FreeLing: Open-Source Natural Language Processing for R&D

Lluís Padró
Centre de Recerca TALP
Universitat Politècnica de Catalunya
padro@lsi.upc.edu

Introduction

- *What is FreeLing ?*

  A configurable and extensible linguistic analysis library, developer-oriented.

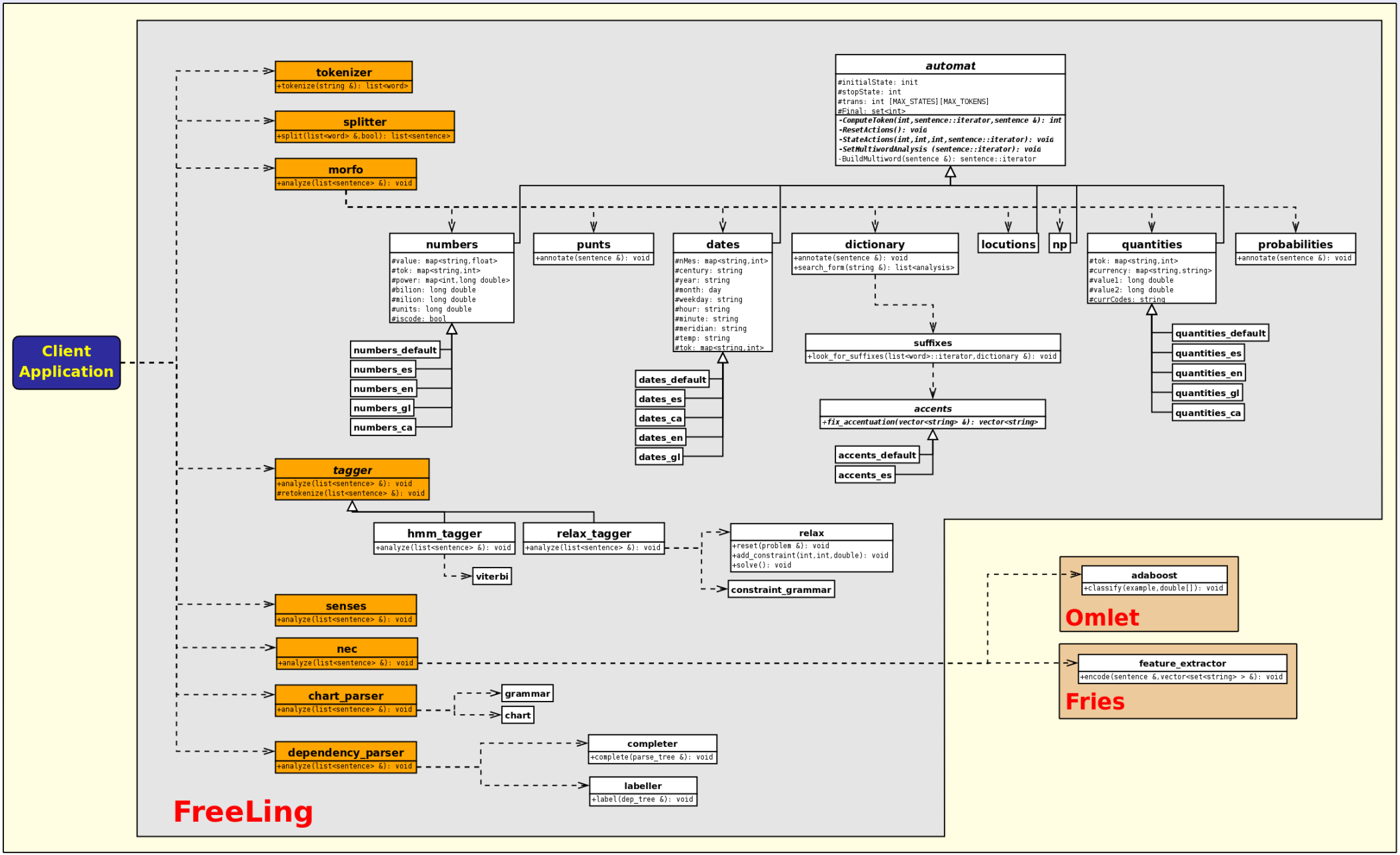
- *What is not FreeLing?*

  A user-oriented off-the-shelf linguistic analyzer.

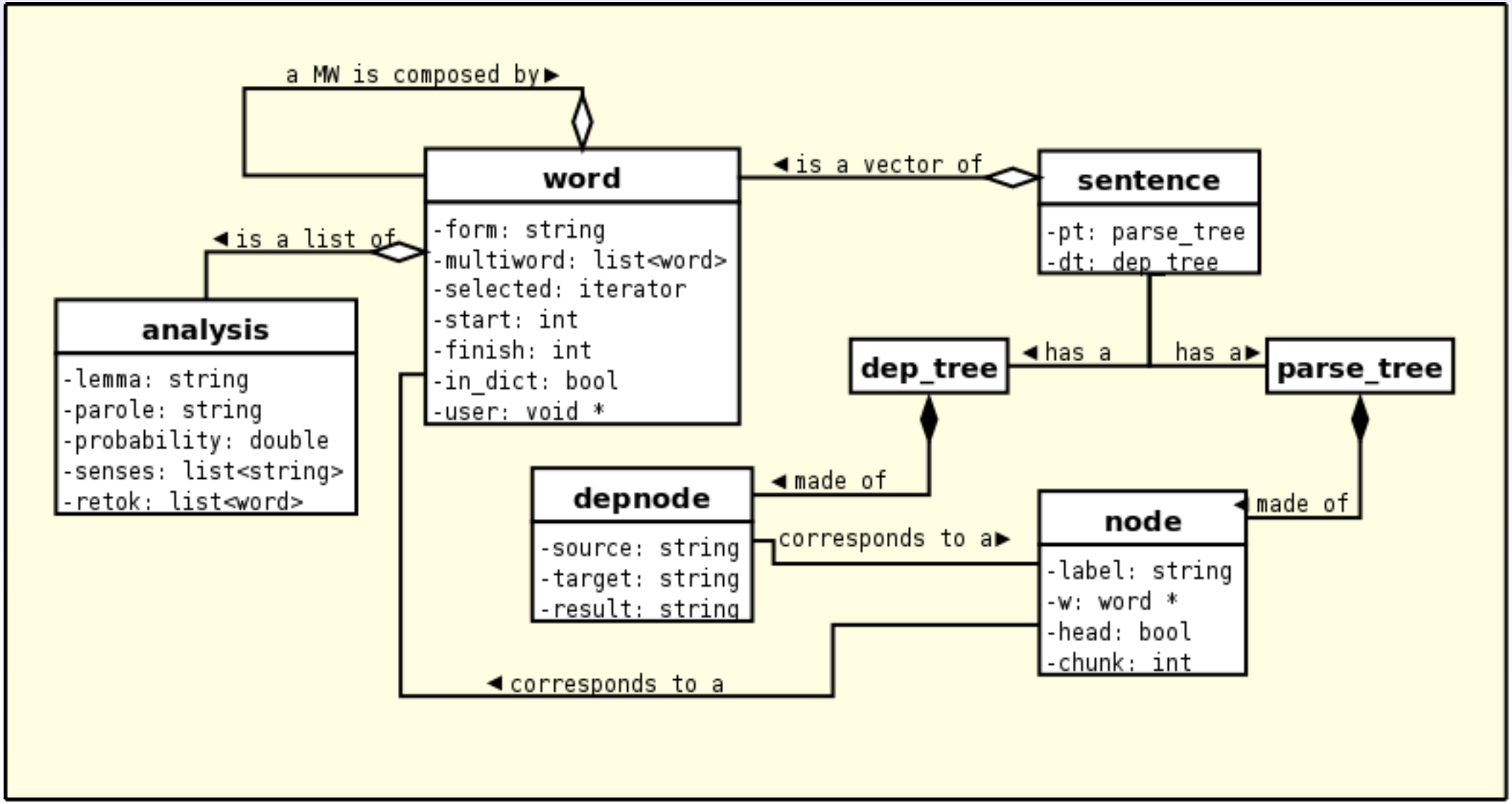
- *What do people use it for?*

  As a user-oriented off-the-shelf linguistic analyzer.

Processing Classes



Linguistic Data Classes



Processing sequence

Main program

Initialization: Create required modules

```
tokenizer tk("tokenizer.dat");
splitter sp("splitter.dat");

maco_options opt("es");
opt.QuantitiesDetection = false;
opt.LocutionsFile="locucions.dat";
opt.SuffixFile="sufixos.dat";
opt.DictionaryFile="dicc.src";
opt.NPdataFile="np.dat";
opt.ProbabilityFile="probabilitats.dat";
opt.PunctuationFile="punct.dat";
maco morfo(opt);

hmm_tagger tagger("es", "tagger.dat", true, 2);
```

Processing sequence

Main program

Read and process text: send each input line through processing chain

```
string text; list<word> lw; list<sentence> ls;

while (getline(cin,text)) {

    lw=tk.tokenize(text);
    ls=sp.split(lw, false);

    morfo.analyze(ls);
    tagger.analyze(ls);

    ProcessAnalyzedSentence(ls)
}
```

Including new languages (1)

- Tokenizer & Splitter:
  - Adapt config files.
- Morphological analyzer:
  - Index form dictionary
  - Adapt suffixation rules
  - Provide (if any) multiwords file
  - ***Develop*** (if needed) date, number, and quantities modules

Including new languages (2)

- Tagger (and probabilities module)
  - Use a tagged corpus to train taggers and compute lexical probabilities. Scripts are provided with FreeLing
- Chart parsers and Dependency parsers
  - Develop appropriate grammars (or adapt some of the existing ones to the new language)

Some NLP applications using FreeLing (1)

- OpenTrad (PROFIT, www.opentrad.org)
  - Spanish & English analysis for es-ba and en-ba syntactic transfer machine translation.
  - Adaptations
    - Improve/develop chunking grammars and dependency parser rules
    - Produce appropriate XML output

Some NLP applications using FreeLing (2)

- ASOMO (Judo Socialware, www.asomo.net)
  - ML-based NER development environment for opinion mining on highly unstructured documents (blogs, forums, etc.)
  - Adaptations:
    - Extend/adapt JAVA API
    - Develop ad-hoc modules to use Omlet&Fries to train NER modules.

Some NLP applications using FreeLing (3)

- VKM (Cromosoma S.A.)
  - CIDEM project to evaluate the viability of using NLP techniques in interactive Videogames. Closed-domain dialogue and QA system.
  - Adaptations:
    - Use semantic dictionary with basic logical forms instead of WN synsets. FreeLing output is processed by a DCG.

Some NLP applications using FreeLing (4)

- T-Incluye (Fundación CTIC, www.tincluye.org)
  - Exclusive language detector
  - Adaptations:
    - Adapt the form dictionary lemma criteria for some words (e.g. *Príncipe-princesa*)
    - Develop an ad-hoc grammar for noun phrases, to pre-filter correct/irrelevant/incorrect phrases.
    - Improve JAVA API for Semantic DB access.

Some NLP applications using FreeLing (5)

- Dixio (Semantix, www.semantix.com/)
  - Embeeded intelligent dictionary
  - Adaptations:
    - Improve client-server operation
    - Develop PHP client.

Other application fields...

- Information Retieval (IR)
- Information Extraction (IE)
- Document management (Text Categorization, Text Clustering, Text Mining, ...)
- Linguistic Research
- Opinion mining
- Dialogue Systems
- etc.

Open Source Benefits

- Used both in academy... :
  - Studies on medieval Spanish evolution
  - CLARIN project
  - Deep parsing (Spanish Resource Grammar)
  - Preprocess to many research applications
- ... and industry:
  - Apertium proper noun recognizer
  - Spell checkers (Galician OpenOffice)
  - Semantic web
  - Legal text treatment

Open Source Benefits

- Visibility:
  - >250 citations
  - ~ 50,000 dowloads since sept'09 (versions 2.1 and 2.2)
- Contributions:
  - Extension up to 8 languages.
  - Porting to other platforms
  - Linguistic data
  - Code (bugfixes, APIs, modules)
  - **Suggestions and bug reports**

Open Source Benefits

- Bussiness
  - Dual License
  - Customization
- Funding
  - R&D projects: EU, Spanish Government.
  - Industry contracts.

Conclusions

- FreeLing is not only an efficient analyzer, but a highly customizable tool.
- It is very helpful in the development of higher level applications or specific-purpose analyzers.
- It is not difficult to set up a basic morpho+PoS tagger kit for a new language.

Conclusions

- 6-year lasting open-source project
- Original goals achieved:
  - Visibility
  - Opportunity creation
  - Widely used
- Partially achieved:
  - Community sustained
- Not achieved yet:
  - “Standard” platform for NLP

FreeLing: Open-Source Natural Language Processing for R&D

Lluís Padró
Centre de Recerca TALP
Universitat Politècnica de Catalunya
padro@lsi.upc.edu