# TreeCloud \& Unitex: an increased synergy 

## Claude Martineau

## To cite this version:

Claude Martineau. TreeCloud \& Unitex: an increased synergy. ECLAVIT Workshop, Nov 2017, Champs-sur-Marne, France. 2017. hal-01702091

## HAL Id: hal-01702091

## https://hal.science/hal-01702091

Submitted on 19 Apr 2019

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.

TreeCloud \& Unitex: an increased synergy

## Claude Martineau

## Projet PEPS CNRS/UPE Eclavit

## TreeCloud is a tree cloud visualization of a text


-The stopword list contains grammatical words and auxiliary or modal verbs.
For example, in Englis
The stopwords are usually the most common words in a language.
tree, all these words must te firstly removed from the input text.

- Whenever a sliding window of analysis is s sed, its size (i.e. a number of words) must be
given as parameter.

Co-occurrence and distance calculatior
Given two words $A$ and $B$ and

- $0_{11}$ observed number of fliding windows contaning both $A$ and $B$
$-0_{12}$ observed umber of s sliding windows contanining $A$ but not $B$
$-0_{2}$,

the following variables are defined:
 , ner number of sliding wwindows not contaiting $A$

- $\left.E_{22}=\left(R_{2} C_{2}\right)\right)_{1}$, expected nefinumber o sitions of co-occurrence formulas are the following:


Several versions of TreeCloud
Downloadable version in Python
2009: TreeCloud 1.3 for Windows, Linux, Mac developed by Philippe Gambette Use of SplitsTree 4.10 to draw the tree
On-line version in C
2009: $11^{\text {ts }} \mathrm{C}$ version developed by Jean-Charles Bontemps
2012: Iransition to Unicode developed by Claude Martineau
2014: 15 timplementation of Unitex developed by Claude Martinea

## Unitex/GramLab is a corpus analyser and annotation too

Multilingual: Up to 22 languages (French, English,..., Greek, ..., Korean, Thai) - Unicode 3.0 (UTF8, UTF16LE, UTF16BE)

Cross-platform: Linux, macOS, Windows

- Open source: https://github.com/UnitexGramLab
- Website and binary installers: http://unitexgramlab.org
- Under development since 2001 by a group of passionate volunteers Unitex/GramLab uses linguistic resources:
- DELA (LADL electronic dictionaries)

A typical DELA entry is composed by a simple or compound inflected form, followed by lemma and grammatical information. Each entry can be associated with syntactic and
semantic a trimututes sand inflection rules: semantic a atributes and inflection rules:
Example: Given the French simple word "avocat" (lawyer) and the compound word "avocat
d'affares" "(business lawyer) a DELA repersen daffaires (business lawyer), a DELA representation would be:

business lawyer

## 1 inflected form

2, canonical form
3. grammatical catit
$4+$ sementictatctributes
5 inflectional information $\qquad$
Syntactic or semantic rules called «local grammars» represented by graphs

- Graphical representations of local grammars are composed by a set of linked boxes.

A successul path is a path between initial and final states.

Some examples of matched and unmatched sequences
Unitex-Gram Lab is a corpus processing suite [MATCH] suite [MATCH]
Unitex-Grammata is an open surce corpus rocessing Unitex-Gram Lab is a hard to learn corpus processing suite [FALI]
Unitex-GramLab is [FAll]

## An example of analysis

Application of a dictionary; the result is the text diction
w contains two search paths:

a noun $(\langle N\rangle)$ followed by verb in progressive form $\langle\langle b e . V\rangle\langle V: G\rangle)$
lexical mask like $\langle V: K\rangle$ refers to the text dictionary.
The recognized sequence
form of concordances.


Notice that Unitex/GramLab can also produce an annotated text that includes all the tagged patterns.

## Several ways to use Unitex/GramLab

Two interfaces written in JAVA
Unitex IDE (classic)
They refer to
Unitex Core written in $\mathrm{C} / \mathrm{C}++$


Use the API C and JAVA (JNI) that provides access to $\cdot \cdot a$ virtual file system

Command lines or system calls with Perl, Python, etc.

How and Why to plug Unitex into TreeCloud? Construction of the tree with Unitex

1) Unitex transforms the input text into a new text with all the
forbidden/stopwords replaced by the $X X X$ «word »
2) The new text is sent to Treecloud with $x X x$ as the unique
forbidden word (the unique word in the stop list)

Get a larger and more accurate coverage of forbidden words


Insert multiwords into the tree
$\begin{aligned} & \text { If a dictionary contains compound words, these words } \\ & \text { can be kept in the tree but all multiwords cannot be }\end{aligned}$
$\begin{aligned} & \text { can be kept in the tre } \\ & \text { listed in a dictionary. }\end{aligned}$
The adjacent grammar contains a path that recognizes
$\begin{aligned} & \text { person names. The variable SPS contains the name of } \\ & \text { person captured by the subgraph Name_of Person. }\end{aligned}$
$\begin{aligned} & \text { The, (Per) } \\ & \text { version). }\end{aligned}$

Make a strong selection of the words kept in the tre


2017 on-line version of TreeCloud: an improved implementation Introduction of the concept of file processing
In the online version of 2014, there is only a single Unitex processing for each language. The
necessary resources were hard-coded into the program.
In the new 2017 version, there can be several processing

these pairs (language, processing) a concept of processing fie has been set up.
For example in the processing file below, the first ini indicates the path tor French resources,
then three dictionaries (French seneral d dictionary, first name dictionary, toponym dictionary)
have to be applied to the text. In the end, the local grammar is applied in the replace mode.

```
MRE:Trecloud_WS/French/src
MFCHHER=Dela/delal.f.pulic.cbi
M
c
```

Take advantage of the work already done by Unite
Unitex/GramLab analysis steps

| program called | d files | At the end of the Unitex analysis pro text.snt contains a cleaned text |
| :---: | :---: | :---: |
| Normalize | text.snt | ( $\mathrm{normalization} \mathrm{of} \mathrm{separator} \mathrm{characters)}$, |
| Tokenize | kens.txt, text.cod | text.cod contains the list of indexe tokens into the tokens.txt file list. |
| Dico | dlf, dll, err | dlf, dlc, err, respectively contain sim |
| Locate | concord.ind | words, compound |
|  |  | concord.ind contains the matched sequences with their position into the (XXX, and multiword units) |

To get the «new text», we retokenize the text with matched sequences of the concord.ind file as the new tokens of the text. New token.txx and text. cod files are created. This proces
prevents double reading of the text and double division into words.
Thanks to the Unitex API and virtual file system, all this work is done in memory.


IEE PaRIS

Some examples of trees in different languages
$\qquad$


Tree of the 20 most frequent compound and person nouns in an English article


Tree of the 45 most frequent compound and person nouns in a Jules Verne's novel


Tree of the 30 most frequent words sin Serbian article

## Conclusion

Plug-in of Unitex into TreeCloud provides:
A more accurate representation of forbidden words

- All kinds of multiwords to be recognized in the text and presented in the tree
A visual representation of some grammatical or semantic categories of the words.
- A faster construction of the tree (via a careful use of the Unitex API) http://treecloud.univ-m/v.fr

