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-upec-upem.archives-ouvertes.fr/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**

**TreeCloud is a tree cloud visualization of a text**

TreeCloud builds a tree visualization of a text, which looks like a tag cloud where the tags are displayed around a tree to reflect the co-occurrence distance between the words in the text.

The adjacent grammar contains some path/bases:
- The gray box represents a call to a subtag that recognizes auxiliary and modal verbs.
- The second box uses some lexical masks («AVD», «TR», «ST»), to match grammatical category.
- The «ANTIFERGROSS»-basic mask in the last box matches any word. The list of stopwords given in the user.

The .{Pers} label is added to the output (2017 online version).

Some examples of trees in different languages

1. **Tree of the 45 most frequent compound and person nouns in a Jules Verne's novel**
2. **Tree of the 50 most frequent words in a German article**
3. **Tree of the 30 most frequent words in a Serbian article**

**Conclusion**

- **Plug-in of Unitex into TreeCloud provides:**
  - A more accurate representation of forbidden words
  - All kinds of multilinguals to be recognized in the text and presented in the tree
  - A visual presentation of some grammatical or semantic categories of the words.
  - A faster construction of the tree (via a careful use of the Unitex API)

**Get a larger and more accurate coverage of forbidden words**

**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 a lemma and grammatical information. Each entry can be associated with syntactic and semantic attributes and inflection rules.

**Example:**

<table>
<thead>
<tr>
<th>Inflected form</th>
<th>Lemma</th>
<th>Grammatical Information</th>
<th>Syntactic Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>avocat</td>
<td>avocat</td>
<td>noun, masculine</td>
<td></td>
</tr>
<tr>
<td>avocat d'affaires</td>
<td>avocat</td>
<td>noun, masculine</td>
<td></td>
</tr>
</tbody>
</table>

**How and Why to plug Unitex into TreeCloud?**

1. **Unitex transforms the input text into a new text with all the forbidden/stopwords replaced by the «XXX» word +
2. The next text is sent to TreeCloud with «XXX» as the unique forbidden word (the unique word in the stop list)

**Several ways to use Unitex/GramLab**

- **Two interfaces written in JAVA:**
  - Unitex IDE (classic)
  - GramLab IDE (project-based)

When you open the «TreeCloud & Unitex**

**Get the view tree, re-balance the text with matched sequences of the concord file in the note-taking of the text.**

**New Unitex & TextCloud files are created.**

- This process 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.

**How to call to SplitsTree**

If a dictionary contains compound words, these words can be kept in the tree but all multirefs cannot be kept in a dictionary.

The adjacent grammar contains a path that recognizes person names. The variable «PS» contains the name of a person captured by the sub-tag name_

Example:

```plaintext```
TreeCloud & Unitex: an increased synergy
Claude Martineau
Projet PEPS CNRS/UPE Eclat
```

**Unitex/GramLab is a corpus analyser and annotation tool**

- **Based on Automata and RTWs with outputs:**
  - Multilingual: Up to 22 languages (French, English,..., Greek, ..., Korean, Thai)
  - Unicode 3.0 (UTF-8, UTF-16LE, UTF-16BE)
  - Cross-platform: Linux, macOS, Windows
  - Open source: https://github.com/Unitex/UnitexGramLab
  - Website and binary installer: http://unitex/gramlab.org

**Under development since 2001 by a group of passionate volunteers**

**Get a larger and more accurate coverage of forbidden words**

**Insert multirefs into the tree**

If a dictionary contains compound words, these words can be kept in the tree but all multirefs cannot be kept in a dictionary.

The adjacent grammar contains a path that recognizes person names. The variable «PS» contains the name of a person captured by the sub-tag name_

Example:

```plaintext```
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 «XXX» word +
2. The next text is sent to TreeCloud with «XXX» as the unique forbidden word (the unique word in the stop list)

**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 version, there can be several processing procedures for each language.

Furthermore, the Serbian language (Latin and Cyrillic) has been added. In order to manage these two languages, a processing file has been set up.

For example, in the processing file below, the first line indicates the path for French resources, and the second one for Serbian. The tree grammar is designed for this goal. A similar process is used for languages with nouns or adjectives with cases (e.g. Serbian or Greek).

Example:

```plaintext```
```
**Tree of the 50 most frequent words in a German article**

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

**Tree of the 30 most frequent words in a Serbian article**

**Conclusion**

- **A more accurate representation of forbidden words**
- **All kinds of multilinguals 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.uni-nvl.fr**

**Some examples of trees in different languages**

1. **Tree of the 45 most frequent compound and person nouns in a Jules Verne's novel**
2. **Tree of the 50 most frequent words in a German article**
3. **Tree of the 30 most frequent words in a Serbian article**