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 Martinou
Projets PENS CNRS/UEP Eclatv

TreeCloud is a tree cloud visualization of a text

TreeCloud builds a tree cloud 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 three pathways:
- The grey box represents a call to a subgraph that recognizes auxiliaries and modal verbs.
- The second box uses some lexical masks (N(JAVI), +COMO, ...). To match grammatical categories.
- The ANTIPOSSER lexical mask in the last box matches the multiword units (XXX, and multiword units).

A more accurate representation of forbidden words
- All kinds of multiwords to be recognized in the 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).

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 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 tree of the 50 most frequent words in a Greek article

The tree of the 45 most frequent compound and person nouns in an English 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
- 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)