The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French

Benoît Sagot
Alpage, INRIA & Université Paris 7, France
Outline

1. Introduction: the Lefff and the other Alexina lexicons
2. Brief description of the Alexina framework
3. Sources of lexical information in the Lefff
4. Evaluation of the Lefff
5. Future work
I. Introduction: the Lefff and the other Alexina lexicons
Many NLP tasks benefit from rich and large-coverage lexical information.

- morphological information is relevant for POS-tagging.
- syntactic information is relevant for parsing.
- Such lexical information is not always freely available, even for major languages such as French.
The Alexina framework

• Alexina is a framework for modeling and acquiring lexical information at the morphological and syntactic levels (valency…)

• Alexina lexicon for French: the Lefff (Lexique des formes fléchies du français)

• The Lefff is used in various tools:
  • morphological info: POS taggers, lemmatizers…
  • morphological and syntactic info: parsers for various formalisms (LTAG, LFG, IG, Pre-group grammars…)


Several other Alexina lexicons already exist:

- large-scale morphological + syntactic lexicon: Spanish (Leffe, ongoing work)
- large-scale morphological lexicons: Polish, Persian (PerLex), Galician (Leffga), English
- medium- or small-scale morphological lexicons: Slovak, SoraLex (Sorani Kurdish)
- imported (morph.) lexicons (Morph-it, Alpino)
- All Alexina lexicons are freely available (LGPL-LR)
2. Brief description of the Alexina framework
A two-level architecture

- **Intensional level**: inflection class + “initial” sub-categorization frame + list of possible redistributions
  - one entry for each sense of each lemma
  - manually or semi-automatically developed

- **Extensional level**:
  - one entry for each inflected form and each redistribution of each intensional entry
  - generated automatically from intensional entries
  - used in NLP tools
An example

Intensional entry:

clarifier₁ v-er:std Lemma;v;
<Suj : cln | scompl | sinf | sn, Obj : ( cla | scompl | sn )>; %actif,%passif,
%se_moyen_impersonnel,%passif_impersonnel,
%ppp_employé_comme_adj

Extensional entry:

clarifiés v [pred=’clarifier₁
<Suj : cln | scompl | sn, Obl2 : ( par-sn )’],
@passif,@pers,@Kmp]; Kmp %passif
The morphological level

• Each intensional entry is associated with an inflection class

• Inflection classes are defined as follows
  • a list of forms defined by a prefix and a suffix
    + a morphological tag
  • sandhi patterns (e.g., mang_ons ➞ mange_ons)
  • tables and forms may be constrained by regular expressions on the stem
Example

<table name="v-er" canonical_tag="W" rads="...*">
    <form suffix="er" tag="W"/>
    <form suffix="a" tag="J3s"/>
    <form suffix="ai" tag="J1s"/>
    <alt>
        <form suffix="2e" tag="PS13s" rads="..*[td]" var="dbl"/>
        <form suffix="e" tag="PS13s" var="std"/>
    </alt>
    ...
</table>

<sandhi source="et_2e$" target="ett_e$"/>
<sandhi source="[:ou:]y_e$" target="[:ou:]i_e$"/>
The syntactic level

- At the intensional level: initial sub-categorization frame + redistributions (mappings from initial to final sub-categorization frames)

- w.r.t. the lexical rules approach, the difference is that it is a one-shot mapping — whereas lexical rules may be applied sequentially

- Subcategorization frame = for each argument:
  - its syntactic function
  - its possible realizations (syntagmatic + clitic)
Example

<suj : cln | scompl | sinf | sn, obj : ( cla | scompl | sn )>

%passif

%passif = {Only PastParticiple}
+ {Macros @pers} + {Macros @passive}
+ {suj < obj[cla>cln, de-sinf > sinf, seréfl > , seréc >]}
+ {suj )()} + {obl2 (par-sn)}
+ {?{@ctrlsuj.* } + {?{@ctrlobjobjà @ctrlsujobjà}}
+ {?{@ctrlobjobjde @ctrlsujobjde} + {?{@ctrlobj.* } }

<suj : cln | scompl | sn, obl2 : ( par-sn )>

@passif,@pers,@kmp
3. Sources of lexical information in the Lefff
Automatic acquisition techniques

- always followed by *manual validation*
- statistical techniques for *extracting morphological entries from raw corpora* (Clément et al., 2004; Sagot, 2005)
- automatic acquisition of *specific syntactic information* (Sagot, 2006)
Error mining techniques

- *manual correction and extension guided by automatic techniques*

- *simple statistics on tagged corpora* for detecting missing entries (Molinero et al., 2009)

- *error mining in parsing results* for correcting the syntactic information (Sagot and de La Clergerie, 2006)

- *manual mining* of the output of Lefff-based NLP tools (parsers, taggers, tokenizers, spell checkers…).
Comparison and merging with other resources (1/2)

- preliminary *linguistic analysis* of specific phenomena and their *modeling* in one or several other resources

- Lexicon-Grammar tables (Gross, 1975), Dicovalence (van den Eynde & Mertens, 2006), Lexique des Verbes Français (Dubois & Dubois-Charlier, 1997)

- conversion into the Alexina representation

- merging with the Lefff
Comparison and merging with other resources (2/2)

- This approach was applied to various classes of entries and/or phenomena such as:
  - **impersonal constructions** (Sagot and Danlos, 2008),
  - **pronominal constructions** (Danlos and Sagot, 2008)
  - **verbs in -iser and -ifier** (Sagot and Fort, 2009)
  - **several classes of frozen verbal expressions** (Danlos et al., 2006)
  - **adverbs in -ment** (Sagot and Fort, 2007)
4. Evaluation of the Lefff
Quantitative comparison with other resources

Number of unique lemmas per category

<table>
<thead>
<tr>
<th>Category</th>
<th>Lefff</th>
<th>Morphalou</th>
<th>Multext</th>
<th>Dicovalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbs</td>
<td>6,825</td>
<td>8,789</td>
<td>4,782</td>
<td>3,729</td>
</tr>
<tr>
<td>nouns</td>
<td>37,530</td>
<td>59,002*</td>
<td>18,495</td>
<td>0</td>
</tr>
<tr>
<td>adjectives</td>
<td>10,483</td>
<td>22,739</td>
<td>5,934</td>
<td>0</td>
</tr>
<tr>
<td>adverbs</td>
<td>3,584</td>
<td>1,579</td>
<td>1,044</td>
<td>0</td>
</tr>
<tr>
<td>prepositions</td>
<td>225</td>
<td>(51)</td>
<td>117</td>
<td>0</td>
</tr>
</tbody>
</table>
The Lefff for POS tagging

- MElt POS tagger (Denis & Sagot, 2009, 2010)
- MaxEnt-based tagger
- contextual features
- surface features extracted from the words
- possibility to add lexical features
- Using the Lefff as a source of lexical features increases the accuracy from 97.25% up to 97.75% (state-of-the-art)
The Leff for parsing

- **FRMG parser** (LTAG, generated from a metagrammar) (Thomasset & de La Clergerie, 2005; de La Clergerie 2010)
- Based on the Leff (esp. syntactic information)
- **Lexicon-Grammar tables** are considered a highly valuable syntactic resource (Gross 1975)
- These tables were converted in the Alexina format (Tolone & Sagot 2009)
- Evaluation according to the EASy metrics and corpus: 59.9% on “relations” with the Leff vs. 56.6% with the converted Lexicon-Grammar tables
5. Future work
Future work on the Lefff

- Ongoing work on a new version of the verbal part
- Sub-categorization information for predicative nouns and adjectives
- Studying new phenomena for merging lexical information with other resources
- Semantic information, in the form of a mapping with the WOLF (Wordnet Libre du Français) (Sagot & Fišer 2008)
Future work on other Alexina lexicons

- Leffe (Spanish) and Leffga (Galician): the Victoria project (Nicolas et al., 2010)

- PerLex (Persian): the PerGram project (Samvelian & Müller)

- EnLex (English): exploitation of existing syntactic resources

- Resource-scarce languages (SoraLex: Sorani Kurdish, Leffga: Galician): work on developing lexical resources for related languages
All Alexina lexicons are freely available (LGPL-LR):
alexina.gforge.inria.fr
or gforge.inria.fr/projects/alexina/
(use the subversion repository, or the tgz packages)

What you need:
– "alexina-tools": the set of tools for compiling the intensional lexicon into the extensional one
– the lexicon proper
All Alexina lexicons are freely available (LGPL-LR):
alexina.gforge.inria.fr
or gforge.inria.fr/projects/alexina/
(use the subversion repository, or the tgz packages)

What you need:
– “alexina-tools”: the set of tools for compiling the intensional lexicon into the extensional one
– the lexicon proper
– my email, in case of problems:
  benoit.sagot@inria.fr