between the *<title></title>* tags.



The notation \$[1,300]\$ above the box that calls this subgraph controls the number of iterations (from one to 300 tokens).⁹² The *patternNotHiddenNotTitleCloseTag.grf* subgraph uses the negative right context to recognize a token only if it is neither a hidden text (token classified as "Hidden"), nor the *title* closing tag.⁹³



We will modify the analysis cascade in which we will replace the previous *toolHidden.grf* graph with the two graphs: the *toolHiddenDateTag.grf* graph followed by the *toolHiddenTitleTag* graph. The latter graph will be applied until the fixed point is reached.

⁹² Like all output, these interval boundaries are written after a slash character. The box actually contains :patternNotHiddenNotTitleCloseTag/\$[1,300]\$.

⁹³ See <u>Chapter 3, Section 1.7.3</u>, page 55, for more details about negative right context.

Unitex Getting Started - Denis Maurel and Cvetana Krstev

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		toolXml.fst2	~			
	~	toolHidden.fst2	~			
2		toolHiddenDateTag.fst2	~			
3		toolHiddenTitleTag.fst2	~		~	
4		absoluteDate.fst2	~			
5		relativeDate.fst2	~			
6		town.fst2	~			
7		townGeneralization.fst2	2			2
3		street.fst2	~			

We will explain in detail how the *toolHiddenTitleTag.grf* graph works: Unitex detects the opening *title* tag, then it iterates and recognizes 300 tokens at most. If it does not come upon the closing *title* tag, as will be the case with our example text, it tags these 300 tokens as a text to be hidden (*Hidden*) with an additional feature (*Hidden=insideTag*) stating that the end of a text was not yet reached. The graph is launched again. In the second run, the graph detects the new token that we created in the previous step, followed by some regular tokens and after that it loops again over at most 300 tokens.⁹⁴ When the closing *title* tag is recognized before reaching the upper limit of iterations 300, the text collected as a title is categorized as *Hidden* with the additional atribute *Hidden=endTag*.

As we saw in <u>Section 1.1.7.2</u>, page 78, the cascade created a new folder:

MyUnitex\English\Corpus\UnitexGettingStarted\ JubileeLoremIpsum\JubileeLoremIpsum.xml_csc

containing a new file for each graph or iteration of graph.

The *toolHiddenTitleTag.grf* graph is the third graph of the cascade and it is iterated four times. So the cascade created four new files named:

- 1. JubileeLoremIpsum.xml_3_0.snt,
- 2. JubileeLoremIpsum.xml_3_1.snt,
- 3. JubileeLoremIpsum.xml_3_2.snt,
- 4. JubileeLoremIpsum.xml_3_3.snt.

To check the result, we will open these files. If we search with *name=title* to gain insight only into interesting categories, we will find:

JubileeLoremIpsum	{name=title+openTag}{Hidden=insideTag} Aliquam convallis
.xml_3_0.snt	sollicitudin purus{name=title+closeTag}
JubileeLoremIpsum	{name=title+openTag}{Hidden=insideTag}{Hidden=insideTag} augue.
.xml_3_1.snt	Integer id felis
JubileeLoremIpsum	{name=title+openTag}{Hidden=insideTag}{Hidden=insideTag}{Hidden
.xml_3_2.snt	=endTag}{name=title+closeTag} ⁹⁵

The JubileeLoremIpsum.xml_3_2.snt and JubileeLoremIpsum.xml_3_3.snt files are identical, which means that the fixed point has been reached and that the iteration of the toolHiddenTitleTag graph stops.

The synthesis cascade (the same as before, see <u>Section 2.2</u>, page 87) will rewrite this text without any internal annotation, whereas, if the hiding graphs had not been used, *Wednesday*

⁹⁴ Actually, the 300 tokens surrounded with braces have become a single token of a category *Hidden*.

⁹⁵ Which means that at this point our text, inside *<title>/</title>* tags, consists of only 5 tokens.

June 1, 2022 would be tagged as an AbsoluteDate inside the <date></date> tags and Nearlondon would be tagged as a Town inside the <title></title> tags.⁹⁶

<xml></xml>
<pre><date>Posted on Wednesday June 1, 2022</date></pre>
<title>From Nearlondon:Lorem ipsum</title>

4.4.2 Deleting part of a text

4.4.2.1 The deleting graphs

Instead of hiding the *date* and *title* tags, suppose we want to remove their content.

For the *date* tag, presumably with a short content, we can write a graph that replaces everything recognized with <date></date>. We will name this graph *toolEraseDateTag.grf*.



For the *title* tag, which can have a very long content in the number of tokens, we will use a loop limited to 300 tokens. We will name this graph *toolEraseTitleTag.grf*.



This graph uses a subgraph that we will name patternNotCloseTitle.grf.



These two graphs are similar to graphs developed in the previous section; they differ only in the produced output. We will build now the cascade. The *toolEraseDateTag* graph and the *toolEraseTitleTag* graph are in *Replace* mode; the *toolEraseTitleTag* graph is repeatedly used until a fixed point.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		toolXml.fst2	2			
	1	toolHidden.fst2	2			
	V	toolHiddenDateTag.fst2	1			
2		toolEraseDateTag.fst2		2		
	V	toolHiddenTitleTag.fst2	1		V	
3		toolEraseTitleTag.fst2		2	V	
4		absoluteDate.fst2	1			
5		relativeDate.fst2	1			
6		town.fst2	1			
7		townGeneralization.fst2	1			r
8		street fst2	V			

We will open the *JubileeLoremIpsum.xml.snt* text and launch the analysis cascade; we will open the *JubileeLoremIpsum.xml_csc.txt* text (without applying dictionaries) and launch the

⁹⁶ Now that we no longer have to care about the length of the text, we can directly hide the whole preamble, between the *<xml>* and *<bdy>* tags, instead of hiding inside the *<date>* and *<title>* tags separately.

synthesis cascade. The beginning of the resulting text *JubileeLoremIpsum.xml_csc_csc.txt* is now:

<xml></xml>		
<date></date>		
<title></title>		
<body></body>		

4.4.2.2 Comparing the variables

To stick to common XML practice, we will add to the synthesis cascade a new graph that transforms two successive tags, opening and closing, into a self-closing tag. We will name this graph *selfClosing.grf*. We will use two variables *\$tag1\$* and *\$tag2\$* and compare them with the formula *\$tag1.EQUAL=tag2\$*.⁹⁷ A self-closing tag is output only if two successive opening and closing tag contain the same element name.



The new graph is added as the third in the synthesis cascade and marked to be used in *Replace* mode.



After applying the synthesis graph to *JubileeLoremIpsum.xml_csc.snt*, we will get what we wanted.

<xml> <date></date> <title></title></xml>	
<body></body>	

4.5 Sub-categorization with cascades

We will work here on the *JubileePlainText* text.

4.5.1 Three modified graphs

We will take the first two graphs of <u>Section 1.1</u>, page 71, named *absoluteDate.grf* and *relativeDate.grf*; we will modify them by introducing sub-categories. Either DayName, DayNumer (as in graphs) or name of a day, number of a day.⁹⁸

⁹⁷ It is also possible to compare a variable to a character string either in a case-sensitive or case-insensitive way. See the Unitex user manual, Section 6.9.2.

⁹⁸ Care must be taken that, on each path, a closing brace corresponds to an opening brace. In a complex graph, we can verify this by using the *FSGraph/Tools/Verify braces* option.



Then, we must also modify the *tag* graph of the synthesis cascade, adding the new categories.⁹⁹



<csc><form>June 2, 2022</form><code>AbsoluteDate</code>code>grf=absoluteDate</code></csc>

When we run the analysis and synthesis cascades with these three new graphs (don't forget to compile the modified graphs!), we get the desired result, a subcategorization of dates.

⁹⁹ When the categories have been added in the corresponding box, it is possible to sort them in alphabetical order by the *FSGraph/Tools/Sort Node Label* option.
This week we celebrate the Platinum Jubilee of the Queen of England: seventy years of reign! The party will start on <AbsoluteDate><Month>June</Month> <DayNumber>2</DayNumber>, <Year>2022</Year></AbsoluteDate> and end on <RelativeDate><Month>June</Month> <DayNumber>5</DayNumber></RelativeDate>. The Queen ascended the British throne on <AbsoluteDate><Month>February</Month> <DayNumber>6</DayNumber>, <Year>1952</Year></AbsoluteDate>. Our small town of <Town>Nearlondon</Town> will also celebrate this event. <Town>Nearlondon</Town> has always been a loyal subject of Her Majesty. Although the city of <Town>London</Town> is dear to us, on <RelativeDate><DayName>Thursday</DayName> <Month>June</Month><DayNumber>2</DayNumber></RelativeDate> we will change the name of <Address>London Street</Address> to <Address>Queen Elizabeth Street</Address>. Everyone will fondly remember 2022! God Save the Queen!

4.5.2 Generalization graphs with restrictions

4.5.2.1 Without substitution of a category

A single number should not be recognized as a year, even if it has 4 digits. This could cause too many errors. This is why 2022 in the penultimate sentence *Everyone will fondly remember 2022* was not tagged as a *Year* (nor as a date). We will use a generalization graph to decide that since 2022 has already been recognized as a date, it has to be recognized again as a date.¹⁰⁰ We will name this graph *yearGeneralization.grf.* Note that the subcategory *Year* is in the box, while category *AbsoluteDate* is the output of the box.



We will place this graph in the third position of the *analysis* cascade. Don't forget to check the *Generaliz*... box.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		absoluteDate.fst2	2			
2		relativeDate.fst2	2			
3		yearGeneralization.fst2	2			~
4		town.fst2	2			
5		townGeneralization.fst2	2			V
6		street fst2	V			

When we run this new analysis cascade and the same synthesis cascade, the desired result is as expected (We display only the penultimate line).

Everyone will fondly remember <AbsoluteDate><Year>2022</Year></AbsoluteDate>!

As explained in <u>Section 4.1</u>, page 94, the generalization graph is automatically replaced in the cascade by a generated graph, having the same name *yearGeneralization.grf*, which is in the folder:

*MyUnitex**English**Corpus**UnitexGettingStarted*\ *JubileePlainText**jubileePlainText_csc**jubileePlainText_3_0_snt*

After rearrangement, this graph looks like this:



¹⁰⁰ See <u>Section 1.1.5</u>, page 75. Remember that the .grf file must stay in the same folder as the .fst2 file.

4.5.2.2 With substitution of a category

It is also possible to modify the category obtained. For example, if we will want to categorize *YearDate* years recognized in this way, we must write in the box *Year.YearDate/,.AbsoluteDate*. Save this graph.

→ G - <u>YearDate</u> → □ ,.AbsoluteDate +grf=yearGeneralization}

Of course, this will only work if the *tag.grf* graph of the synthesis cascade is also modified to recognize the new tag *YearDate*.



<csc><form>June 2, 2022</form><code>AbsoluteDate</code><code>grf=absoluteDate</code></csc>

The result is now different (don't forget to compile the modified graphs!).

Everyone will fondly remember <YearDate><Year>2022</Year></YearDate>!

After rearrangement, the replaced generalization graph looks like this:



Chapter 5: dictionary creation

Before starting this chapter

Go to the folder:MyUnitex\English\InflectionSelect and copy the two files named Equivalences.txt and Morphology.txt.Create a new folder named UnitexGettingStarted (without space).Go to the folder:MyUnitex\English\Inflection\UnitexGettingStartedPaste the two files Equivalences.txt and Morphology.txt.¹⁰¹

1 Introduction

We have already discussed the use of the dictionaries from the Unitex distribution in <u>Chapter 2, Section 1.2.4</u>, page 14.

Before we begin, we will open in Unitex our novel, *DombeyAndSon.txt*, located in the folder:

We will answer No to the question Do you want to preprocess the text.

In this chapter, we will create a new dictionary following these three steps:

- 1. We will manually create a dictionary of representative (principal) forms (or lemmas) and rules for their inflection;
- 2. We will create automatically a dictionary of inflected forms, that is, all forms that can occur in texts, like singular and plural forms of nouns;
- 3. We will automatically create a dictionary in format .*bin* that will appear in the *User resources* field.

For the first step, we will use the Unitex editor. We will open the *File Edition/New File* menu and save the file in the *Dela* folder using the name *myDictionary.dic*. We will make a difference between the inflection of monolexical words and polylexical words.¹⁰²

2 Inflection of monolexical words

2.1 Inflection by simple suffixation

2.1.1 The nouns with s in the plural

We will enter the following three lines in the *myDictionary.dic* file. Note that these entries should not contain spaces. In addition, we will finish the third line (the last line) by clicking the *Enter* key.

¹⁰¹ These two files define the morphological codes used by Unitex for a specific language. It is not recommended to modify them, except if a user wants to use different morphological codes or wants to create resources for a new language.

¹⁰² It is also possible to create the inflection rules based on word stems, which is convenient for the Semitic languages; see the Unitex user manual, Section 3.5.4.

advantage,N1 agent,N1 firm,N1					
-------------------------------------	--	--	--	--	--

Each line represents a dictionary entry consisting of a lemma (representative form), followed after a comma by the associated inflectional graph (in the case of our entries *N1.grf*). In the name of this graph, *N* represents the grammatical category (Part-Of-Speech) noun. We will now create this graph and save it in the folder:

We will save and close the myDictionary.dic file.



The *N1* graph enables the inflection of nouns for which the plural is obtained by adding an *s* to the singular form. We will use, although it is not mandatory, the same codes used by the dictionary provided with the Unitex distribution *dela-en-public.bin*. One can see these codes in the *Word List* window: *s* is used for the singular forms and *p* for the plural forms. We will add a comment (a box starting with the slash /) with some examples of nouns that can be inflected by this graph.

This inflectional graph is read like this: in the upper part, there is an empty box, signifying that a word (lemma) to which it applies does not change; in the lower part, there is a box containing an *s*, signifying that it is concatenated to the word to which it applies. The output of these boxes are values of grammatical categories associated with inflected forms.



This graph must be compiled, so we will click the *Compile graph* button. We will obtain an error message: *Main graph matches epsilon! Error: the main graph NO recognize <E>*. The meaning of this message is that we cannot use this graph for search (the *Locate pattern* option) because it recognizes an empty symbol.¹⁰³ Since we have not developed this graph for search, we can ignore this message. We will simply click the *OK* button.

We will now launch the inflection program. This program will create a dictionary of all inflected forms, named *myDictionaryflx.dic* (*flx* is automatically added to the file name). In order to do this, we will open the *myDictionary.dic* file in the *DELA* menu, which gives access to programs

¹⁰³ Which would cause an infinite loop and crash the computer.

dealing with dictionaries. Note that now we will not use the *File Edition* menu, because we are not going to create or modify a dictionary of lemmas.¹⁰⁴

We will now inflect the *myDictionary.dic* dictionary by using the *DELA/Inflect* menu option. First, we will set the folder in which the inflectional graphs are stored. Click the *SET* button, select the folder:¹⁰⁵

MyUnitex\English\Inflexion\UnitexGettingStarted

and click the Open button. Next, click the Inflect Dictionary button.

Inflection				
Directory where inflectional FST2 are stored:				
nitex\English\Inflection\UnitexGettingStarted Set				
Allow both simple and compound words				
 Allow only simple words 				
 Allow only compound words 				
Sort and factorize inflectional codes				
Cancel Inflect Dictionary				

The myDictionaryflx.dic dictionary is created and displayed on the screen.



If we want to use this dictionary on our novel (or any other text), one more step is necessary. We must choose the *DELA/Compress into FST* menu option and confirm by clicking *OK*. This will create two more files: *myDictionaryflx.bin* and *myDictionaryflx.inf*.¹⁰⁶ Note that if you want to use this dictionary on some other computer you need to copy both files.

We will now close both windows containing the *myDictionary.dic* and *myDictionaryflx.dic* dictionaries. We want to apply our newly created dictionary to the novel (already open). We will open the *Text/Apply Lexical Resources* menu and click the *Clear* button (in order to cancel all previous selections). We will choose the *myDictionaryflx.bin* file (appearing in the left field) and click the *Apply* button. As a result, the *Word List* window will be displayed.

¹⁰⁴ If you detect an error in your *myDictionary.dic* file, you have to close it in the *DELA* menu and reopen it in the *File Edition* menu.

¹⁰⁵ If you are not going to go through this entire section in one session, remember to set the folder containing the inflectional graphs when you start to work again.

¹⁰⁶ The *myDictionaryflx.bin* file contains words, but it is a binary file and a text editor cannot read it. The *myDictionaryflx.inf* file contains the factorization of tags associated with entries: it is readable.

Unitex Getting Started - Denis Maurel and Cvetana Krstev

Word Lists in U:\Documents\MyUnitex\En	glish\Corpus\DombeyAndSon 🗗 🔟 🛛
DLF: 5 simple-word lexical entries	ERR: 16427 unknown simple words
advantages, advantage.N:p agent, N:s firm, N:s firms, firm.N:p	A a a aback
	abaft abandon abandoned
	abandoning abandonment
	abandons abase abased
This file is empty.	abasement
	abate abated abatement
	abbreviations abducted
	abear abhorrence abide
	abilition V

Since five of our six inflected forms (two for each lemma) appear in this novel, they will be listed in the *DLF* area. *DLC* area is empty, while the *ERR* area lists 16 427 unknown simple words. This is natural since we applied only our tiny dictionary to the text.

We will close the Word List window.

2.1.2 Two more examples

We will open the *myDictionary.dic* file again, using the *File Edition/Open.../Dictionaries* menu and answer *OK* to the warning *This is not necessarily the text being processed by Unitex* stating that this file cannot be used for the text analysis.¹⁰⁷ We will add to our dictionary of lemmas some more illustrative examples:

- for the inflection of the nouns *branch*, *bus*, *box* having plural forms produced by adding *es* to the singular form we will use the *N2* graph;
- for the inflection of the verbs *accept*, *connect*, *work* having forms produced by adding *s*, *ed* and *ing* to the infinitive form we will use the *V1* graph.¹⁰⁸

We will use the same format as before – remember that you should not use spaces and to press the *Enter* key after the last entry.

advantage,N1 agent,N1 firm,N1 branch,N2 bus,N2 box,N2 accept,V1 connect,V1 work,V1

We will save and close the myDictionary.dic file.

¹⁰⁷ Except if we close it and reopen using the menu *Text/Open*.

¹⁰⁸ We saw in the <u>Chapter 2</u>, <u>Section 1.2.5</u>, page 15, that *G* codes the gerund of a verb. We saw in the <u>Chapter 3</u>, <u>Section 1.2</u>, page 47, that *K* codes the past participle and *W* codes the infinitive of a verb. We also need to know that the code *P* is used for the present tense, *I* for the perfect tense, while 1, 2 and 3 are codes for persons.

We will create and compile the N2 et V1 graphs. We can notice that the N2 graph is very similar to the previous N1 graph (the content of the lower box is *es* instead of *s*); so we can facilitate our work by opening N1.grf, using the FSGraphs/Save as... menu to rename it as N2.grf; and then changing the content of this box. Do not forget to replace the comment as well by typing new examples and to compile graphs.



We can now open the *myDictionary.dic* dictionary by using the *DELA/Open Recent* menu and inflect it using the *DELA/Inflect* option. The new dictionary named *myDictionaryflx.dic* will be displayed in a separate window with its 57 entries. You can see below the content of this window with some of the produced inflected forms.



We will verify the inflected forms and, if we find them correct, we can sort them by using the *DELA/Sort Dictionary* option.¹⁰⁹

¹⁰⁹ Note that we could have inflected this dictionary after saving the first graph, *N1.grf*. The inflection program issue error notifications for all entries using graphs that do not exist and produces a dictionary of inflected forms only for those entries for which inflectional graphs exist.

Unitex Getting Started - Denis Maurel and Cvetana Krstev

U:\Documents\MyUnitex\English\Dela\myDict	ionaryflx.dio	:	 0"	X
	Find	Reload	•	
accept, accept.V:Plp				•
accept, accept.V:Pls				
accept, accept.V:P2p				
accept, accept.V:P2s				
accept, accept.V:P3p				
accept,accept.V:W				
accepted, accept.V:Ilp				
accepted, accept.V:Ils				
accepted, accept.V:I2p				
accepted, accept.V:I2s				
accepted, accept.V:I3p				
accepted, accept.V:I3s				
accepted, accept.V:K				
accepting, accept.V:G				
accepts,accept.V:P3s				
advantage, advantage.N:s				
advantages, advantage.N:p				
agent, agent.N:s				
agents, agent.N:p				
box, box.N:s				
boxes, box.N:p				
branch, branch.N:s				
branches, branch.N:p				Ŧ
			•	

When inflecting a dictionary of lemmas, we can check the *Sort and factorize inflectional codes* option in the *DELA/Inflect* menu. A more compact dictionary of inflected forms will be produced. Namely, there will be only one entry for each form of a lemma, followed by a sequence of all possible values of grammatical codes. If we apply this to our *myDictionary.dic* dictionary, we will get the *myDictionaryflx.dic* dictionary of inflected forms consisting of 24 entries, some of which are displayed in the figure below.



In order to make this dictionary functional, we must apply the *DELA/Compress into FST* option. We will now close the windows containing *myDictionary.dic* and *myDictionaryflx.dic*. We will test this new dictionary by applying it to our novel (do not forget to click the *Clear* button in the *Text/ApplyLexical Resources*) menu. In the *DLF* zone of the *Word List* window, 19 entries will be listed (the word forms *accepts, bus* and *buses* do not occur in the novel) compared to the five entries obtained by the previous version of the dictionary. Therefore, there are less entries in the *ERR* zone: 16 411. Close the *Word List* window.

Chapter 5: Dictionary creation

Word Lists in U:\Documents\MyUnitex\En	glish\Corpus\DombeyAndSon 🖬 🗹
DLF: 19 simple-word lexical entries	ERR: 16411 unknown simple words
accept,.V:Plp:Pls:P2p:P2s:	Eilter unknown words with tags ind
accepted, accept.V:Ilp:Ils:	
accepting, accept.V:G	
advantage,.N:s	a
advantages, advantage.N:p	aback
agent, .N:s	abait
box,.N:s	abandon
boxes,box.N:p	abandoned
branch, .N:s	abandoning
branches, branch.N:p	abandonment
	abandons
	abase
DLC: compound lexical entries	abased
This file is empty.	abasement
	abashed
	abate
	abated
	abatement
	abbreviations
	abducted
	abear
	abhorrence
	abide
	abilition V

2.2 The *L* operator

In order to concatenate the suffix to a substring of a word, that is, not after its last letter (e.g. *compan* from *company*, see below), we use the *L* operator (standing for *Left*) in order to move from right to left starting from the end of a word.

2.2.1 The nouns with final y

Plural forms of the nouns *company, complexity, family* are produced by replacing the final y with the suffix *ies*. Starting from the end of the word, we move one letter to the left (operator L) and then we add *ies*. This rule will be written as *Lies* in the box for plural forms in the N3 graph.¹¹⁰

We will now add three more lines to the *myDictionary.dic* file (having now 12 entries). Remember to open it in the Unitex editor by using the *File Edition/Open.../Dictionaries* menu.



We will save this dictionary and close the editor. Now we will create the *N3* graph, using as a base the *N1.grf* graph, modifying it and saving with the *Save as...* menu option under the name *N3.grf*. Remember to replace the comment with new examples and compile the graph.



In order to use this new version of the dictionary we have to inflect and compress it. These are the steps to follow: open the *myDictionary.dic* file (the *DELA/Open Recent* menu) and inflect it (the *DELA/Inflect* menu, *Sort and factorize inflectional codes* option). The new version of the *myDictionaryflx.dic* dictionary will be displayed containing 30 lines. If inflected forms

¹¹⁰ Since the inflection endings are written in lower-case letters, the upper-case letters are reserved for operators.

are correct, we can proceed with the production of the dictionary version that can be used in text analysis: the *DELA/Compress into FST* menu. We will close the windows displaying *myDictionary.dic* and *myDictionaryflx.dic*. We can now apply the new dictionary to our text: the *Text/ApplyLexical Resources* menu (do not forget to click the *Clear* button). The *DLF* zone of the *Word List* window shows that 23 words have been recognized by the *myDictionaryflx.dic* dictionary. We can also see that the word *complexity* does not occur in our text. Before proceeding with new examples, we will close the *Word List* window.

2.2.2 Several more examples

We will again open the *myDictionary.dic* file in the Unitex editor (*File Edition/Open.../Dictionaries*). We will add to the dictionary some more examples of inflections requiring the *L* operator once or several times.

- *leaf, thief, calf* are nouns in which the final *f* is replaced by *ves* in plural (*N4.grf* graph).
- *wife, knife, life* are nouns in which the final *fe* is replaced by *ves* in plural (*N5.grf* graph, the *L* operator is applied twice).
- *mouse, louse* are nouns in which the final *ouse* is replaced by *ice* in plural (*N6.grf* graph, the *L* operator is applied four times).
- *narcissus, terminus, tumulus* are nouns in which the final *us* is replaced by *i* in plural (*N7.grf* graph, the *L* operator is applied twice).
- *phenomenon, criterion, datum* are nouns in which the two final letters are replaced by *a* in plural (*N8.grf* graph, the *L* operator is applied twice).
- *love*, *live*, *give* are verbs for which the final *e* is deleted before adding the ending *ing* (*V2.grf* graph, the *L* operator is applied once for this inflected form).
- *accompany, try, worry* are verbs for which the final letter *y* is replaced by *i* in all inflected forms (*V3.grf* graph, the *L* operator is applied once for all inflected forms).

We will now add entries for these nouns and verbs into the *myDictionary.dic* dictionary (do not forget to press the *Enter* key after the last entry). This dictionary now has 32 entries.

 leaf,N4	
 worry,V3	

We will save the dictionary and close the Unitex editor. After that, we will create and compile new graphs for the inflection of nouns and verbs.





As the last step, we will open the *myDictionary.dic* file using the *DELA/Open Recent* option and inflect it choosing *DELA/Inflect*. The new *myDictionaryflx.dic* dictionary contains 82 lines (if the Sort and factorize inflectional codes option is on). After compressing this dictionary (*DELA/Compress into FST* menu), we will close both windows (*myDictionary.dic* and *myDictionaryflx.dic*). We will open the *Text/ApplyLexical Resources* menu (do not forget to clear all preselected dictionaries by clicking the *Clear* button and select the *myDictionary.bin* dictionary). The *Word List* now displays 57 recognized words. We will close *Word List*.

2.3 Operators R, C and D

We use the *R* operator (*Right*) in combination with the *L* operator, when we need to substitute part of a lemma while preserving another part on the right.

What do the nouns *goose* and *tooth* have in common? In the plural, the last two letters remain as they are while the string *oo* is replaced by *ee*. It means that starting from the end of a lemma, we have to move to the left by four letters (*LLLL*), then write *ee*, and then move to the right by two letters (*RR*). This procedure is used by the *N9.grf* graph.



Let us see how it works. The vertical bar indicates the current position in the processed word (the position of the cursor) at each step in the sequence of operations.

LLLLeeRR	goose	tooth
L LLLeeRR	goos e	toot h
LL LLeeRR	goo se	too th
LLL LeeRR	go ose	to oth
LLLL eeRR	g oose	t ooth
LLLLe eRR	ge ose	te oth
LLLLee RR	gee se	tee th
LLLLeeR R	gees e	teet h
LLLLeeRR	geese	teeth

Other examples:

- *foot* is a noun for which the last letter remains in the plural while the string *oo* is replaced by *ee* (*N10.grf* graph).
- *man, woman, hypothesis* are nouns for which the penultimate letter becomes *e* in the plural (*N11.grf* graph).



The *C* operator (*Copy*) duplicates the current letter, that is the letter situated on the left of the cursor, which permits the inflection with a single graph of the verbs *jog*, *plan*, *stop*, *prefer* for which the final consonant is duplicated in the past tense, the past participle and the gerund (*V4.grf* graph).

Other examples:

- past tense is identical to the present tense for verbs *slit, split* while their final consonant is duplicated in the gerund (*V5.grf* graph).
- the final consonant is duplicated in the plural before adding *es* for the nouns *quiz* or *gas* (*N12.grf* graph).



The *D* operator (*Delete*) deletes the current letter (and shifts the right-hand part of the word to the left), that is the letter situated to the left of the cursor, which permits the inflection with a single graph of the verbs *shoot, speed, meet* for which the double letter (*oo, ee*) is simplified in the past tense and the past participle (*V6.grf* graph).¹¹¹ We will demonstrate step by step how the pattern *LDR* used in the *V6.grf* graph works.

¹¹¹ Note that some other operators exist that are described in the Unitex user manual, Section 3.5.1.

LDR	shoot	meet
L DR	shoo t	mee t
LD R	sho t	me t
LDR	shot	met

Other examples:

• the verbs *kneel*, *sleep*, *creep* for which the double letter (*oo*, *ee*) is simplified in the past tense and the past participle and the letter *t* is added (*V7.grf* graph).



We will now add these eight nouns and twelve verbs into the *myDictionary.dic* dictionary (do not forget to press the *Enter* key after the last one). This dictionary now has 52 entries.



We will save the dictionary and close the Unitex editor.

After opening the *myDictionary.dic* file, using the *DELA/Open Recent* menu, and inflecting it, choosing the *DELA/Inflect* menu, the new *myDictionaryflx.dic* dictionary contains 144 lines (if the *Sort and factorize inflectional codes* option is on). After compressing this dictionary (*DELA/Compress into FST* menu), we will close both windows (*myDictionary.dic* and *myDictionaryflx.dic*). We will apply the new *myDictionaryflx.bin* dictionary (*Text/ApplyLexical Resources* menu). The *Word List* now displays 103 recognized words. We will close the *Word List*.

3 Inflection of multi-word units

Before starting this section

If you have not modified the *Norm.txt* file as explained in <u>Chapter 2, Section 3.1.2</u>, page 35, do so, or download it, then close and reopen Unitex.

In order to inflect multi-word units (MWU), or compound words, we will create graphs that have as many boxes as there are tokens in an MWU with the addition of one box for the result. One has to remember that processing units, or *tokens*, for Unitex are strings of alphabetic characters (letters) and any other individual character, including the space.¹¹²

3.1 Example of the MWU *air of mystery*

MWUs *air of mystery, sort of thing* and *tree of knowledge*, each comprised of five tokens (three alphabetic strings and two spaces), of which only the first one is inflected to obtain the plural form. We will inflect these MWUs with the graph named *NC_XWW1*.¹¹³

We will now add three entries to the *myDictionary.dic* file (and click *ENTER* after the third line) using the Unitex editor (*File Edition/Open.../Dictionaries*; answer *OK*). These entries contain the information about the inflection of the MWU as a whole (in this case *NC_XWW1*) and between parentheses (without any space) information about the inflection of each word token (if it inflects): lemma, inflectional graph and inflectional information of a simple word form used in a MWU being described.

air(air.N1:s) of mystery,NC_XWW1 sort(sort.N1:s) of thing,NC_XWW1 tree(tree.N1:s) of knowledge,NC_XWW1

The *NC_XWW1.grf* graph consists of six boxes: the first five boxes refer to the five tokens in a MWU, while the sixth box is empty, it is used only for output.



The second (<\$2>) and the fourth (<\$4>) token correspond to spaces, so they do not inflect. The third (<\$3>) and the fifth (<\$5>) token refer to word tokens that do not inflect in described MWUs. Only the first token inflects in number. In order to indicate this, we will use the keyword *Nb* representing the grammatical category of number.¹¹⁴ The formula *Nb=\$n* means that the variable *\$n* is instantiated to all possible values of the keyword *Nb*, in this case, *s* and *p*. This means that the single path in the graph corresponds to both singular and plural forms of a MWU being inflected. This formula in the first box indicates that all forms (actually two) of the first word token will be taken in turn, while its use in the sixth box indicates that the corresponding values will be assigned to the inflected MWU forms in the output.

¹¹² See <u>Chapter 2, Section 1.2.3</u>, page 13.

¹¹³ The prefix NC_ in this name indicates that a noun to be inflected is a compound. The suffix is free.

¹¹⁴ The keywords and their possible values are given for each language in the *Morphology.txt* file that has to be in the same folder as the inflectional graphs.



We will close this graph after compiling it. We will open the *myDictionary.dic* file using the *DELA/Open Recent* menu and inflect it using the inflectional graphs in the folder:

$MyUnitex \ English \ Inflection \ UnitexGettingStarted$

The new *myDictionaryflx.dic* dictionary has 150 entries, two for each of the newly added MWU lemmas. After its transformation (*DELA/Compress into FST* menu), we will close the windows with the *myDictionary.dic* and *myDictionaryflx.dic* dictionaries; and open the *Text/ApplyLexical Resources* menu, click the *Clear* button, select the *myDictionaryflx.dic* dictionary and then click the *Apply* button. The number of recognized simple words listed in *Word List* is still 103 (*DLF*), but there are also three MWUs (*DLC*).



We will close the Word List.

3.2 Some other examples

We will open for the last time the *myDictionary.dic* file using the *File Edition/Open.../Dictionaries* menu, and clicking the *OK* button. We will add to it lemmas for some other illustrative examples of MWUs.

- crow's nest, death's-head, hair's breadth; these are MWUs composed, like before, of five tokens, but this time only the last one is inflected (NC_WWX1.grf graph); the only difference between this graph and the NC_XWW1.grf graph is that now the fifth instead of the first box contains the formula Nb=\$n indicating the inflection in number.
- *boa constrictor, prince regent, secretary general;* these are three token MWUs in which only the first word token inflects (*NC_XW1.grf* graph).
- *altar rail, Anglo-Norman, bad boy;* only the last word token inflects (*NC_WX1.grf* graph).
- after-life, eye-cup, sky-blue; these MWUs inflect like the previous except that the first and the third word tokens can be written together (NC_WX2.grf graph); the fact that two words can be written without the hyphen is indicated by the new path that connects directly the first and the third token.

- three quarter, blood red, self glorification; these MWUs inflect like the previous except that the first and the third token can be connected with a hyphen (NC_WX3.grf graph); the new path is added to the NC_WX3.grf graph which connects the <\$1> and <\$3> boxes with a hyphen in between.
- *carry out, take off, wash out* are MWU which unites two previous cases, namely, two word tokens can be either written together or separated by a hyphen or a space, while only the last word inflects (*NC_WX4.grf* graph).
- *attorney general, notary public* are MWUs that allow inflection of either the first or the second word (but not both!) (*NC_XX1.grf* graph); this graph has two paths, the upper taken from the *NC_XW1.grf* graph and the lower from the *NC_WX1.grf* graph.¹¹⁵

We will get the final version of the *myDictionary.dic* dictionary by adding these twenty entries:

.... crow's nest(nest.N1:s),NC WWX1 death's-head(head.N1:s),NC WWX1 hair's breadth(breadth.N1:s),NC WWX1 boa(boa.N1:s) constrictor,NC XW1 prince(prince.N1:s) regent.NC XW1 secretary(secretary.N3:s) general,NC XW1 altar rail(rail.N1:s),NC_WX1 Anglo-Norman(Norman.N1:s),NC WX1 bad boy(boy.N1:s),NC WX1 after-life(life.N5:s),NC_WX2 eye-cup(cup.N1:s),NC WX2 sky-blue(blue.N1:s),NC WX2 three quarter(quarter.N1:s),NC_WX3 blood red(red.N1:s),NC WX3 self glorification(glorification.N1:s),NC WX3 carry out(out.N1:s),NC WX4 take off(off.N1:s),NC WX4 wash out(out.N1:s),NC WX4 attorney(attorney.N1:s) general(general.N1:s),NC XX1 notary(notary.N3:s) public(public.N1:s),NC_XX1

We will construct and compile these seven inflectional graphs.

	crow's nest death's-head hair's breadth
	<pre></pre>
NC_WWX1	

¹¹⁵ We thank Agata Savary for these three examples: *carry out, take off* and *attorney general*.



We will now close all graphs using the *Graphs/Close all* option. Like before, we will open the *myDictionary.dic* file using the *DELA/Open Recent* menu and inflect it (*DELA/Inflect* menu). The final *myDictionaryflx.dic* dictionary consists of 216 entries. After transforming it (*DELA/Compress into FST* menu), we will close both *myDictionary.dic* and *myDictionaryflx.dic* windows. We will apply our dictionary to the text (*Text/ApplyLexical Resources*, clear all preselections and select *myDictionaryflx.dic*). The number of recognized MWUs listed in the *Word List* is now 21 (*DLC*).

4 Some additional remarks

4.1 Adding features in dictionary entries

Features with additional information can be added to Unitex dictionaries. In the Unitex user manual, these features are referred to as *semantic codes*. In a dictionary of lemmas these features follow an inflectional class (a name of an inflectional graph) and are preceded by the + character. Features that are added to the entries in a dictionary of lemmas are transferred to the entries of a dictionary of inflected forms.

Let us turn again to our dictionary of lemmas (*myDictionary.dic*) that we will modify by adding semantic codes: *abstract* for abstract nouns, *animal* for nouns denoting animal, *concrete* for concrete inanimate nouns and *human* for nouns denoting human.

advantage,N1+abstract agent,N1+human firm,N1+concrete ... wash out(out.N1:s),NC_WX4+abstract notary(notary.N3:s) public(public.N1:s),NC_XX1+human attorney(attorney.N1:s) general(general.N1:s),NC_XX1+human

If you add one of these semantic codes to all noun entries in the lemma dictionary, then our form dictionary (*myDictionaryflx.dic*) will have 133 semantically described entries.

```
Anglo-Norman,Anglo-Norman.N+human:s
Anglo-Normans,Anglo-Norman.N+human:p
...
advantage,advantage.N+abstract:s
advantages,advantage.N+abstract:p
after-life,after-life.N+abstract:s
after-lifes,after-life.N+abstract:p
afterlife,after-life.N+abstract:p
afterlifes,after-life.N+abstract:p
agent,agent.N+human:s
agents,agent.N+human:p
...
```

These semantic codes can be used to refine a lexical mask, for instance by writing <*N*+*abstract*>.

4.2 Direct creation of inflected entries

It is possible to add entries directly into the *myDictionaryflx.dic* dictionary, which can be useful, for instance, for words that do not inflect. However, one has to be careful because if

the *myDictionaryflx.dic* dictionary is recreated from *myDictionary.dic*, these manually added entries would be lost. It is better thus to have a new inflectional graph, named N13.grf, describing nouns that do not change in plural, like *moose*, *sheep* or *aircraft*.



On the other hand, one can directly create a dictionary of inflected forms, which can be useful, for instance, for a dictionary of proper names that usually do not inflect in number. For instance, a *myProperNameDictionary.dic* dictionary can be created directly containing the following entries, collected on the text as you go.



This dictionary can then be sorted by the *Dela/Sort Dictionary* menu.

Dombey,.N+surname:ms:fs Florence,.N+forename:fs Paul,.N+forename:ms Peps,.N+surname:ms:fs

After transforming it (*DELA/Compress into FST* menu), we will close the *myProperNameDictionary.dic* window. We will apply our two dictionaries to the text (*Text/ApplyLexical Resources*, clear all pre-selections and select *myDictionaryflx.dic* and *myProperNameDictionary.dic* by *CTRL*-clicking the mouse button). The number of recognized simple word listed in the *Word List* is now 108 (*DLF*).





4.3 Adding comments

4.3.1 Adding comments to entries

It is possible to add comments to entries in a dictionary of inflected forms. They are placed at the end of a line, after a slash "/". For instance, we can supply some explanatory information in the *myProperNameDictionary.dic* dictionary. This information does not appear in the *Word List*.

Dombey,.N+surname:ms:fs Florence,.N+forename:fs/the daughter of Mr Dombey Paul,.N+forename:ms/the son of Mr Dombey Peps,.N+surname:ms:fs/the doctor

It is also possible to add comments to the dictionary of lemmas, but these comments are not transferred to the inflected dictionary. It can be sometimes important to memorize who added an entry and when. If we use instead of *myDictionary.dic* the following dictionary of lemmas with added comments, we would still obtain the same *flx* dictionary. Let us modify, for example, the first three lines, then inflect the dictionary. The comments are not in the dictionary of inflected forms.





4.3.2 Documenting dictionaries

Finally, we can provide a description of our two dictionaries in two files, *myDictionaryflx.txt* and *myProperNameDictionary.txt*, which placed in the same folder as dictionaries themselves. For instance (for both):

Unitex Getting Started Denis Maurel University of Tours Cvetana Krstev University of Belgrade

This information can be displayed by accessing the *Text/ApplyLexical Resources* menu and right-clicking the name of the dictionary.

4.4 Automatically verifying the format of a dictionary

4.4.1 Inflected dictionary

After adding new entries to a dictionary, we can verify if it is well formed, that is, in the format expected by Unitex. Let us take, for example, the *myDictionaryflx.dic* dictionary and open it using the *DELA/Open Recent* menu. If we select the *DELA/Check Format* menu and click the *Check Dictionary* button, the file named *CHECK_DIC.TXT* will appear in the *DELA* folder and its content will be displayed in a window.

The content of this file in the case of our dictionary will be:

- 1. Quantitative information: 219 lines read, 159 simple entries for 61 distinct lemmas, 60 compound entries for 23 distinct lemmas.
- 2. The list of used characters and their UTF8 code.
- 3. The used grammatical and semantic codes: 6 grammatical/semantic codes used in dictionary: N, human, V, abstract, concrete and animal.
- 4. The used morphological codes: 17 inflectional codes used in dictionary, s, p, P1p, P1s...

However, if we have not manually added anything to the *myDictionaryflx.dic* dictionary, that is, if it has been automatically produced from *myDictionary.dic*, there is no need to verify its format, since it must be correct.

4.4.2 Lemma dictionary

What we can do is to check the correctness of the dictionary of lemmas (*myDictionary.dic*). After opening it in the *DELA* menu and choosing *DELA/Check Format* we can verify its format by checking the *DELAS/DELAC* option.¹¹⁶ After applying this option to our dictionary of lemmas, the content of the *CHECK_DIC.TXT* file displayed on the screen will contain the similar information (and replace the previous one):

- 1. Quantitative information: 78 lines read, 55 simple entries for 55 distinct lemmas, 23 compound entries for 23 distinct lemmas.
- 2. The list of used characters and their UTF8 code.
- 3. The used grammatical and semantic codes: 32 grammatical/semantic codes used in dictionary: *N1*, *abstract*, *human*, *concrete*...

¹¹⁶ This option corresponds to dictionaries of lemmas, while the *DELAC/DELACF* option that is selected by default corresponds to dictionaries of inflected forms.

4. The used morphological codes: *0 inflectional codes used in dictionary*.

4.5 Morphological-mode dictionary (only for confident users)

Before starting this sectionGo to the folder:MyUnitex\English\Graphs\UnitexGettingStartedCreate a new folder named Dictionaries.

Dictionary information can be inserted into a text by a graph, for example, to tag it. To do this, you must know how to insert the morphological mode (see <u>Chapter 3, Section 3</u>, page 63) and use variables (see <u>Chapter 4, Section 1.2</u>, page 79).

We will use the *myProperNameDictionary.dic* dictionary to markup proper names.

First (don't forget this point!), this dictionary must be allowed for use in the morphological mode: open the *Info/Preferences.../Morphological mode dictionaries* menu and click the *Add* button to select the *myProperNameDictionary.bin* dictionary.

We will also use the *myProperNameDictionaryTag.grf* graph below, saved in the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Dictionaries*

PN is a dictionary-entry variable, placed as output text. In the second output, PN.CODE.SEM is the semantic code corresponding to the dictionary entry (*forename* or *surname*); PN.INFLECTED is the inflected form corresponding to the dictionary entry (i.e. the proper noun itself).¹¹⁷



This graph will produce, in *Replace* mode and with the *Text Order* option, the concordance with 3 493 proper names displayed below. If graphs that you will produce and use later no longer use this dictionary in morphological mode, remember to remove it from the list.

🔲 Concordance: U:Documents/MyUnitex/English/Corpus/DombeyAndSon/DombeyAndSon_snttconcord.html 🖉 🗖
3493 matches
what was a girl to < <u>ProperName type="surname">Dombey</u> and Son!
oy-nothing more. Mr < <u>ProperName type="surname">Dombey</u> 's cup of 📒
ghter. So he said, "< <u>ProperName type="forename">Florence</u> , you m
hing else," said Mr < <u>ProperName type="surname">Dombey</u> to himsel
bless me!" said Mr < <u>ProperName type="surname">Dombey</u> , rising t
ring there for Miss <u><propername type="forename">Florence</propername></u> 's nurs
d better ask Doctor <u><propername type="surname">Peps</propername></u> if he'll ha
remember when Miss < <u>ProperName type="forename">Florence</u> was bo:
y, ay, ay," said Mr <u><propername type="surname">Dombey</propername></u> , bending
he same time. "Miss < <u>ProperName type="forename">Florence</u> was al
away. Doctor Parker < <u>ProperName type="surname">Peps</u> , one of the
said Doctor Parker <u><propername type="surname">Peps</propername></u> in a round,

4.6 **Priorities**

If we use multiple dictionaries, we can set up three priority levels. To do this, we will add the symbols - (maximum priority) or + (minimum priority) to the name of the dictionary file.¹¹⁸

¹¹⁷ See the Unitex user manual, Section 6.4.4, for more information on dictionary-entry variables.

¹¹⁸ We have to change the name of two files (*.bin* and *.inf*). If it exists, we have to do the same for the *.txt* file.

For instance, if we want to use our two dictionaries, *myDictionary.dic* and *myProperNameDictionary.dic*:

With the dictionary selection:	A word present in both dictionaries will be labeled
myDictionaryflx.bin myProperNameDictionary.bin	by each dictionary
myDictionaryflx.bin myProperNameDictionary+.bin	only by the <i>myDictionary.bin</i> dictionary
myDictionaryflx.bin myProperNameDictionarybin	only by the <i>myProperNameDictionarybin</i> dictionary

4.7 Additional notes

The inflectional morphology of English is relatively simple. In French, common nouns inflect, not only in number, but some of them also in gender, and adjectives inflect too. In Serbian, nouns have gender, animacy, number and case, while all these categories inflect for adjectives, in addition to definiteness and comparison. All this can be modeled using the inflectional graphs introduced before.

When inflection also touches the prefix, then the more complex rules introduced in <u>Section 6</u>, page 132 must be used.

For Semitic languages, like Arabic, Hebrew and so on, in which inflection often operates between consonants in the root, specific rules were developed (see the Unitex user manual, Section 3.5.4).

5 Dictionary graphs

5.1 Example of Roman Numeral dictionary

You will recall that we have constructed the *RN2-3999.grf* graph to tag Roman numerals (see <u>Chapter 3, Section 3.4.2</u>, page 67). We now propose to create a dictionary graph for Roman numerals. We will name it *RomanNumeralDictionary.grf* and we will base it on the *RN2-3999.grf* graph. So, we can copy the four graphs: *RN1-9-subgraph.grf*, *RN1-99-subgraph.grf*, *RN1-999-subgraph.grf* and *RN2-3999.grf* from the folder:

and paste them in the folder:

MyUnitex\English\Dela

then rename the RN2-3999.grf graph as RomanNumeralDictionary.grf.

We are going to remove the insertions of an opening tag *<romanNumeral>* and a closing tag *</romanNumeral>*; and we replace them with empty boxes (*<E>*). We will also replace the call to *:RN2-999* with a call to *:RN1-999-subgraph* and place this call inside morphological tags as well. We will add a new box at the end of the graph with the output *,.RN*.



We will compile this graph, open the *Text/Apply Lexical Resources* menu and click the *Clear* button (in order to cancel all previous selections). We will choose the *RomanNumeralDictionary.fst2* file (appearing in the left field) and click the *Apply* button.¹¹⁹ As a result the *Word List* window will be displayed.



5.2 Morphological extension of the used dictionaries (only for confident users)

5.2.1 Morphological-mode dictionary declaration

It is also possible to extend the used dictionaries by creating morphological extensions of some entries. The root dictionary must be declared as a morphological-mode dictionary (don't forget this point!). See <u>Section 4.5</u>, page 127.

If we consider that the default English dictionary, *dela-en-public*, is the root dictionary, we have to open the *Info/Preferences.../Morphological mode dictionaries* menu and click the *Add* button to select the *dela-en-public.bin* file. This file is located in your installation folder (see <u>Chapter 2, Section 1.1.2</u>, page 7):

C:\Users\userName\AppData\Local\Unitex-GramLab¹²⁰

or

C:\Program Files (x86)\Unitex-GramLab\App

¹¹⁹ We can also create a text file to write a documentation of the dictionary graph, *RomanNumeralDictionary.txt*, see <u>Section 4.3.2</u>, page 126.

¹²⁰ *userName* is your name under the Windows system.

5.2.2 Example of the -less suffix

Generally, a noun can generate an adjective with the addition of the *-less* suffix. For instance, in our novel some such adjectives occur: *remorseless* from *remorse*, *restless* from *rest*, *colourless* from *colour* and so on. Many of these adjectives are entries in the default English dictionary, *dela-en-public*, as we can see if we use the *Dela/Lookup...* menu (see <u>Chapter 4</u>, <u>Section 3</u>, page 88). However, in our novel, three adjectives derived from nouns with *-less* suffix occur that are not in this dictionary: *brotherless*, *nevyless*¹²¹ and *whiskerless*. The *Nless.grf* dictionary graph recognizes and tags them appropriately. In addition to the POS tag *A*, it adds two tags: *Derivative* to indicate that this is an input constructed by a graph and *Root=\$n.LEMMA\$* to specify the root.¹²²

This graph should be added in the list of resources used to analyze the text: we must save and compile it in the folder:

$MyUnitex \ English \ Dela$

It recognizes adjectives composed of a noun (in a morphological-mode dictionary, not necessarily present in the applied dictionaries) and the *-less* suffix. We will now apply lexical resources to our text, this time adding to preselected dictionaries the *NLess.fst2* dictionary graph.¹²³ The number of simple-word lexical entries is now 22 674 (compared to 22 604 obtained before).¹²⁴





¹²¹ The word *nevy* is an old form of *nephew*.

¹²² The dictionary-entry variables are explained in <u>Section 4.5</u>, page 127.

¹²³ We can also create a text file to document the dictionary graph, *NLess.txt*, see <u>Section 4.3.2</u>, page 126.

¹²⁴ See <u>Chapter 2, Section 1.2.3</u>, page 13.

Word Lists in U:Documents/MyUnitex/English/Corpus/DombeyAndSon/Dombe	eyAndSon_snt 🛛 🗖 🛛	Word Lists in U:\Documents\MyUnitex\English\Corpus\DombeyAndSon\Dombe	yAndSon_snt 🛛 🗖 🛛
Word Lists in U/Documents/MyUnitexEnglish/Corpus/Dombey/AndSonDombo DUE: 22874 simple-word ledcal entries neutral, .A neutral, .NHum:s nevertheles, .ADV new, .A new, .AUV new, .AUV newly, .NUV newly, .NUV	eyAndSon_snt r r r r r r r r r r r r r r r r r r r	<pre>Word Lists in U.DocumentsWhyUntexEnglishCorpusDombeyAndSonDombe DLF:22674 simple-word lexical entries Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless, Whitkerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker Whiskerless.A=Deviative=Root=Whisker ULC:950 compound lexical entries</pre>	AndSon_snt of D 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Loc So Composition Monte Solutions able seman, M-XM+Hum+11s ablent friend, M-XM+Hum+11s active service, M+XM+11s active service, M+XM+11s after lock, M+XM+11s after lock, M+XM+11s ai cony, X+XM+11s ai alone, A+XM+11s al alone, A+XM+1 al alone, A+XM+1 Active service Solutions al alone, A+XM+1 Active Solutions Active service Solutions Active Solutions Acti	Norwoll Norwood obleege obserwation Oho Oppgrad ood ooman oop	DLC: 950 Compound ionce enternes abbent friend, N+XN+Hum=115 abbent friend, N+XN+Hum=115 Act of Parliament, N+NN+E115 Act of Parliament, N+NN+E115 after dark, N+XN+E115 after dark, N+XN+E115 after dark, N+XN+E115 after dark, N+XN+E115 aft of mystery, N+NN+E115 all alone, A+DA+E1 all around, A+DA+E1 y	Nickam viddle Niggs vidernesses vider volce volce vollume vonderings vonderest

We notice that our graph recognized two of the three afore-mentioned adjectives. It could not recognize the adjective *nevyless* because the noun *nevy* itself is not in the applied dictionary.

This dictionary graph adds seventy entries to the *Word List*. We found two, what are the other sixty eight?

First, two errors, from the *b*,.*N* entry.

```
bless, bless. A+Derivative+Root=b
Bless, bless. A+Derivative+Root=b
```

Second, four entries starting with an uppercase letter, although already present in lowercase.¹²⁵

Homeless, homeless. A+Derivative+Root=home Restless, restless. A+Derivative+Root=rest Speechless, speechless. A+Derivative+Root=speech Supperless, supperless. A+Derivative+Root=supper

Third, sixty four duplications of entries already present.

```
artless,.A
artless,artless.A+Derivative+Root=art
...
voiceless,.A
voiceless,voiceless.A+Derivative+Root=voice
```

We will now make a small addition to our graph. The *NlessCorrected.grf* dictionary graph only recognizes the two adjectives ending with *-less* that are not already in applied dictionaries (*brotherless* and *whiskerless*). It uses a negative right context (see <u>Chapter 3, Section 1.7.3</u>, page 55) and a <DIC> mask (see <u>Chapter 2, Section 2.4.3</u>, page 34).



We will now apply lexical resources to our text, this time adding to preselected dictionaries the *NlessCorrected.fst2* dictionary graph.

¹²⁵ If the graph does not add a lemma (the output *,*.*A* instead of the output *,*\$*n*.*LEMMA\$less*.*A*), the graph creates four new lemmas, *Homeless*, *Restless*, *Speechless* and *Supperless*, which is best avoided.

Unitex Getting Started - Denis Maurel and Cvetana Krstev

🗒 Lexical Reso	urces		X
Select the diction using the arrows program before t	aries to be applie . Note that system he user ones.	d. You can sort the dictionaries are g	m one by one iven to the Dico
User resources		System resource	es
myDictionaryftx.t myProperName NIess.fst2 NIessCorrected RomanNumera	Dictiona fst2 Dictiona	dela-en-public.b Dnum-ord.fst2 Dnum.fst2	in A
4			
Right-click a dictio	nary to get inform	ation about it :	
rsigne-ciiCK à dictio	mary to get inform	auon auout it :	
Clear	Default	Set Default	Apply

The number of simple-word lexical entries is now 22 606 entries.

6 Additional possibilities (only for confident users)

6.1 A test on a suffix

By testing the last letter of a verb, we can join the V2.grf and V3.grf graphs into one graph, named V101.grf. In order to illustrate this, we will create the myAdvancedDictionary.dic file containing the following six entries (remember that there should be no spaces, and that you should press the *Enter* key at the end of the last line). You will remember that the first three of these verbs were inflected with the V2.grf graph in our myDictionary.dic dictionary, while the last three verbs were inflected with the V3.grf graph.

In the new V101.grf graph, we will use tests on the verb suffixes and inflect all six verbs. The occurrence of *<e>* in an inflectional graph means that the test is made on the verb suffix; if this suffix is *e*, it is removed and the inflection proceeds further. For instance, in order to produce forms of the third person singular of the present tense:

- the <e> box reads the suffix e, removes it and the next box adds es:P3s
- the <y> box reads the suffix y, removes it and the following boxes write first i and then es:P3s

which yields, for instance, *lovees:P3s* and *accompanyies:P3s*.



Now we can inflect the *myAdvancedDictionary.dic* dictionary by choosing the *DELA* menu and the *Inflect* option. The file containing the *myAdvancedDictionaryflx.dic* inflected dictionary will be created and displayed on the screen.



6.2 Use of variables

6.2.1 Example of goose, tooth, foot

Let us look at these three nouns: goose, tooth, foot. With the N9.grf graph we can inflect the first two nouns, but not the third for which a new graph should be written in which the rule LLLLeeRR would be replaced by LLLLeeR (there is one consonant, instead of two at the end of the noun). There is another way to formulate the rules that is more complex but also more powerful. It is based on two variables, $\pounds n$ and \$ n, that memorize the longest and the shortest strings, respectively, that occur where they are placed in the pattern (put in angular brackets)

in a box of the graph. This pattern is matched with the input word, starting from the right end.¹²⁶ Several such variables can be used and numbered from n=1 to n=9.

In the *N101.grf* graph the first box contains the pattern < f100f2 > which means that the variable f2 will be assigned the longest string preceded by *oo* in the processed word (that is, *se*, *th*, *t*, respectively), then the variable f1 will be assigned the longest string before *oo* (that is, *g*, *t*, *f*, respectively). In the case of the singular, the word is rewritten, while for the plural *oo* is replaced by *ee* keeping everything before and after it as it was, which seems as the most natural way to express this rule.¹²⁷



6.2.2 Example of bring, fight, think

The use of variables can be further illustrated with the inflection of the verbs *bring*, *fight*, *think* for which the letters before *i* are suffixed by *ought* in the past tense and the past participle (*V102.grf* graph).



¹²⁶ For detailed explanation of the meaning and use of these variables, see the Unitex user manual, Section 3.5.2.

¹²⁷ Note that in the string £100£2, which is not inside angular brackets, variables £1 and £2 are not assigned values, but their previously assigned values are used.

We will add to our *myAdvancedDictionary.dic* dictionary the following six entries (without spaces and with clicking the *Enter* key at the end of the last line):

goose,N101 tooth, N101 foot, N101 bring,V102 fight,V102 think,V102

We will follow the preceding steps to inflect the *myAdvancedDictionary.dic* dictionary. The created file named *myAdvancedDictionaryflx.dic* will be displayed on the screen.

D:\Documents\MyUnitex\English\Dela\myAdv	ancedDictio	naryflx.dic		o' 🛛			
	Find	Reload	•	•			
accompanied, accompany.V:Ilp:Ils:I	2p:I2s:I	3p:I3s:K		-			
accompanies, accompany.V:P3s							
accompaning, accompany.V:G							
accompany, accompany.V:Plp:Pls:P2p	:P2s:P3p	· · W					
bring, bring.V:Plp:Pls:P2p:P2s:P3p	: W						
bringing, bring.V:G							
brings,bring.V:P3s				=			
brought, bring.V:Ilp:Ils:I2p:I2s:I	3p:I3s:F	t i i i i i i i i i i i i i i i i i i i					
feet,foot.N:p							
fight, fight.V:Plp:Pls:P2p:P2s:P3p	: W						
fighting,fight.V:G							
fights,fight.V:P3s							
foot,foot.N:s							
fought,fight.V:Ilp:Ils:I2p:I2s:I3	p:I3s:K						
geese,goose.N:p							
give,give.V:Plp:Pls:P2p:P2s:P3p:W							
<pre>gived,give.V:Ilp:Ils:I2p:I2s:I3p:</pre>	I3s:K						
gives,give.V:P3s							
giving, give.V:G							
goose,goose.N:s							
live,live.V:Plp:Pls:P2p:P2s:P3p:W							
lived,live.V:Ilp:Ils:I2p:I2s:I3p:I3s:K							
lives,live.V:P3s				•			

1 Lexicon-grammar

Before starting this section

Go to the folder:MyUnitex\English\Graphs\UnitexGettingStartedCreate a new folder named LexiconGrammar (without space).Go to the folder:MyUnitex\English\Graphs\UnitexGettingStarted\LexiconGrammarCreate a new folder named ParticleVerbs (without space).Go to the folder:MyUnitex\English\Graphs\UnitexGettingStarted\LexiconGrammar\ParticleVerbsGo to the folder:MyUnitex\English\Graphs\UnitexGettingStarted\LexiconGrammar\ParticleVerbsDownload the particleVerbTable.xlsx file.

1.1 Example: particle verbs

We will click the *Text/Open...* menu to open the file:

MyUnitex\English\Corpus\UnitexGettingStarted\DombeyAndSon\DombeyAndSon.txt

Then answer Yes to the question *Do you want to preprocess the text?* A window will open in which we will remove the first two checkmarks, leaving just the last one (*Apply All default Dictionaries*).

The aim of this section is to automatically transform a lexicon-grammar table into a graph. We will prepare in some spreadsheet program a table that describes how some verbs combine with a set of particles to obtain particle verbs. This table *particleVerbTable.xlsx* is in the folder:

	Α	В	С	D	E	F	G	н	1	J	К
1	Particle verbs										
2			around	away	back	down	in	off	on	out	up
3	break	+	-	+	+	+	+	+	-	+	+
4	carry	+	+	-	+	+	+	+	+	+	-
5	get	+	+	+	+	+	+	+	+	+	+
6	give	+	-	+	+	+	+	+	+	+	+
7	put	+	+	+	+	+	+	+	+	+	+
8	speak	+	-	+	+	-	-	+	+	+	+
9	turn	+	+	+	+	+	+	+	+	+	+
10											

The A column in the table contains particle verbs. The plus sign in B column indicates that the verb can be used without particles (all verbs in our table can be used without a particle as well). The C to L columns signal particles with which a verb can be used. The plus sign in a cell indicates that the verb and the particle form a particle verb, while the minus sign means the opposite.

1.2 Transformation of the spreadsheet table

The table that we prepared in a spreadsheet program must be transformed into a Unitex lexicon-grammar table.

First, in the *particleVerbTable.xlsx* spreadsheet, verbs in *A* column must be enclosed in angular brackets so that they should match all forms of the listed verbs. *B2* cell is empty because it is the header of the column about the use of the verbs without a particle; we will write here the symbol *<E>* that corresponds to an empty box in a graph.This file should consist only of table

data with only one header line. We will therefore select A2:K9 (the 2 to 9 lines and the A to K columns) and copy its content. We will now open the Unitex editor (the File Edition/New File menu) and paste the content from the table.¹²⁸

We will save this file with the name *particleVerbTable.txt* in the same folder:

MyUnitex\English\Graphs\UnitexGettingStarted\LexiconGrammar\ParticleVerbs

U:\Documents\M	yUnitex\Engli	sh\Graphs	\UnitexGetti	ngStarted\L	exiconGra	mmar\Particl	eVerbs\6-1	2\particleVe	rbTable.txt	∭ ¤ Ø
File Edit										
<e></e>	around	away	back	down	in	off	on	out	up	
<break> +</break>	-	+	+	+	+	+	-	+	+	
<carry> +</carry>	+	-	+	+	+	+	+	+	-	
<get> +</get>	+	+	+	+	+	+	+	+	+	
<give> +</give>	-	+	+	+	+	+	+	+	+	
<put> +</put>	+	+	+	+	+	+	+	+	+	
<speak> +</speak>	-	+	+	-	-	+	+	+	+	
	1 A A A A A A A A A A A A A A A A A A A	1.1	1 A A A A A A A A A A A A A A A A A A A	1.1	1 A A A A A A A A A A A A A A A A A A A			+		

1.3 Parameterized graphs

We will write a graph that retrieves in a text the constructions described in the table and we will save it in the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**LexiconGrammar**ParticleVerbs*

using the name *particleVerbParameterized.grf*. This graph uses variables @A, @B, @C... in which the content of cells A, B, C... will be stored. If the variables contain a text it will be copied in the working graph produced from this parameterized graph. If the value of the variable is a plus or a minus sign, the box functions as a test: with +, the test is passed and the graph proceeds with what follows, while with - the path is interrupted. There is one more variable, @%, that contains the number of the line, which we will use for the name of subgraphs.



1.4 Generation of graphs

We will make Unitex generate a graph from this parameterized graph and the table. This graph will invoke one subgraph for each line in the table. In order to do this, we must open the

¹²⁸ By using the Unitex editor, we are sure that the lines and columns are in the Tabulation/Line break format and that the text is in UTF-8.

Lexicon-Grammar/Open menu and select the particleVerbTable.txt table. Next, we must use the Lexicon-Grammar/Compile to GRF... menu to open (by clicking the SET button in the first line) the particleVerbParameterized.grf parameterized graph. In the second line, we can choose the name of this graph and modify it to particleVerb.grf. In the third line, we will choose the previous graph name and change it to particleVerb@%.grf (remember that @% is the special variable containing the line number). Finally, we will click the Compile button.

Compile Lexicon-Grammar to GRF	X					
Reference Graph (in GRF format):						
mmar\ParticleVerbs\particleVerbParameterized.grf	Set					
Resulting GRF grammar:						
ted\LexiconGrammar\ParticleVerbs\particleVerb.grf Set						
Name of produced subgraphs:						
d\LexiconGrammar\ParticleVerbs\particleVerb@%. Set						
Cancel Compile						

Now we can open the *particleVerb.grf* graph. It invokes seven subgraphs, one for each line in our table (except the header line) which are numbered 0001 to 0007.



We can open the first subgraph by selecting it and clicking the right mouse-button and choosing the *Open subgraph* option. We see that this graph has six paths for six particles used by the verb *break* and one path for the verb without a particle.



By using the produced *particleVerb.grf* graph in the *Text/Locate Pattern* menu we will get 1 983 concordance lines.
Unitex Getting Started - Denis Maurel and Cvetana Krstev

Concordance: U:\Documents\My	Unitex/English/Corpus/DombeyAndSon/DombeyAndSon_snttconcord.html 🗖 🛛
1983 matches	
her daughter should	break away from her and leave her behind; and the daughte
heir talk, he would	break off, to try to understand what it was that the wave
eing where he might	break out, or what authority he might not quote against h
w that Johnson must	break out again, and who knew that the Doctor would never
e a little bow, "to	break up hearts of oak, whether they're in brigs or buzzu
t I can do?" Edith,	breaking her silence, without moving eye or limb, answere
the suffering of a	breaking heart; who, in her rejected love, had never brea
ing all he has, and	breaking his heart. And that if you would, in your kindne
eing present at the	breaking open of the door, and of giving evidence in full
d even to assist in	breaking it up, required a great effort of his resolution
urable a moment for	breaking the West Indian intelligence to his friend, as a
ard; and the horses	breaking loose down below, early in the gale, and tearing
his pocket without	breaking the seal. "These continual references to Mr Car}
nicative," he said,	breaking off, and speaking in a charming tone of franknes
" said the old man,	breaking off, and rubbing his hands in his old dreamy way
usan, reddening and	breaking off. "Much he knows about it!" Florence raised h
" said Miss Blimber	breaking off, "the word analysis as opposed to synthesis,
th the perspiration	breaking out upon his forehead. There was nothing but a v
g, and occasionally	breaking out into a gruff half-smothered fragment of a ba
lent expression as '	' <u>breaking up</u> ," would have been quite inapplicable to that 🖕

1.5 Annotations

The parameterized graph can be more complex than the previous example. For instance, the *particleVerbAnnotationParameterized.grf* graph annotates the recognized verbs (with particles or not).



After repeating the process of graph generation, and its application to the text in the *Merge* mode, we will obtain a new concordance.

🗇 Concordance: U:Documents)MyUnitex/English)Corpus\DombeyAndSon\DombeyAndSon_snt\concord.html 🗖 🗖 🛛
1983 matches
her daughter should < <u>verb>break away</u> from her and leave her behind; and t
heir talk, he would <u><verb>break off</verb></u> , to try to understand what it was tha
eing where he might < <u>verb>break out</u> , or what authority he might not quote
w that Johnson must < <u>verb>break out</u> again, and who knew that the Doctor w
e a little bow, "to < <u>verb>break up</u> hearts of oak, whether they're in brig
. Every tooth was a < <u>verb>break</u> -neck fence, and every pimple in the measle
Fie, fie, it would < <u>verb>break</u> my heart," returned her mother, making a

2 XAlign

2.1 Installation

The aim of this software that is integrated into Unitex is to align two similar texts, for instance two drafts of the same document or a text and its translation. We will demonstrate how to use this software with the example of a text and its translation.

It is important to note that in order to reproduce the exercises from this chapter one should have downloaded the French resources for Unitex. If that is not the case, one should repeat the process of the installation of Unitex. However, it is not necessary to repeat the whole installation.



One should only add a language.

hoose Components Choose which features of Uni	itex/GramLab you want to install.	Choose Components Choose which features of Unitex/GramLab you want to install.
Check the components you w nstall. Click Next to continue.	ant to install and uncheck the components you don't want to	Check the components you want to install and uncheck the components you don't want to install. Click Next to continue.
Select the type of install:	Custom ~	Select the type of install: Custom
Dr, select the optional components you wish to install:	Core Components (required) Visual Integrated Environments Surce Code User Manual Start Menu and Desktop Shortcuts	Or, select the optional components you wish to instal:
	Description	Description
Space required: 30.4MB	Position your mouse over a component to see its description.	Space required: 23.5MB Position your mouse over a component to see its description.
versité Paris-Est Marne-la-Val	liée	Université Paris-Est Marne-la-Vallée
		c Bady Marsh 2 Cons

2.2 The *French* folder

After automatically creating the French folder:

- if Unitex is closed, we should open it by choosing *French* as the working language;
- if it is already open, we will choose *French* as the working language by using the *Text/Change Language...* menu.

It is of course possible to create the French folder yourself, before launching Unitex. But, in this case, Unitex will not copy the three files necessary for its operation (*Norm.txt*,

Alphabet.txt and Alphabet_sort.txt).¹²⁹ You must do this yourself before running Unitex (or after closing and restarting it).

Before proceeding with this section 1. Go to the folder: MyUnitex\English\Corpus\UnitexGettingStarted Create a new folder named DombeyAndSonEng (without space). Go to the folder:*MyUnitex**English**Corpus**UnitexGettingStarted**DombeyAndSonEng* Download the DombeyAndSonEng.txt file. 2. Go to the folder: *MvUnitex**French**Corpus**UnitexGettingStarted* Create a new folder named DombeyAndSonFra (without space). Go to the folder: MyUnitex \English \Corpus \Unitex GettingStarted \DombeyAndSonFra Download the DombeyAndSonFra.txt file. 3. Go to the folder: MyUnitex Create a new folder named XAlign (without space). Go to the folder: MyUnitex\XAlign Create two new folders named Corpus and Sentence. Go to the folder: *MyUnitex*\XAlign\Corpus\UnitexGettingStarted Create a new folder named DombeyAndSonEngFra.

2.3 File preparation

We will use files from three folders: *English, French* and *XAlign*. We will choose *English* as the working language by using the *Text/Change Language...* menu.

2.3.1 English folder

We will open the file (File Edition/Open/Text files menu):¹³⁰

We will ignore the message *This is not necessarily the text being processed by Unitex*. We will save this file using the name *DombeyAndSonEng.xml* (we will change from the *.txt* format to *.xml* format).

We will add TEI tags at the beginning and at the end of this file:¹³¹

<tei></tei>
<teiheader></teiheader>
<body></body>
<text></text>
<div></div>
Dombey and Son by Charles Dickens
A transient flush of faint surprise overspread the sick lady's face as she raised her eyes towards
him.

¹²⁹ About the *Norm.txt* file, see <u>Chapter 2, Section 3.1.2</u>, page 35; about the *Alphabet.txt* and *Alphabet_sort.txt* files, see <u>Chapter 2, Section 3.3.3</u>, page 41.

¹³⁰ We produced these samples of text in English and French just in order to illustrate the alignment process. ¹³¹ XAlign does not handle correctly files with a TEI header, since it will delete it and replace it with the empty tag <*teiHeader/>*. If we want that our result files have a header, we can add it after the alignment.

We will save this file and close the editor.

2.3.2 French folder

Now we will repeat the process for the *French* version of the text: we will open the file (*File Edition/Open/Text files* menu):

We will save this file using the name *DombeyAndSonFra.xml*. Finally, we will add the same tags at the beginning and the end of this file.

<tei></tei>
<teiheader></teiheader>
<body></body>
<text></text>
<div></div>
Dombey et fils traduit de Charles Dickens
Une rougeur passagère, causée par la surprise, colora légèrement les joues de la malade qui leva
les yeux vers son mari.

We will save this file and close the editor.

2.3.3 XAlign folder

We will now open a new file using the editor *File Edition/New File*. In this file, we will copy the following lines:

<teiheader></teiheader> 	<tei></tei>			
	<teiheader></teiheader>			

We will save this file (*Save text* button), using the name *DombeyAndSonEngFra.xml*, in the folder:

MyUnitex\XAlign\Corpus\UnitexGettingStarted\DombeyAndSonEngFra

During the alignment XAlign uses a file for sentence splitting for the source language (which is, in our case, *English*); we will use the graph that comes with the Unitex distribution. We will copy the *Sentence.fst2* file from the folder

*MyUnitex**English**Graphs**Preprocessing**Sentence*

And paste it in the folder:

MyUnitex\XAlign\Sentence

Once this is done, we will rename this file into SentenceXAlign.fst2.

2.4 Automatic alignment

We will click the *Text/Open...* menu to open the file:

and answer Yes to the question Do you want to preprocess the text. A window will open in which we will remove the first two checkmarks, leaving just the last one Apply All default Dictionaries.

We will open the *Xalign/Open Files* menu and fill the first line (by selecting the *Set* button) with the file:

And we fill the second line with the file:

🔲 XAlign	X
Source text	
h\Corpus\DombeyAndSonEng\DombeyAndSonEng.txt	set
Target text	
1ch\Corpus\DombeyAndSonFra\DombeyAndSonFra.txt	set
Alignment file (optional)	
	set
ОКС	ancel

We will click the *OK* button. A window will ask us to choose the XML file that corresponds to the text file in the first line (*source file*). After clicking *OK*, we will choose the first of the previously developed files:

 $\label{eq:myUnitex} MyUnitex \end{SonEng} \label{eq:myUnitex} Oombey \end{SonEng} \end{SonEng}$

Again, another window will open asking us to select an XML file corresponding to the text file given in the second line (*target file*). We will choose the second previously developed file:

 $MyUnitex \ French \ Corpus \ Unitex \ Getting \\ Started \ Dombey \\ And \\ Son \\ Fra \ Dombey \\ And \\ Son \\ Fra \ xml \\ Son \\ Son \\ Fra \ xml \\ Son \\ Son \\ Son \\ Fra \ xml \\$

The new window will open:

E) X/	Align								X
	0	Dombey and Son b	y Charles Dicker	ns	Dom	bey et fils tradui	t de Charles			^
		CHAPTER I. Dombe	y and Son.		DIC	kens			0	
	1	1 Destaura (CHA	PITRE PREMIER.			1	
		darkened room in	the great arm-	chair	Dom	bey père et Dombey	fils.	_		
		by the bedside,	and Son lay tuck	ked up					2	=
=		carefully dispos	ed on a low set	tee	Dan ava	s un coin de la ch it laissé peu de j	ambre ou l'om our, était a	n ssis		
	2	immediately in f	ront of the fire	e and	dan	s un grand fauteui	l, près du l	it,	3	
		close to it, as were analogous t	o that of a muf:	fin,	Dom	bey père.	chaudement			
		and it was essen	tial to toast h	im	env	eloppé , dans un p	etit berceau			
		brown while he w	as very new.		d'o	sier, placé avec s	oin sur un s	ofa		
H		Dombey was about	eight-and-fort	У	du peu	feu. on eût dit un	petit gâtea	pres u	4	
	³ years of age.				auq	uel il fallait fai	re prendre			
	4	Son about eight-	and-forty minute	es.	cou	leur, pendant qu'i dre.	l était enco:	re		
		and though a han	dsome well-made	man,						
	1	too stern and po	mpous in appear	ance,	Dom	bey père avait env	iron		_	
		Son was very bal	d, and very red	, and	qua env	iron quarante-huit	minutes.		5	
	6	though (of cours	e) an undeniabl	y fine	Dom	bey père était un	peu chauve, 1	un		
		infant, somewhat	ffect, as vet.	otty	peu	rouge, et quoiqu' n fait, il avait l	il fût grand 'air trop du	et ret	6	
		On the brow of D	ombey, Time and	his	tro	p guindé pour plai	re à premièr	e	ĭ	
-		brother Care had	set some marks	, as	vue					•
۲	All sentences/Plain text All sentences/Plain text						۲			
C) Ma	atched sentences					Matched s	senten	ces	$^{\circ}$
C	O All sentences/HTML All sentences/HTML All sentences/HTML				ML	$^{\circ}$				
\subseteq	 Aligned with target concordance Aligned with source concordance 				nce	0				
	Lo	cate	Clear alignment	Align	Save alignment	Save alignment as		Loca	te	

We will click the *Align* button that will initiate the alignment process. A new window will ask us to select the XML file for the result. We will choose the third and the last file that we created before, namely the file:

> *MyUnitex\XAlign\Corpus\UnitexGettingStarted\ DombeyAndSonEngFra\DombeyAndSonEngFra.xml*

After the alignment process is finished, some error messages will be displayed, which we will ignore, and click the *OK* button

 Dombey and Son by Charles Dickens CHAFTER I. Dombey and Son. CHAFTER I. Dombey and Son. Dombey sat in the corner of the darkened room in the great arm-chair by the bedside, and Son lay tucked up warm in a little basket bedstead, carefully disposed on a low settee 2 immediately in front of the fire and close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. Bombey was about eight-and-forty warms of age. Bombey was about eight-and-forty minutes. Dombey was rether bald, rather red, and though a handsome well-made man, to be prepossessing. Son was very bald, and very red, and though (of course) an uneliably fine infant, somewhat crushed and spotty in his general effect, as yet. Mathed sentences/Plain ted All sentences/Plain ted Matched sentences All sentences/HTML 	E	U:Documents\MyUnitex\XAlign\Corpus\DombeyAndSonEngFra\DombeyAndSonEngFra.xml				
CHAPTER I. Dombey and Son. Dickens 0 Dombey sat in the corner of the darkened room in the great arm-chair by the bedside, and Son lay tucked up warm in a little basket bedstead, carefully disposed on a low settee Dombey père et Dombey fils. 0 2 immediately in font of the fire and close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. Dombey was about eight-and-forty minutes. Dombey was about eight-and-forty minutes. Dombey was rather bald, rather red, and though a handsome well-made man, too be preposessing. Dombey père avait environ guarante-huit minutes. Son about eight-and-forty red, and though (of course) an undiably find red to be preposessing. Dombey père avait environ guarante-huit minutes. Dombey off Care had set some marks, as All sentences/Plain text @ In the brow of Dombey, Time and his prother Care had set some marks, as All sentences/HIML		Dombey and Son by Charles Dickens	Dombey et fils	traduit de Charles		
<pre>CHAPITE 1. Joundey and Soll. Dombey sat in the corner of the darkened room in the great arm-chair by the bedside, and Son lay tucked up warm in a little basket bedstead, carefully disposed on a low settee limmediately in front of the fire and close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. Dombey was about eight-and-forty greats of age. Son about eight-and-forty minutes. Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be preposessing. Son was very bald, and very red, and though (of course) an undniably fine and though of Dombey, Time and his brother Care had set some marks, as All sentences/Plain text All sentences/Plain text All sentences/HIML A</pre>	Ш	CHARTER I Dombou and Son	Dickens	0		
Dombey sat in the corner of the darkened room in the great arm-chair by the bedside, and Son lay tucked up warm in a little basket bedstead, carefully disposed on a low settee 2 immediately in front of the fire and close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. Dombey was about eight-and-forty years of age. Dombey was rather bald, rather red, and though a handsome well-made man, to be prepossessing. Dombey yere avait environ quarante-huit minutes. Dombey pre avait environ quarante-huit minutes. Dombey pre etait un peut difterand though (of course) an undeniably fine in fint, somewhat crushed and sporty in his general effect, as yet. On the brow of Dombey, Time and his prother Care had set some marks, as Matched sentences/Plain text All sentences/Plain text All sentences/HTML All sentences/HTML	Ш	1 1 1 1 Doubey and Son.	CHAPITRE PREMI	ER.		
darkened room in the great arm-chair Dombey pere et Dombey fils. 2 by the bedside, and Son lay tucked up Dombey pere et Dombey fils. 2 arefully disposed on a low settee Dans un coin de la chambre où l'on avait laissé peu de jour, était assis 3 carefully disposed on a low settee Dombey pere et Dombey fils. Dans un coin de la chambre où l'on avait laissé peu de jour, était assis and it was essential to toast him brown while he was very new. Dombey pils reposait, chaudement enveloppé, dans un petit berceau d'osier, placé avec soin sur un sofa peu élevé, juste en face et tout près du fit, genes at andpous handsome well-made man, too tern and pompous in appearance, to be prepossessing. 0 Son was very bald, and very red, and though (of course) an unelaibly fine infant, somewhat crushed and spotty in his general effect, as yet. Dombey père était un peu chauve, un peu rouge, et quoiqu'il fût grand et his prother Care had set some marks, as Dombey père était un peu chauve, un peu rouge, et quoiqu'il fût grand et his prother Care had set some marks, as Image: All sentences/Plain text All sentences/Plain text All sentences/Plain text @ All sentences/HTML All sentences/HTML	Ш	Dombey sat in the corner of the		1		
Dy the Declate, and Son 149 tucked up warm in a little basket bedstead, carefully disposed on a low settee 1 mmediately in font of the fire and close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. Dans un coin de la chambre où l'on avait laissé peu de jour, était assis dans un grand fauteul, près du lit, Dombey père. 3 3 Dombey to that of a muffin, and it was essential to toast him brown while he was very new. Dombey fils reposait, chaudement enveloppé, dans un peit berceau d'osier, placed avec soin sur un sofa peu élevé, juste en face et tout près du feu on edit dit un petit gâteau auquel il faliait faire prendre couleur, pendant qu'il était encore tendre. 3 mathed some well-made man, too be prepossesing. Dombey père avait environ quarante-huit ans, Dombey fils environ quarante-huit minutes. 5 on was very bald, and very red, and though (of course) an undniably fine in fait, somewhat crushed and spotty in his general effect, as yet. Dombey père était un peu chauve, un peu rouge, et quoiqu'il fut grand et trop guindé pour plaire à première vue. 5 • All sentences/Plain text • All sentences/Plain text • • Matched sentences Matched sentences /	Ш	darkened room in the great arm-chai	Dombey père et	Dombey fils.		
<pre>2 carefully disposed on a low setty carefully disposed on a low setty close to it, as if his constitution were analogous to that of a mulfin, and it was essential to to coast him brown while he was very new. Dombey fils reposait, chaudement enveloppé, dans un peit berceau d'osier, placed avec soin sur un sofa peu élevé, juste en face et tout près du feu. on eft dit un petit gâteau auque il fallait faire prendre couleur, pendant qu'il était encore tendre. Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be preposessing. Son was very bald, and very red, and though (of course) an undniably fine and though (of course) an undniably fine in fair, somewhat crushed and spotty in his general effect, as yet. On the brow of Dombey, Time and his brother Care had set some marks, as All sentences/Plain text O Matched sentences All sentences/HTML All sentences/HTML All sentences/HTML O and though a handence of the course of all sentences/HTML O matche disentences (HTML)</pre>	Ш	by the bedside, and Son lay tucked warm in a little basket bedstead.	p Dang up coin c	le la chambre où llon		
alimmediately in front of the fire and close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. Dombey was about eight-and-forty years of age. Bombey was about eight-and-forty years of age. Bombey was rather bald, rather red, and though a handsome well-made man, to be preposessing. Son was very bald, and very red, and though (of course) an undeniably fine infant, somewhat crushed and spotty in his general effect, as yet. On the brow of Dombey, Time and his prother Care had set some marks, as All sentences/Plain text All sentences/Plain text All sentences/HTML All sentences/HTML And though text And the sentences/HTML		carefully disposed on a low settee	avait laissé p	eu de jour, était assis		
Close to it, as if his constitution were analogous to that of a muffin, and it was essential to toast him brown while he was very new. 3 Dombey was about eight-and-forty yield it and to un petit dire construction of age. 3 Dombey was about eight-and-forty minutes. Dombey was retro bald, rather red, or etern and opmous in appearance, to be prepossessing. Dombey pere avait environ quarante-huit minutes. Dombey was retry bald, and very red, and though (of course) an undeniably fine infant, somewhat crushed and spotty in his general effect, as yet. Dombey pere était un peu chave, un peu rouge, et quoiqu'il fut grand et trop guindé pour plaire à première yue. • All sentences/Plain text All sentences/Plain text • All sentences/HTML All sentences/HTML	Ш	2 immediately in front of the fire an	dans un grand	fauteuil, près du lit, 3		
 and it was essential to toast him brown while he was very new. Dombey was about eight-and-forty years of age. <u>Son about eight-and-forty minutes.</u> Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be prepossessing. Son was very bald, and very red, and though (of course) an undeniably fine dinfant, somewhat crushed and spotty in faits general effect, as yet. On the brow of Dombey, Time and his brother Care had set some marks, as All sentences/Plain text All sentences/Plain text All sentences/HTML 	Ш	close to it, as if his constitution were analogous to that of a muffin	Dombey père.			
 brown while he was very new. Dombey was about eight-and-forty years of age. 4 Son about eight-and-forty minutes. Dombey was rather bald, rather red, and though a handsome well-made man, to be preposessing. Son was very bald, and very red, and though (of course) an undeniably fine infant, somewhat crushed and spotty in his general effect, as yet. Matched sentences/Plain text All sentences/Plain text All sentences/Plain text All sentences/HTML 	Ш	and it was essential to toast him	enveloppé, da	posait, chaudement		
 Jombey was about eight-and-forty years of age. Jombey was rot age. Son about eight-and-forty minutes. Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be prepossessing. Son was very bald, and very red, and through (of course) an undeniably fine infant, somewhat crushed and spotty in his general effect, as yet. On the brow of Dombey, Time and his prother Care had sent some marks, as All sentences/Plain text All sentences/Plain text All sentences/HTML 	Ш	brown while he was very new.	d'osier, placé	avec soin sur un sofa		
a) Double y was about eight-and-forty minutes. au red ait un petit gateau 4 4) Son about eight-and-forty minutes. auquel il fallait faire prendre couleur, pendant qu'il était encore tendre. Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be prepossessing. Dombey père avait environ quarante-huit ans, Dombey fils environ quarante-huit nan, Somewhat crushed and spotty in his general effect, as yet. Dombey père était un peu chauve, un peu rouge, et quoiqu'il fût grand et bien fait, il avait l'air trop dur et 6 • Don the brow of Dombey, Time and his portother Care had set some marks, as All sentences/Plain text • • All sentences/Plain text All sentences/Plain text • • All sentences/HTML All sentences/HTML •	μ	Deplete the short sight and forth	peu élevé, jus	te en face et tout près		
 4 Son about eight-and-forty minutes. Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be preposessing. Son was very bald, and very red, and though (of course) an undeniably fine in fair, somewhat crushed and spotty in his general effect, as yet. In the brow of Dombey, Time and his brother Care had some marks, as All sentences/Plain text Matched sentences All sentences/HTML 		3 vears of age.	du feu. on eut	dit un petit gâteau 4		
Dombey was rather bald, rather red, and though a handsome well-made man, too stern and pompous in appearance, to be preposessing. Dombey pire avait environ quarante-huit minutes. Son was very bald, and very red, and though (of course) an undeniably fine in fast, somewhat crushed and spotty in his general effect, as yet. Dombey pire avait environ quarante-huit minutes. On the brow of Dombey, Time and his brother Care had set some marks, as Dombey pire était un peu chauve, un peu rouge, et quoiqu'il fût grand et bien fast, il avait l'air trop dur et trop guindé pour plaire à première vue. 6 Image: All sentences/Plain text All sentences/Plain text Image: All sentences/Plain text All sentences/HTML All sentences/HTML		4 Son about eight-and-forty minutes.	couleur, penda	int qu'il était encore		
and though a handsome well-made man, foo stern and pompous in appearance, Son was very bald, and very red, and though (of course) an undeniably fine infant, somewhat crushed and spotty in the brow of Dombey, Time and his prother Care had set some marks, as Image: All sentences/Plain text All sentences/Plain text All sentences/HTML		Dombey was rather bald, rather red,	tendre.			
bob stern and pomputs in appearance, it observes and prepare sessing. Dombey pere avait environ it appearance, it oppeares the set of the set		and though a handsome well-made man	Dembers a constant			
Son was very bald, and very red, and though (of course) an undeniably fine infant, somewhat crushed and spotty in his general effect, as yet. Dombey père était un peu chauve, un peu rouge, et quoiqu'il fût grand et bien fait, il avait l'air trop dur et bien fait, il avait l'air trop dur et course. All sentences/Plain text All sentences/Plain text Matched sentences Matched sentences All sentences/HTML All sentences/HTML		to be prepossessing.	guarante-huit	ans. Dombey fils 5		
All sentences/Plain text All sentences/HTML All sentences/HTML All sentences/HTML		Son was very bald, and very red, an	environ quarar	te-huit minutes.		
Infant, somewhat crushed and spotty in his general effect, as yet. On the brow of Dombey, Time and his prother Care had set some marks, as All sentences/Plain text		6 though (of course) an undeniably fi	e Dombey père ét	ait un peu chauve, un		
All sentences/Plain text		infant, somewhat crushed and spotty	peu rouge, et	quoiqu'il fût grand et		
vue. vue. Image: All sentences/Plain text All sentences/Plain text Image: All sentences Matched sentences Image: All sentences Matched sentences Image: All sentences All sentences		On the brow of Dombey. Time and his	trop guindé po	our plaire à première		
All sentences/Plain text All sentences/Plain text Matched sentences Matched sentences All sentences/HTML All sentences/HTML	-	brother Care had set some marks, as	vue.	-		
Matched sentences Matched sentences All sentences/HTML All sentences/HTML		All sentences/Plain text All sentences/Plain text				
○ All sentences/HTML All sentences/HTML ○	C	O Matched sentences Matched sentences C				
	C	○ All sentences/HTML All sentences/HTML				
 Aligned with target concordance Aligned with source concordance 	C					
Locate Clear alignment Align Save alignment Save alignment as Locate		Locate Clear alignment Ali	n Save alignment Save alignm	ent as Locate		

Click the Save alignment button.

You should note that if you close Unitex after saving the alignment, you can open them again by filling all three lines with the appropriate XML files.

T XAlign	X
Source text	
\Corpus\DombeyAndSonEng\DombeyAndSonEng.xml	set
Target text	
h\Corpus\DombeyAndSonFra\DombeyAndSonFra.xml	set
Alignment file (optional)	
Is\DombeyAndSonEngFra\DombeyAndSonEngFra.xml	set
ок с	ancel

2.5 Reading and manual correction

We have done the first alignment. It is necessary now to read and correct it. It is enough to click the *English* sentence (or phrase) and then the *French* sentence (or phrase) to establish the link, if it does not exist, or to remove it if it exists:¹³²

¹³² In order to remove the link it is enough to simply click on it.

remove 4-6;
add 4-5;
remove 5-7;
add 5-6;
remove 6-8;
remove 6-9;
add 6-7;
add 7-8;
add 8-14;
remove 9-14;
add 9-15;
add 10-16;
remove 11-16;
remove 11-17;
add 12-15.

Click the Save alignment button.

Generally, a sentence is translated into another sentence: (0-0), (5-6), (6-7) (9-15) and (13-18). However, there are exceptions:

- three sentences in the source file, 1, 2 and 10, are translated into two sentences: (1-1, 1-2), (2-3, 2-4) and (10-15, 10-16);
- three sentences in the source file, 7, 8 and 12, are translated into three sentences: (7-8, 7-10, 7-11), (8-12, 8-13, 8-14) and (12-15, 12-16, 12-17);
- a sentence in the target file, 9, is added;
- a sentence in the source file, 11, is not translated;
- two sentences in the source file, 3 and 4, are translated together into a single sentence: (3-5, 4-5);
- two sentences in the source file, 10 and 12, are reversed: (10-16, 12-15).

2.6 Search performed on one of two texts

In XAlign we can use Unitex search functions via *Text/Locate Pattern* and get as a result the visualized alignment. For instance, we can be interested in whether the name of Mr Dombey was systematically represented in the French translation. We will click the *Locate...* button, that can be found on the left side of the *XAlign* window, and then answer *Yes* to the question that appears. Unitex proposes to open a text deduced from the XML file, text which it names with the suffix *_xalign* (*DombeyAndSonEng_xalign.txt*).

We will type *Dombey* in the *Regular expression* line and click the *SEARCH* button. The _*xalign* file will be processed by Unitex and the result will appear in the *XAlign* window, after checking the *All sentences/HTML* option (at the left-hand side). If the *Matched sentences* option is chosen only the sentences that correspond to our search query will be displayed.



In the obtained results we can see that, in the translation (12-16), the proper name was translated by *son épouse* (*his wife*).

We will choose the *All sentences/Plain text* option at the left side of the *XAlign* window in order to start a new search. This time we will use a graph, for instance the *willV.grf* graph that we developed in <u>Chapter 3, Section 1.8</u>, page 56.¹³³ We will search for this graph (*Set* button) in the *Graph* line and click the *SEARCH* button. We will select the *Matched sentences* option and we will get only one match in the source file sentence 7.

The same options exist on the right of the *XAlign* window. It is therefore possible to search in the target language.

However, it is not possible to launch a cascade on an aligned text, except if we proceed as explained in the next section (only for confident users).

2.7 Cascade search and XAlign (only for confident users)

The main idea is to launch cascades on the *DombeyAndSonEng.xml* file without touching the and <s> tags placed by XAlign.

2.7.1 Launching cascades

We will go to the folder:

*MyUnitex**English**Corpus**UnitexGettingStarted**DombeyAndSonEng*

and rename the DombeyAndSonEng.xml file as DombeyAndSonEng.xml.txt (see <u>Chapter 4</u>, <u>Section 2</u>, page 84). We will open this text (*Text/Open* menu) and answer Yes to the question Do you want to preprocess the text. A window will open in which we will remove the first two checkmarks, leaving just the last one Apply All default Dictionaries. We will find the analysis cascade (Apply CasSys Cascade... menu) in the folder:

$MyUnitex \ English \ CasSys \ UnitexGettingStarted \ Measures$

And we will launch this cascade. The result file is the *DombeyAndSonEng.xml_csc.txt* file. We will open this text with Unitex, without processing it, and launch the *synthesis* cascade. This second result file is the *DombeyAndSonEng.xml_csc_csc.txt* file with two recognized measures (sentences 4 and 5).

¹³³ Or <u>Chapter 3, Section 1.9.3</u>, page 58, if you have done this section.

<s id="n4" xml:id="d1p1s4">Dombey was about <Measure type="Duration"><Number>eightand-forty</Number> years</Measure> of age. </s><s id="n5" xml:id="d1p1s5">Son about <Measure type="Duration"><Number>eight-and-forty</Number> minutes</Measure>. </s>

2.7.2 File transformation

We will rename the *DombeyAndSonEng.xml_csc_csc.txt* file as *DombeyAndSonEng.xml*. If we open this file with XAlign, the tags disappear. To keep them, we will use the HTML codes *<* for < and *>* for >. We will pay attention to modify only our tags (*Measure* and *Number*) and not the and <s> tags produced by XAlign.

<s id="n4" xml:id="d1p1s4">Dombey was about <Measure type="Duration"><Number>eight-and-forty</Number> years</Measure> of age. </s><s id="n5" xml:id="d1p1s5">Son about <Measure type="Duration"><Number>eight-and-forty</Number> minutes</Measure>. </s>

We will save this file.

2.7.3 Using XAlign

If we open XAlign with this new file:

🗂 XAlign	
Source text	
\Corpus\DombeyAndSonEng\DombeyAndSonEng.xml	set
Target text	
:h\Corpus\DombeyAndSonFra\DombeyAndSonFra.xml	set
Alignment file (optional)	
Is\DombeyAndSonEngFra\DombeyAndSonEngFra.xml	set
OK Can	cel

We obtain the desired result.

Ē) XA	lign									×
	0	Dombey and Son b	y Charles Dicke	ns		Dom	bey et fils traduit	t de Charle:	s		
		CHAPTER I Dombe	y and Son	_		DIC	Kens			0	
	1	CHARTER I. DONDC	y and bon.			CHAI	PITRE PREMIER.			4	
		Dombey sat in th	e corner of the							1	
=		darkened room in	the great arm-	chair		Domi	bey père et Dombey	fils.		2	
		warm in a little	basket bedstea	d,		Dans	s un coin de la ch	ambre où l'e	on		
		carefully dispos	ed on a low set	tee		avai	it laissé peu de jo	our, était d	assis	2	
	2	immediately in f	ront of the fire	e and tion		dans	s un grand fauteui	l, près du l	lit,	3	-
		were analogous t	o that of a muf:	fin,		Dom	bey pere.	chaudement			
		and it was essen	tial to toast h	im		enve	eloppé , dans un p	etit bercea	u		
		brown while he w	as very new.			d'os	sier, placé avec s	oin sur un a	sofa		
		Dombey was about	<measure< td=""><td></td><td></td><td>peu</td><td>élevé, juste en fa</td><td>ace et tout</td><td>près</td><td></td><td></td></measure<>			peu	élevé, juste en fa	ace et tout	près		
	3	type="Duration">	<number>eight-a</number>	nd-for		aum	uel il fallait fai:	re prendre	au	4	
		ty year	s of a	age.		coul	leur, pendant qu'i	l était enc	ore		
	Ι.	Son about <measu< td=""><td>re</td><td></td><td></td><td>tend</td><td>dre.</td><td></td><td></td><td></td><td></td></measu<>	re			tend	dre.				
	4	type="Duration"> ty minu	<number>eight-a tes.</number>	na-ror		Dom	neu nàra susit enu	iron			
		Dombey was rathe	r bald, rather :	red,		qua	rante-huit ans, Dor	mbey fils		5	
	5	and though a han	dsome well-made	man,		envi	iron quarante-huit	minutes.			
	ľ	too stern and po	mpous in appear:	ance,		Dom	bey père était un p	peu chauve,	un		
		Son was very bal	d. and very red	and		peu bier	rouge, et quoiqu': n fait, il avait l	ll Iut gran 'air trop d	a et ur et	6	
		though (of cours	e) an undeniabl;	y fine	-	trop	p guindé pour plai:	re à premiè	re	Ŭ	
-	0	infant, somewhat	crushed and sp	otty		vue.					-
۲	All	sentences/Plain text						All sentence	es/Plain	text	۲
C	O Matched sentences Matched sentences 0						0				
C	○ All sentences/HTML All sentences/HTML						0				
C	 Aligned with target concordance Aligned with source concordance 						0				
	Locate Clear alignment Align Save alignment Save alignment as										
							-				

Before reading this chapter, please, read this caution:

This chapter requires some computer knowledge. You must at least be comfortable with the notion of scripts.

Before starting this ch	apter: ¹³⁴	
1. Go to the folder:	MyUnitex	
Create a new folder na	med Scripts.	
Go to the folder:	$MyUnitex \ Scripts$	
Create a new folder na	med ScriptEntities (without space).	
Go to the folder:	$MyUnitex \Scripts \ScriptEntities$	
Create three new folde	ers named Input, Output and ScriptEntitiesLingpkg.	
Go to the folder:	$MyUnitex \Script Script Entities \Script Entities \Lingpkg$	
Create two new folder	s named Resources and Scripts.	
2. Go to the folder: <i>My</i> a Copy the <i>jubileePlain</i> 7	Unitex\English\Corpus\UnitexGettingStarted\JubileePlainText Text.txt file.	
Go to the folder: Paste this file.	$MyUnitex \ScriptS \ScriptEntities \Input$	
3. Go to the folder: <i>My</i> Copy the <i>jubileeXmlTe</i>	Unitex \English \Corpus \UnitexGettingStarted \JubileeXmlText xt.xml file.	
Go to the folder: Paste this file.	$MyUnitex \ScriptS \ScriptEntities \Input$	

1 Introduction

The Unitex interface is a tool for developing graphs and graph cascades. Once these resources have been developed and tested, it is possible to create Unitex scripts which apply the developed tools on a corpus of several documents stored in several files. These Unitex scripts are launched from a command line and they access a compressed file that can contain all types of resources: Unitex scripts, graphs, cascades and dictionaries. We will take as an example the cascades developed in <u>Chapter 4, Section 1</u>, page 71 and <u>Section 2</u>, page 84. If you have allowed a dictionary for use in the morphological mode, please, remove them now from the list (see <u>Chapter 5, Section 4.5</u>, page 127).

In general, users will copy in the *Input* folder all texts they want to process, as we have done.

2 Creation of the *linguistic package*

The *linguistic package* (*scriptEntitiesLingpkg* folder) will contain everything needed to launch the analysis and synthesis cascades.

¹³⁴ Remember that file names should not contain spaces and are case sensitive. The same goes for folder names. See <u>Chapter 2, Section 1.1.1</u>, page 7.

2.1 The *Resources* folder

We will replicate our personal working folder in the *Resources* folder.

Go to the folder: MyUnitex\ScriptEntities\ScriptEntities\ScriptEntitiesLingpkg\Resources
 Create a new folder named English.
 Go to the folder: MyUnitex\ScriptEntities\ScriptEntitiesLingpkg\Resources\English
 Create three new folders named CasSys, Dela and Graphs.
 Go to the folder: MyUnitex\English
 Copy the three Alphabet.txt, Alphabet_sort.txt and Norm.txt files.
 Go to the folder: MyUnitex\ScriptEntities\ScriptEntitiesLingpkg\Resources\English

Paste these files.

In general, one will copy in the CasSys folder the cascades that one wants to use for the processing.

Go to the folder: $MyUnitex \English \CasSys \UnitexGettingStarted \XmlTextEntities$ Copy the two analysis.csc and synthesis.csc files. Go to the folder: $MyUnitex \ScriptEntities \ScriptEntitiesLingpkg \Resources \English \CasSys$ Paste these files.

In general, users copy the graphs used by cascades in the Graphs folder. In fact, it is enough to copy the compiled files (*.fst2*). However, the presence of source graph files (*.grf*) makes it possible to use the language package as a backup of one's work.

 Go to the folder: MyUnitex\ScriptEntities\ScriptEntitiesLingpkg\Resources\English\Graphs
Create a new folder named UnitexGettingStarted.
 Go to the folder: MyUnitex\English\Graphs\UnitexGettingStarted
Copy the folder Entities. Go to the folder: MyUnitex\Scripts\ScriptEntities\ScriptEntitiesLingpkg\ Resources\English\Graphs\UnitexGettingStarted
Paste this folder.

Neither of the cascades use dictionaries, so we do not need to copy anything to the folder:

 $MyUnitex \ Script Entities \ Script Entities \ Lingpkg \ Resource \ English \ Dela$

Otherwise, if the cascades used dictionaries, see <u>Section 2.4.1</u>, page 155.

2.2 The scripts folder

2.2.1 Using the console

The *scripts* folder will contain the Unitex scripts. The scripts can be developed manually, but the Unitex console will help us.

If Unitex is already open, we need to close and reopen it in order to clean the console. After that, everything we do with the menu will be recorded by the console. We will launch our two cascades.

1. We will use the *Text/Open* menu (answering *No* to the question *Do you want to preprocess the text*) to open the file:

- 2. We will use the *Text/ApplyCasSysCascade*... menu, then subsequently double-click the *UnitexGettingStarted* and *XmlTextEntities* folders to open them; we will select the *analysis.csc* cascade and click the *Launch* button (do not build the concordance).¹³⁵
- 3. We will use the *Text/Open* menu (answering *No* to the question *Do you want to preprocess the text*) to open the file:

*MyUnitex**English**Corpus**UnitexGettingStarted**JubileePlainText**JubileePlainText_csc.txt*

- 4. We will use the *Text/ApplyCasSysCascade...* menu, then select the *synthesis.csc* cascade and click the *Launch* button (do not build the concordance).
- 5. We will use the *Info/Console* menu and copy its content (*CTRL-C*).
- 6. We will use the *File Edition/New File* menu, then paste the commands recorded in the console.
- 7. We will use the *Save text* button and save this file with the name *scriptEntities.uniscript*, placed in the folder:

MyUnitex\Scripts\ScriptEntities\ScriptEntitiesLingpkg\Scripts

¹³⁵ We will use the *XmlTextEntities* cascade, as we will later parse also the *XmlTextEntities.xml* file.

This file contains eight commands, which look like this:136

mkdir "C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText_snt"
"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" Normalize "C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ jubileePlainText.txt" "-rC:\Documents\MyUnitex\English\Norm.txt" "output_offsets=C:\Documents\MyUnitex\ English\Corpus\UnitexGettingStarted\JubileePlainText\JubileePlainText_snt\ normalize.out.offsets" -qutf8-no-bom
"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" Tokenize "C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText.snt" "-aC:\Documents\MyUnitex\English\Alphabet.txt" -qutf8-no-bom
"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" CasSys "-aC:\Documents\MyUnitex\English\Alphabet.txt" "-tC:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText.snt" "-IC:\Documents\MyUnitex\English\CasSys\UnitexGettingStarted\XmlTextEntities\analysis.csc" -v "-rC:\Documents\MyUnitex\English\Graphs\" "
input_offsets=C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText_snt\normalize.out.offsets" -qutf8-no-bom
mkdir "C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText_csc_snt"
"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" Normalize "C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText_csc.txt" "-rC:\Documents\MyUnitex\English\Norm.txt" "
output_offsets=C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText_csc_snt\normalize.out.offsets" -qutf8-no-bom
"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" Tokenize "C:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText_csc.snt" "-aC:\Documents\MyUnitex\English\Alphabet.txt" -qutf8-no-bom
"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" CasSys "-aC:\Documents\MyUnitex\English\Alphabet.txt" "-tC:\Documents\MyUnitex\English\Corpus\UnitexGettingStarted\JubileePlainText\ JubileePlainText_csc.snt" "-IC:\Documents\MyUnitex\English\CasSys\UnitexGettingStarted\XmlTextEntities\synthesis.csc" -v "-rC:\Documents\MyUnitex\English\Graphs\"
"

¹³⁶ Users can obtain a different content if the location and names of the folders in question differ. Do not forget this when preparing a script.

2.2.2 File adjustment

We must adjust this file because of differences between the operating systems Windows, MacOs and Unix:

- 1. We will remove the 56 quotation marks;
- 2. We will replace the 148 backslashes (\) with slashes (/). 137
- 3. We will add two lines at the beginning of the script to create a virtual working folder where the files will be processed one after the other (*CURRENT_WORK_DIR*). The virtual folder and its virtual files will not be stored on the disk, but in random-access memory, which will speed up the processing.

CURRENT_WORK_DIR = {CORPUS_WORK_DIR}/{UNIQUE_VALUE} DuplicateFile -p {CURRENT_WORK_DIR}

4. We will delete the two lines starting with *mkdir* command and replace them by these two lines at the beginning of the script, after the previous two inserted lines.

DuplicateFile --make-dir {CURRENT_WORK_DIR}/corpus_snt DuplicateFile --make-dir {CURRENT_WORK_DIR}/corpus_csc_snt

2.2.3 Script generalization

In order to make this script more general, we will transform this sequence of commands in the following way:

1. We will add a line (it will become the fifth line) to process a file from our folder.

DuplicateFile -i {INPUT_FILE_1} {CURRENT_WORK_DIR}/corpus.txt

2. We will replace the ten occurrences of the beginning of the text file name:

C:/Documents/MyUnitex/English/Corpus/JubileePlainText/jubileePlainText

with:

{CURRENT_WORK_DIR}/corpus

3. We will add a line (the penultimate of the script) to save the result file.

DuplicateFile -i {CURRENT_WORK_DIR}/corpus_csc_csc.txt {OUTPUT_FILE_1}

4. We will remove the six calls to the *UnitexToolLogger* program and the path that leads to it:

C:/Program Files (x86)/Unitex-GramLab/App/UnitexToolLogger.exe

5. We will replace the ten paths that lead to our personal Unitex working folder:

C:/Documents/MyUnitex

with the folder of resources of the package

¹³⁷ This is due to the inversion of these symbols between Windows, on the one hand, and MacOs or Unix, on the other. If you are not using Windows, you can therefore ignore this step. You may have a different number of backslashes if the hierarchy of folders on your computer is different.

{PACKAGE_DIR}/Resources

6. We will modify the two paths leading to our cascades (the two *Cassys* lines) by removing (before *analysis.csc* and *synthesis.csc*):

UnitexGettingStarted/XmlTextEntities/

7. We will add a line at the end of the script, to free up some space before processing the next file in the input folder.

DuplicateFile --recursive-delete {CURRENT_WORK_DIR}

Our script file (thirteen lines) will now look like this:

CURRENT_WORK_DIR = {CORPUS_WORK_DIR}/{UNIQUE_VALUE}
DuplicateFile -p {CURRENT_WORK_DIR}
DuplicateFilemake-dir {CURRENT_WORK_DIR}/corpus_snt
DuplicateFilemake-dir {CURRENT_WORK_DIR}/corpus_csc_snt
DuplicateFile -i {INPUT_FILE_1} {CURRENT_WORK_DIR}/corpus.txt
Normalize {CURRENT_WORK_DIR}/corpus.txt -r{PACKAGE_DIR}/Resources/English/Norm.txt output_offsets={CURRENT_WORK_DIR}/corpus_snt/normalize.out.offsets -qutf8-no-bom
Tokenize {CURRENT_WORK_DIR}/corpus.snt -a{PACKAGE_DIR}/Resources/English/Alphabet.txt - qutf8-no-bom
Cassys -a{PACKAGE_DIR}/Resources/English/Alphabet.txt -t{CURRENT_WORK_DIR}/corpus.snt - I{PACKAGE_DIR}/Resources/English/CasSys/analysis.csc -v - r{PACKAGE_DIR}/Resources/English/Graphs/ input_offsets={CURRENT_WORK_DIR}/corpus_snt/normalize.out.offsets -qutf8-no-bom
Normalize {CURRENT_WORK_DIR}/corpus_csc.txt -r{PACKAGE_DIR}/Resources/English/Norm.txt output_offsets={CURRENT_WORK_DIR}/corpus_csc_snt/normalize.out.offsets -qutf8-no-bom
Tokenize {CURRENT_WORK_DIR}/corpus_csc.snt - a{PACKAGE_DIR}/Resources/English/Alphabet.txt -qutf8-no-bom
Cassys -a{PACKAGE_DIR}/Resources/English/Alphabet.txt - t{CURRENT_WORK_DIR}/corpus_csc.snt - I{PACKAGE_DIR}/Resources/English/CasSys/synthesis.csc -v - r{PACKAGE_DIR}/Resources/English/Graphs/ input_offsets={CURRENT_WORK_DIR}/corpus_csc_snt/normalize.out.offsets -qutf8-no-bom
DuplicateFile -i {CURRENT_WORK_DIR}/corpus_csc_csc.txt {OUTPUT_FILE_1}
DuplicateFilerecursive-delete {CURRENT_WORK_DIR}

This script successively launches two cascades on a file and will produce the final result without writing on the disk numerous intermediate files, which are processed in RAM instead. It is the *DuplicateFile* command which allows access to a virtual folder and its virtual files.

2.2.4 Final touches

Since the *corpus.txt* file has already been normalized, the file resulting from the first cascade, *corpus_csc.txt*, is also normalized. Therefore, it is not necessary to do it again, although it would not be wrong. Since the second cascade uses the *corpus_csc.snt* file, we must rename the *corpus_csc.txt* file produced by the first cascade.

We are therefore going to replace the second *Normalize* command which follows the first *CasSys* command, on line 9, by:

DuplicateFile -i {CURRENT_WORK_DIR}/corpus_csc.txt {CURRENT_WORK_DIR}/corpus_csc.snt

In this case, the offset file is that of the first, and now only, Normalize command.

Normalize {PACKAGE_DIR}/Resources/corpus.txt -r{PACKAGE_DIR}/Resources/English/Norm.txt --output_offsets=<u>{CURRENT_WORK_DIR}/corpus_snt/normalize.out.offsets</u> -gutf8-no-bom

We need to replace the *input_offsets* option of the second *CasSys* command with the *output_offsets* option of the *Normalize* command.

Cassys -a{PACKAGE_DIR}/Resources/English/Alphabet.txt -t{PACKAGE_DIR}/Resources/corpus_csc.snt -l{PACKAGE_DIR}/Resources/English/CasSys/synthesis.csc -v -r{PACKAGE_DIR}/Resources/English/Graphs/ --input_offsets=<u>{CURRENT_WORK_DIR}/corpus_snt/normalize.out.offsets</u> -qutf8-no-bom

With these modifications we obtain the last version of our script.

2.3 Compression

The *scriptEntitiesLingpkg* folder is now complete. The only thing that remains to be done is to create the compressed file. In order to do that we have to open this folder, select the two subfolders, named *Resources* and *Scripts*, compress them together and name the compressed *scriptEntitiesLingpkg.zip* file.¹³⁸ We will cut this compressed file and paste it in the *MyUnitex\ScriptEntities* folder.

2.4 Another possibility (only for confident users)

2.4.1 The use of dictionaries

2.4.1.1 The default English dictionary

If the cascades use the default English dictionary, *dela-en-public*:

Go to the folder where Unitex is installed:¹³⁹ C:\Users\userName\AppData\Local\Unitex-GramLab or C:\Program Files (x86)\Unitex-GramLab\App Copy the three files named dela-en-public.bin, dela-en-public.inf and dela-en-public.txt. Go to the folder:MyUnitex\ScriptEntities\ScriptEntitiesLingpkg\Resources\English\Dela Paste these files.

We must add to our script five new lines between the first *Tokenize* command and the first *Cassys* command. The first one for the *Dico* command and the other four for the *SortTxt* commands that create four files *dlf.n, dlc.n, err.n* and *tags_err.n*.

¹³⁸ It must be a *.zip* file, *.rar* files are not accepted.

¹³⁹ See <u>Chapter 2, Section 1.1.2</u>, page 7.

Dico -t{CURRENT_WORK_DIR}/corpus.snt -a{PACKAGE_DIR}/Resources/English/Alphabet.txt {PACKAGE_DIR}/Resources/English/Dela/<u>dela-en-public.bin</u> SortTxt {CURRENT_WORK_DIR}/corpus_snt/dlf -l{CURRENT_WORK_DIR}/corpus_snt/dlf.n o{PACKAGE_DIR}/Resources/English/Alphabet_sort.txt SortTxt {CURRENT_WORK_DIR}/corpus_snt/dlc -l{CURRENT_WORK_DIR}/corpus_snt/dlc.n o{PACKAGE_DIR}/Resources/English/Alphabet_sort.txt SortTxt {CURRENT_WORK_DIR}/corpus_snt/err -l{CURRENT_WORK_DIR}/corpus_snt/err.n o{PACKAGE_DIR}/Resources/English/Alphabet_sort.txt SortTxt {CURRENT_WORK_DIR}/corpus_snt/err -l{CURRENT_WORK_DIR}/corpus_snt/err.n o{PACKAGE_DIR}/Resources/English/Alphabet_sort.txt SortTxt {CURRENT_WORK_DIR}/corpus_snt/tags_err l{CURRENT_WORK_DIR}/corpus_snt/tags_err.n o{PACKAGE_DIR}/Resources/English/Alphabet_sort.txt

2.4.1.2 A private dictionary

If the cascades use a private dictionary of forms, for example the *myProperNameDictionary.dic* dictionary, created in <u>Chapter 5, Section 4.2</u>, page 123.

Go to the folder $MyUnitex \setminus English \setminus Dela$ Copy the two files named myProperNameDictionary.bin and myProperNameDictionary. inf.Go to the folder: $MyUnitex \setminus ScriptEntities \setminus ScriptEntitiesLingpkg \setminus Resources \setminus English \setminus Dela$ Paste these files.¹⁴⁰

Just the *Dico* command changes.

Dico -t{CURRENT_WORK_DIR}/corpus.snt -a{PACKAGE_DIR}/Resources/English/Alphabet.txt {PACKAGE_DIR}/Resources/English/Dela/<u>myProperNameDictionary.bin</u>

2.4.1.3 Several dictionaries

If the cascades use both the default English dictionary and the private dictionary, the both dictionaries are placed in the same line.

Dico -t{CURRENT_WORK_DIR}/corpus.snt -a{PACKAGE_DIR}/Resources/English/Alphabet.txt {PACKAGE_DIR}/Resources/English/Dela/<u>dela-en-public.bin</u> {PACKAGE_DIR}/Resources/English/ Dela/<u>myProperNameDictionary.bin</u>

2.4.1.4 A morphological dictionary

It is also possible to allow the use of dictionaries in the morphological mode (see <u>Chapter 5</u>, <u>Section 4.5</u>, page 127). For example, if the private dictionary used in the previous section, *myProperNameDictionary.dic*, is to be declared also as a morphological dictionary, its name must be written a second time, preceded by the *-m* prefix.

Dico -t{CURRENT_WORK_DIR}/corpus.snt -a{PACKAGE_DIR}/Resources/English/Alphabet.txt {PACKAGE_DIR}/Resources/English/Dela/dela-en-public.bin {PACKAGE_DIR}/Resources/English/ Dela/myProperNameDictionary.bin <u>-m{PACKAGE_DIR}/Resources/English/Dela/</u> <u>myProperNameDictionary.bin</u>

In any case, if the user's cascades use dictionaries, the user can always repeat the procedure described in <u>Section 2.2.1</u>, page 150, which consists of applying the cascades through the Unitex interface, copying the console to the editor and adjusting the resulting commands.

¹⁴⁰ Optionaly, you can also copy and paste the *myProperNameDictionary.dic* file and, if it exists, *myProperNameDictionary.txt*, if you want the language package to be also a backup of your work.

2.4.2 The last file from the _csc folder

It is possible to retrieve the last file from the *_csc* folder. This is useful because a bug has crept into the stable version 3.3 of Unitex: the output files are not correct if the cascade annotates categories with graphs, then deletes them with other graphs. This rarely happens, but in order to be sure that we have the correct output, the right file to recover is the last file from the *_csc* folder.

For that:

1. Add, at the beginning of the script, a new DuplicateFile line.

DuplicateFile --make-dir {CURRENT_WORK_DIR}/corpus_csc_csc

2. Modify the output file.

DuplicateFile -i {CURRENT_WORK_DIR}/corpus_csc_csc/corpus_0_0.snt {OUTPUT_FILE_1}

Since intermediate files are not kept, the number of the last file processed is always 0_0.

3 The command line

3.1 The batch launching file

We will use the *File Edition/New File* menu to create a file that will contain the following command line (it has to be adapted for each computer, depending on where the Unitex software is installed):¹⁴¹

"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" { BatchRunScript -i .\Input -e -o .\Output -t1 .\scriptEntitiesLingpkg.zip -f -s Scripts\scriptEntities.uniscript }

We will save this file, naming it *multiLaunchEntities.bat*, in the folder:

*MyUnitex**Scripts**ScriptEntities*

3.2 Running the script

We will now open the command prompt window and choose as the working folder:

MyUnitex\Scripts\ScriptEntities

And then launch the *multiLaunchEntities.bat* batch file.



As a result, we will obtain two files (*jubileePlainText.result.txt* and *JubileePlainText.xml.result.txt*) in the folder:

MyUnitex\Scripts\ScriptEntities\Output

¹⁴¹ A space is obligatory after the opening brace and before the closing brace. The t1 option means that you are working with one core. If your computer has several and your corpus is very large, you can work by parsing the texts of the corpus in parallel, by modifying this option: t2, t3...

3.3 A command line with comments and several possible outputs

The command line above has two drawbacks: errors are not reported and only one file is possible as output. To overcome this, it is possible to launch the files one by one with the command line below which can be saved under the name *launchEntities.bat* in the folder:¹⁴²

MvUnitex	\Scripts	\Script	Entities
my onnor	wei ipus	weipt	LIIUIUICE

"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe"
{ SelectOutput --output=on }¹⁴³ { InstallLingResourcePackage
-p C:\Documents\MyUnitex\Scripts\ScriptEntities\scriptEntitiesLingpkg.zip
-x \$:UnitexPkgResource -v } { RunScript -v
-a INPUT_FILE_1=C:\Documents\MyUnitex\Scripts\ScriptEntities\Input\jubileePlainText.txt
-a CORPUS_WORK_DIR=\$:UnitexPkgWork -a PACKAGE_DIR=\$:UnitexPkgResource
-a OUTPUT_FILE_1=
C:\Documents\MyUnitex\Scripts\ScriptEntities\Output\jubileePlainText.result1.txt
-a OUTPUT_FILE_2=
C:\Documents\MyUnitex\Scripts\ScriptEntities\Output\jubileePlainText.result2.txt
\$:UnitexPkgResource\Scripts\ScriptEntities.uniscript } { InstallLingResourcePackage
-p C:\Documents\MyUnitex\Scripts\ScriptEntities\scriptEntities\Lingpkg.zip
-x \$:UnitexPkgResource -u -v }

If you want to get a single output file, but without comments, you need to remove the -v option. It can also be used to have comments with a single output file by removing the -a OUTPUT_FILE_2 command. Finally, we can also add the commands -a OUTPUT_FILE_3, and so on, to have more than two output files.

But the use of this script is not mandatory.

4 An inventory of tag occurrences

In a cascade that processes XML-like input text, the standoff option launches the creation of an inventory of tag occurrences. For example, a standoff cascade can be used after the annotation with two cascades (analysis and synthesis). These standoff cascades are not integrated in the interface version of Unitex and can therefore be launched only by a script. To illustrate their use, we will create a new script by modifying the previously developed script.

The data used for this step is a cascade of graphs (which may possibly consist of only one graph) and a template file for formatting the results.

In order not to lose the results of the previous section (the annotated files), we will create a new folder for the results of the standoff cascade: *StandoffOutput* in the *Scripts* folder:

 $MyUnitex \Scripts \ScriptEntities$

4.1 The standoff cascade

The *standoff* cascade graphs indicate which tags we would like to count. For these graphs we will create a *Standoff* folder in the folder:

¹⁴² Do not forget to adapt it for your computer, e.g. the paths to Unitex and to your personal Unitex folder.

¹⁴³ This script can also be used without comment with the { *SelectOutput --output=off* } option.

If we want to process the same examples where we have inserted the tags <*AbsoluteDate></AbsoluteDate>, <RelativeDate></RelativeDate>, <Town></Town> and <<i>Address></Address>, ^144* we need four graphs that we can combine into one, with four subgraphs, which we will name *grfStandoff.grf.*¹⁴⁵ The graph and its subgraphs will be recorded in the folder:

*MyUnitex**English**Graphs**UnitexGettingStarted**Entities**Standoff*



Each subgraph recognizes the start tag, the end tag and the content between them, and at the same time prevents nesting of tags of the same type.



We will now create a cascade that consists of only one graph (*grfStandoff.fst2* in *Merge* mode). We will record it using the name *standoff.csc* in the folder:

#	Disabled	Name	Merge	Replace	Until Fix Po	Generaliz
1		grfStandoff.fst2	2			

4.2 The template file

The template file sets the design of the report that inventories the tag occurrences. It is divided into three parts, the second part being itself divided into two.

¹⁴⁴ <u>Chapter 4, Section 1</u>, page 71 and <u>Section 2</u>, page 84.

¹⁴⁵ This grouping makes it possible to process the text only once, which saves time. There may be a bias in case of nesting.

Introduction Inventory (#LINE... #REST): to be repeated for each type and type+subtype combination Inventoried item: {TYPE} or {TYPE} << and {SUBTYPE}>>¹⁴⁶ Inventoried term: {TERM} and its number of occurrences in the document: {COUNT} Conclusion

We will use the *File Edition* menu to create a new file with the name *standoffPattern.txt* in the folder:¹⁴⁷

 $MyUnitex \ Script Entities \ Script Entities \ Lingpkg \ Resources \ English$

```
<rml>
#LINE
(listAnnotation type="{TYPE}" <<subtype="{SUBTYPE}">>>
#BLOCK
<term frequency="{COUNT}">{TERM}</term>
#END
</listAnnotation>
#REST
</xml>
```

4.3 The script

We will use the final version of the script we developed in <u>Section 2.2.4</u>, page 154. We will rename this file under the name standoffScriptEntities.uniscript in the folder:

MyUnitex\Scripts\ScriptEntities\scriptEntitiesLingpkg\Scripts

We will add in it a line before Normalize to indicate the location of the template file:

RunScript {PACKAGE_DIR}/Resources/English/standoffPattern.txt -R -o {CURRENT_WORK_DIR}/patternResult.txt --no_translate_separator

This line uses a specific Unitex function *RunScript*, which loads the *standoffPattern.txt* file under the name *patternResult.txt*. The *no_translate_separator* option prohibits replacement of slashes by backslashes.

We will copy the last three lines *DuplicateFile*, *Tokenize* and *Cassys*; we will paste them right after. We will modify them: the working file is now *corpus_csc_csc.snt*; the *standoff* option replaces the *input_offsets* option; and the name of cascade to be used is *standoff.csc*. The result of the standoff cascade, the *corpus_csc_csc_standoff.txt* file, replace the previous output in the ultimate line of the new script.

DuplicateFile -i {CURRENT_WORK_DIR}/corpus_csc_csc.txt {CURRENT_WORK_DIR}/corpus_csc_csc.snt Tokenize {CURRENT_WORK_DIR}/corpus_csc_csc.snt -a{PACKAGE_DIR}/Resources/English/Alphabet.txt -qutf8-no-bom Cassys -a{PACKAGE_DIR}/Resources/English/Alphabet.txt -t{CURRENT_WORK_DIR}/corpus_csc_csc.snt -I{PACKAGE_DIR}/Resources/English/CasSys/standoff.csc -v -r{PACKAGE_DIR}/Resources/English/Graphs/ --standoff={CURRENT_WORK_DIR}/corpus_csc_csc_standoff.txt OUTPUT_FILE_1}

¹⁴⁶ The text between double angle brackets will only be written in the presence of a subtype.

¹⁴⁷ This example of template file is in XML format, which is not the obligatory format.

4.4 The new linguistic package

We will now copy graphs in the linguistic package.

Go to the folder:MyUnitex\English\Graphs\UnitexGettingStarted\EntitiesCopy the Standoff folder.Go to the folder:MyUnitex\Scripts\ScriptEntities\ScriptEntitiesLingpkg\Resources\English\Graphs\UnitexGettingStarted\EntitiesPaste this folder.

As well as the new cascade.

Go to the folder: $MyUnitex \English \Cassys \UnitexGettingStarted \XmlTextEntities$ Copy the standoff.csc file. Go to the folder: $MyUnitex \ScriptEntities \ScriptEntitiesLingpkg \Resources \English \CasSys$ Paste this file.

We will create the *scriptEntitiesLingpkg.zip* file in the same way as in <u>Section 2.3</u>, page 155.

4.5 The new batch file

We will save the *multiLaunchEntities.bat* file under the name *multiLaunchStandoffEntities.bat*. We need to change in this file the name of the script and name of the folder for results (bold and underlined):

"C:\Program Files (x86)\Unitex-GramLab\App\UnitexToolLogger.exe" { BatchRunScript -i .\Input -e -o .\<u>StandoffOutput</u> -t1 .\scriptEntitiesLingpkg.zip -f -s Scripts\<u>standoffScriptEntities.uniscript</u> }

As a result of running this script, we will obtain two identical files (*jubileePlainText.result.txt* and *JubileePlainText.xml.result.txt*) in the folder:¹⁴⁸

MyUnitex\Scripts\ScriptEntities\StandoffOutput

Note that with the command line with several possible outputs in <u>Section 3.3</u>, page 158, you could merge both the script of <u>Section 2.2.4</u>, page 154, and the script of <u>Section 4.3</u>, page 160, to output the tagged file and the inventory of tag occurrences file.

¹⁴⁸ These two files are identical since the entities are the same in both.

<xml></xml>
stAnnotation type="AbsoluteDate">
<term frequency="1">June 2, 2022</term>
<term frequency="1">February 6, 1952</term>
stAnnotation type="Address">
<term frequency="1">Queen Elizabeth Street</term>
<term frequency="1">London Street</term>
stAnnotation type="RelativeDate">
<term frequency="1">Thursday June 2</term>
<term frequency="1">June 5</term>
stAnnotation type="Town">
<term frequency="2">Nearlondon</term>
<term frequency="1">London</term>

4.6 Some additional information

It is possible to add information to the *standoff* file by using specific variables. We can retrieve the name of the file and, if it is in XML format, also information from its header.¹⁴⁹

4.6.1 The name of the input file

The file name can be included in the template file. To do this, we must add the *{FILENAME}* variable to the *standoffPattern.txt* file, which is in the folder:

 $MyUnitex \ Script Entities \ Script \ Scrip \ Script \ Script \ Script \ Script \ Script \ Script \$

```
<<u><</u>

<u><sourceFile></u>

<u><fileName>{FILENAME}</fileName></u>

<u></sourceFile></u>

#LINE

listAnnotation type="{TYPE}" <<subtype="{SUBTYPE}">>>

#BLOCK

<term frequency="{COUNT}">{TERM}</term>

#END

</listAnnotation>

#REST

</xml>
```

Besides this, we must add the option -*a* FILENAME={INPUT_FILE_1} to the RunScript line of the file:

 $MyUnitex \ Script Entities \ Script Entities \ Lingpkg \ Scripts \ standoff \ Script Entities. uniscript$

¹⁴⁹ It is also possible to retrieve the Unitex version and its data by using a specific Unitex function *VersionInfo*, but we will not go into details here.

RunScript {PACKAGE_DIR}/Resources/English/standoffPattern.txt -R -o {CURRENT_WORK_DIR}/patternResult.txt <u>-a FILENAME={INPUT_FILE_1}</u> --no_translate_separator

After repeating the compression explained in <u>Section 2.3</u>, page 155, and running the script, we again get two identical files. These files contain at the beginning the name of the processed file.

4.6.2 Extracting information from the XML header

When the source file is in XML format, some information located before the body of the text can be interesting to include in the standoff file.

For instance, if we want to retrieve the content inside tags <*title*></*title*> of the *jubileeXmlText.xml.txt* file, we must:

1. Create a file containing the path of the XML element we want to retrieve.¹⁵⁰ This file, which we will name, for instance *pathTitle.txt*, must be recorded in the folder:

Unitex\Scripts\ScriptEntities\ScriptEntitiesLingpkg\Resources\English

/xml/title

2. Add to the script just before the *RunScript* line a new line that invokes the specific Unitex function *Unxmlize* that copies the content of a tagged element:

Unxmlize -x --selxpath_file={PACKAGE_DIR}/Resources/English/pathTitle.txt -o {CURRENT_WORK_DIR}/title.txt -n IGNORE {CURRENT_WORK_DIR}/corpus.txt

3. Also add the option -c TITLE={CURRENT_WORK_DIR}/title.txt to the RunScript line:

RunScript {PACKAGE_DIR}/Resources/English/standoffPattern.txt -R -o {CURRENT_WORK_DIR}/patternResult.txt -a FILENAME={INPUT_FILE_1} <u>-c</u> <u>TITLE={CURRENT_WORK_DIR}/title.txt</u> --no_translate_separator

4. Add the variable {*TITLE*} in the *standoffPattern.txt* file:

```
<ml>
<ml>
<ml>
<sourceFile>
<ml>
<fileName>{FILENAME}</fileName>
<ml>
<ml>
<ml>
</ml>
</ml>
</ml>
</ml>
</ml>
</ml>
```

¹⁵⁰ The path structure follows the XPath syntax: this path indicates how XML tags are nested in the document. Since the *<date>* tag is at the same level as the *<title>* tag, the path to this tag will be /xml/date and not /xml/title/date. If the information to be retrieved is in a tag with a specific value of an attribute, for instance *<title type="newspaper">*, the path to it becomes: */xml/title[@type="newspaper"]*.

After repeating the compression process and launching the script, we get two different files, one with empty <*title*>{*TITLE*}</*title*> tags and the other with the extracted text, enclosed in <*title*></*title*> tags. For the *jubileeXmlText.xml* file:

<xml></xml>
<sourcefile></sourcefile>
<filename>.\Input\jubileeXmlText.xml</filename>
<title>The municipal newspaper of the city of Nearlondon</title>
<listannotation type="AbsoluteDate"></listannotation>
<term frequency="1">June 2, 2022</term>
<term frequency="1">February 6, 1952</term>
<listannotation type="Address"></listannotation>
<term frequency="1">Queen Elizabeth Street</term>
<term frequency="1">London Street</term>
<listannotation type="RelativeDate"></listannotation>
<term frequency="1">Thursday June 2</term>
<term frequency="1">June 5</term>
<listannotation type="Town"></listannotation>
<term frequency="2">Nearlondon</term>
<term frequency="1">London</term>

5 Two small remarks on the optimization of a cascade

Let us say we want to launch our cascade over a very large number of texts, all placed in the same directory. Cascade time, even with multiple cores, can become an important parameter.

5.1 Number of graphs

Each time a graph of a cascade is applied, the whole text is reread. If two graphs are not ambiguous or if the first does not prepare data to be processed by the second, it is therefore better to merge them into a single graph which calls two subgraphs.

For instance, the five graphs of the analysis cascade of <u>Chapter 4, Section 1.1</u>, page 71, cannot be merged.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz
1		absoluteDate.fst2	V			
2		relativeDate.fst2	V			
3		town.fst2	V			
4		townGeneralization.fst2	V			2
5		street.fst2	V			

The two *absoluteDate.grf* and *relativeDate.grf* graphs are ambiguous and they must be applied in this order; the *town.grf* graph uses the previous two; the *townGeneralization.grf* graph uses the previous one; and the *street.grf* graph uses all previous graphs.

5.2 Number of boxes

It takes a little longer to read two boxes of a graph than one, It is therefore better to merge them, if it is possible.

For instance, the *month.grf* graph below:



could be replaced by the *optimizedMonth.grf* graph:



Note that such adjustments reducing the number of graphs or boxes for optimization may hinder the development of the entire work. It is therefore recommended to do it near the end of the elaboration, only if the number of files to process is large.

Contents

Cha	pter 1: i	ntroduction	1
1	A short	history	1
2	A quick	overview of the book	1
3	Organiz	ation of correct files	2
4	A workt	book?	3
	4.1 Gr	aph exercises	3
	Exerci	se 1	3
	Exerci	se 2	3
	Exerci	se 3	3
	Exerci	se 4	3
	4.2 Ca	scade exercises	3
	Exerci	se 5	3
	Exerci	se 6	4
	Exerci	se 7	4
	4.3 Dio	ctionary exercises	4
	Exerci	se 8	4
	Exerci	se 9	5
	Exerci	se 10	5
	4.4 Lex	xicon grammar exercise	6
	Exerci	se 11	6
	4.5 XA	lign exercise	6
	Exerci	se 12	6
	4.6 Sci	ript exercises	6
	Exerci	se 13	6
	Exerci	se 14	6
5	Acknow	ledgement	6
Cha	nter 2 [.] f	irst stens	7
4			
T	Discove	ring Unitex	/
	1.1 Un	litex installation	/
	1.1.1	Creating a private working folder	/
	1.1.2	Installation process	1
	1.1.3	Initial choices	10
	1.1.4	ing Lipitov	11
	1.2 US	Ctarting Unitex	11
	1.2.1	Starting Unitex	11
	1.2.2	Opening a text	12 12
	1.2.5	Applying dictionaries	CL
	1.2.4	Applying dictionaries	
	1.2.5	The DLF window	15
	1.2.0	The DLC window	10
r	1.2./ A dotail	THE ERR WITHOW	10
Z		eu example	10
	2.1 CO	ncordances with a lomma	10
	2.2 CO	neordances with a graph quary	18 10
	2.3 (0	A first such	19
	2.3.1	A IIIst graph.	
	2.3.2	Adding the future tence and the conditional	
	2.3.3	המעוווב נווב וענעו ב נכווזב מווע נווב נטוועונוטוומו	

2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	680002245555667770112
2 3 3 3 3 3 3 3 3 3 3 3 3 3	8000224555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	0002245555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	002245555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	02245555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	2245555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	245555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	45555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4	5555667770112
3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4	555667770112
3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4	55667770112
3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4	5667770112
3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4	667770112
3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4	6 7 7 7 0 1 2 7
3 3 3 4 4 4 4 4 4 4 4 4 4	57 57 57 57 57 57 57 57 57 57 57 57 57 5
3 3 4 4 4 4 4 4 4 4 4 4 4	.7 .7 .7 .0 .1 .1
3 	.7 .0 .1 .1
4 	.0 1 1
4 4 4 4 4 4	1 1 1
4 	1 1
	2
4	
4	. 2
	5
4	
	.5
	.5
	./
	.8
	.9
	0
	0
5	1
5	,2
5	,2
5	,2
5	,3
5	,4
5	,4
5	,4
5	,5
5	6
5	7
5	7
5	7
5	8
5	9
	9
5	9
5	9
C	0
	1
	51
	1
	51 52
	51 52 52
6 	51 52 52 52
	5

Contents

	3.1 First step: from 2 to 9	63
	3.2 Second step: from 2 to 99	64
	3.2.1 The subgraph from 1 to 9	64
	3.2.2 The graph from 2 to 99	
	3.3 Third step: from 2 to 999	65
	3.3.1 The subgraph from 1 to 99	
	3.3.2 The graph from 2 to 999	
	3.4 Fourth step: from 2 to 3 999	
	3 4 1 The subgraph from 1 to 999	67
	3.4.2 The graph from 2 to 3.999	67
	3.5 Ambiguities of Roman numerals (only for confident users)	68
	3.5.1 The subgraph of ambiguous Roman numerals	69
	3.5.2 The use of weights	69
Cha	pter 4: CasSys	71
1	First example: named entity recognition in a plain text	71
1	1.1 Analysis cascado	
	1.1 Analysis cascade	
	1.1.1 The absoluce die.git graph	72
	1.1.2 The relative Date grid diapit	5 / دح
	1.1.3 The analysis.csc cascade	
	1.1.4 The town.gri graph	
	1.1.5 Generalization graphs	
	1.1.6 The street.grf graph	
	1.1.7 Created files	
	1.1.7.1 The csc files	
	1.1.7.2 The _csc folder	78
	1.2 Synthesis cascade	79
	1.2.1 The tag.grf graph	79
	1.2.1.1 A first attempt	80
	1.2.1.2 A second attempt	
	1.2.1.3 The Until fix point option	
	1.2.2 The internalDeletion.grf graph	
2	Second example: named entity recognition in an XML text	
	2.1 Analysis cascade	
	2.1.1 The toolXml.grf graph	
	2.1.2 The toolHidden.grf graph	
	2.2 Synthesis cascade	
3	Third example: measure recognition	
	3.1 Analysis cascade	
	3.1.1 The number.grf graph	
	3.1.2 Cascade	
	3.1.3 The measure.grf graph	
	3.2 Synthesis cascade	
	3.2.1 The tag.grf graph	
	3.2.2 Cascade	
	3 2 3 The numberDeletion grf granh	93
	3.2.4 The number AndNumber of graph	93
4	Additional possibilities (only for confident users)	92
-	4.1 How does a generalization graph work?	Ω <i>I</i>
	4.2 Testing the variables	
	4.3 Reuse the same graph	+ر ۹۶
	4 3 1 Output variable	ور عم
	4.3.2 Granh renositony	
	4.3.2 Graph repusitory	رو ۵۵
	1.4 Think of determining part of a text	00
	7.7.1 Thing part of a text	
	7.7.2 Determing part of a text	101
	T.T.2.1 IIIC UCICUITS SI APIIS	

	4.4.2.2 Comparing the variables	102
	4.5 Sub-categorization with cascades	102
	4.5.1 Three modified graphs	
	4.5.2 Generalization graphs with restrictions	104
	4.5.2.1 Without substitution of a category	104
	4.5.2.2 With substitution of a category	105
Chap	pter 5: dictionary creation	107
1	Introduction	
2	Inflection of monolexical words	
	2.1 Inflection by simple suffixation	
	2.1.1 The nouns with s in the plural	107
	2.1.2 Two more examples	110
	2.2 The <i>L</i> operator	113
	2.2.1 The nouns with final y	113
	2.2.2 Several more examples	114
	2.3 Operators R, C and D	116
3	Inflection of multi-word units	118
	3.1 Example of the MWU air of mystery	119
	3.2 Some other examples	120
4	Some additional remarks	123
	4.1 Adding features in dictionary entries	123
	4.2 Direct creation of inflected entries	123
	4.3 Adding comments	125
	4.3.1 Adding comments to entries	125
	4.3.2 Documenting dictionaries	126
	4.4 Automatically verifying the format of a dictionary	126
	4.4.1 Inflected dictionary	126
	4.4.2 Lemma dictionary	
	4.5 Morphological-mode dictionary (only for confident users)	
	4.6 Priorities	
_	4.7 Additional notes	
5	Dictionary graphs	
	5.1 Example of Roman Numeral dictionary	
	5.2 Morphological extension of the used dictionaries (only for confident users)	
	5.2.1 Morphological-mode dictionary declaration	
~	5.2.2 Example of the -less suffix	
6	Additional possibilities (only for confident users)	
	6.1 A test on a suffix	
	6.2 Use of variables	
	6.2.1 Example of goose, tooth, toot	
	6.2.2 Example of bring, light, think	134
Chap	pter 6: other tools	137
1	Lexicon-grammar	137
	1.1 Example: particle verbs	137
	1.2 Transformation of the spreadsheet table	137
	1.3 Parameterized graphs	138
	1.4 Generation of graphs	138
	1.5 Annotations	140
2	XAlign	141
	2.1 Installation	141
	2.2 The <i>French</i> folder	141
	2.3 File preparation	142
	2.3.1 English folder	142
	2.3.2 French folder	143
	2.3.3 XAlign folder	143

Contents

2.4 Automatic alignment	1/12
2.5 Reading and manual correction	1/15
2.5 Rearch performed on one of two texts	1/6
2.7 Cascade search and XAlign (only for confident users)	140
2.7 Cusedue sedien and Arligh (only for confident users)	147
2.7.2 File transformation	148
2.7.2 The datatornation	148
Chapter 7: scripts	149
1 Introduction	
2 Creation of the <i>linguistic package</i>	
2.1 The <i>Resources</i> folder	
2.2 The <i>scripts</i> folder	
2.2.1 Using the console	
2.2.2 File adjustment	
2.2.3 Script generalization	
2.2.4 Final touches	
2.3 Compression	
2.4 Another possibility (only for confident users)	
2.4.1 The use of dictionaries	
2.4.1.1 The default English dictionary	
2.4.1.2 A private dictionary	
2.4.1.3 Several dictionaries	156
2.4.1.4 A morphological dictionary	156
2.4.2 The last file from the _csc folder	157
3 The command line	157
3.1 The batch launching file	157
3.2 Running the script	157
3.3 A command line with comments and several possible outputs	158
4 An inventory of tag occurrences	158
4.1 The standoff cascade	158
4.2 The template file	159
4.3 The script	160
4.4 The new linguistic package	161
4.5 The new batch file	161
4.6 Some additional information	162
4.6.1 The name of the input file	
4.6.2 Extracting information from the XML header	
5 Two small remarks on the optimization of a cascade	164
5.1 Number of graphs	164
5.2 Number of boxes	164

Index

Alphabet.txt	41
Build concordance	17
CODE.SEM	127
Copy operator	117
correct files	2
DELA/Compress into FST	109
DELA/Inflect	109
Dela/Lookup	89
dela-en-public.bin	14
Delete operator	117
DIC	
dictionary graph	128
DLC	15
DLF	15
Dominique Perrin	1
EQUAL	102
ERR	15
File Edition	35
FIRST	34
FSA	1
FSGraph/New	20
generalization graph	75
graph repository	
INFLECTED	127
input variable	80
installation	7
Intex	1
inventory of tag occurrences	158
LADL	1
left context	30
Left operator	113
LEMMA	130
lexical mask	
lexicon-grammar	1, 137
linguistic package	149
Located sequences	17
LOWER	
Maurice Gross	1
Max Silberztein	1
Menus	12
Denis Maurel

is emeritus professor at the University of Tours, member of Lifat (Laboratory of Fundamental and Applied Computer Science of Tours) and associate member of the LLL (Laboratory of Ligerian Linguistics).

He defended his thesis in 1989 under the supervision of Professor Maurice Gross.

His research focuses on NLP (Natural Language Processing), mainly using Unitex, as well as the creation of linguistic resources: Proper Names Processing (*Prolex* project), Named Entity Recognition (*CasEN* project), scientific documents mining (*Istex* and *Abliss* projects).

He taught Unitex in linguistics and digital humanity masters courses, as well as in computer science engineering schools. He gave several tutorials to doctoral students and researchers.



Cvetana Krstev

served as a professor at the University of Belgrade, Faculty of Philology from 1998 to 2020. She is a co-founder and current president of the Language Resources and Technologies Society *JeRTeh*.

Her research focuses on the development of NLP resources and tools. She developed e-dictionaries of Serbian as well as numerous tools for Unitex. These tools, which target Serbian texts, include the named entity recognition system, correction of OCR errors, restoration of diacritics, extraction of definitions, and many more.

She gave tutorials on using Unitex to researchers and PhD students studying Linguistics, Philology and Information Sciences.

