www.inesc-id.pt

technology
from seed

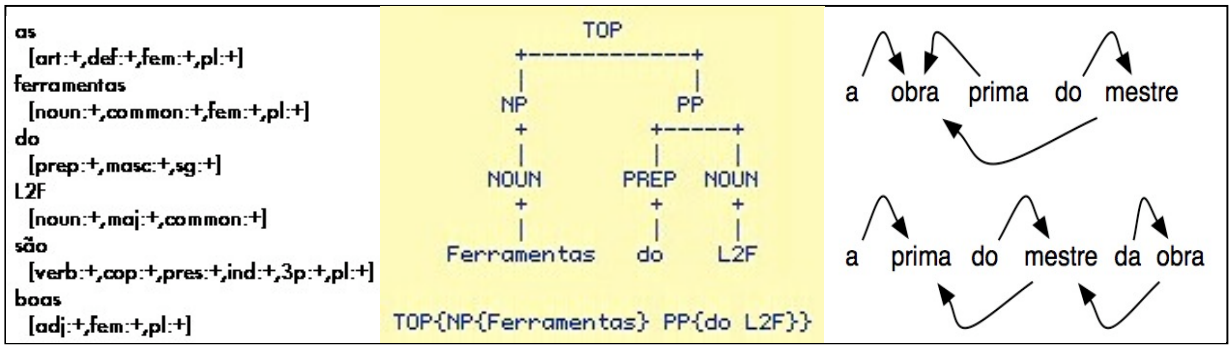
inesc id lisboa

INESC-ID
**Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa**

Rua Alves Redol, 9
1000-029 Lisboa Portugal
Tel. +351.213100300
Fax: +351.213145843
Email: info@inesc-id.pt

## L²F – Spoken Language Systems Laboratory

# Natural Language Processing Tools




**Language processing can take advantage of several tools, such as syntactic and semantic analyzers. In order to perform their tasks, some of these tools use linguistic information (for instance, dictionaries and grammars), making natural language processing by computers closer to the human process.**

We are using natural language processing tools in many of our applications, namely in dialog management, automatic summarization, information retrieval, question answering, discourse analysis, term and emotion extraction. Besides applying these tools to text we are applying them to automatic transcriptions of spoken documents, leading to new challenges.

**Goal.** Build a library of state of the art natural language processing tools, that can be used in many different applications.

**Description.** Natural Language processing tools can be grouped as:

(a) *Morphological Tools* are responsible for the first steps in the processing chain, such as splitting sentences, detecting composed terms (e.g. "*guarda-chuva*", "*Presidente da República*") and classify or disambiguate words (e.g. "*canto*": a noun or a verb? "*a*": a preposition or an article?). Many of these tools take advantage of dictionaries and hand-crafted or data-driven rules;

(b) *Syntactic Tools* parse texts and return sentences organized in phrases (e.g. nominal or verb phrases) or dependency structures relating words. Typically they use grammars that can be inferred or built by linguistic experts;

(c) *Semantic Tools* perform the last step that enable a sentence to be understood by a computer. This step uses the syntactic information previously obtained and the semantic information associated with words and/or groups of words. These tools generate a formal representation (such as a frame or a formula in some logic) for each sentence.

These basic tools can be integrated in complex architectures to achieve more sophisticated tools, such as, for instance, identifying who is talking in children stories.

**More information** is available by email to info@l2f.inesc-id.pt or directly from the website http://www.l2f.inesc-id.pt/.

www.inesc-id.pt

inesc id lisboa

# L²F: Natural Language Processing Tools

## Team

David Matos, PhD
Nuno J. Mamede, PhD
Luísa Coheur, PhD
Jorge Baptista, PhD
Ricardo Ribeiro, PhD Student
Joana Paulo, PhD Student
Fernando Batista, PhD Student
João Graça, PhD Student

## Tools

(Grouping is more or less ad-hoc, but intends to reflect the level of processing for which the tool is useful)

Morphological Tools

*MARv* - a morphossyntactic disambiguation tool;

*monge* - a word form generator;

*PAsMo* - a rule-based morphology processor, tag converter, and sentence splitter;

*RuDriCo* - a rule-based morphology processor;

*SMorph* - a morphological analyser;

*XA* - a morphological analyser similar to ispell and jspell;

*YAH* - yet another hyphenator.

Syntactical Tools

*ParVO* - a C++ implementation of Earley's algorithm with attribute unification;

*SuSAna* - a chunk analyzer;

*TiraTeimas* - verifies if a set of chunks satisfies a set of constraints;

*Algas* - establishes dependency relations between chunks and words.

Syntax/Semantics Interface

*Ogre* - transforms a structure where both chunks and words are connected into a dependency structure;

*AsDeCopas* - applies contextual rules (possibly hierarchically organized) to a graph.

Discourse Analysis

*DID* – identifies indirect and direct speech in children stories. It also attributes a character to each direct speech utterance.

Multi-purpose

*Galinha* - a portal for building and running applications;

*LRDB* - a language resources database and access framework;

*FSTk* - a finite-state transducer library;

*ShReP* - a framework for simplifying the process of constructing NLP systems.

## More Information

By email to info@l2f.inesc-id.pt or
directly from the website http://www.l2f.inesc-id.pt/.