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 & Unitex: an increased synergy
Claude Martineau
Projet PEPS CNRS/UPF Eclatix

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 grammude gives an overview of Barack Obama’s 2008 presidential campaign speeches.

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-16, UTF-32)
- Cross-platform: Linux, macOS, Windows
- Open source: https://github.com/Unitex/gramlab
- Website and binary installers: http://www.treecloud.net/gramlab.org

Under development since 2001 by a group of passionate volunteers

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 e
2) The new text is sent to TreeCloud with XXX as the unique forbidden word (the unique word in the stop list)

Get a larger and more accurate coverage of forbidden words

Under development since 2001 by a group of passionate volunteers

Some examples of trees in different languages

The 45 most frequent compound and person nouns in a Jules Verne’s novel

The 30 most frequent compound words in a Serbian article

The 30 most frequent compound words in a Greek article

Conclusion

Plug-in of Unitex into TreeCloud provides:
- A more accurate representation of forbidden words
- All kinds of multiplicity to be recognized in the input text
- 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).