Learning Morphology of Romance, Germanic, and Slavic languages with the tool *Linguistica*

Helena Blancafort
LREC 2010
Outline

1. Introduction
2. State of the art
3. Linguistica: How it works
4. Experiments and Results
5. Conclusions and further work
Introduction

Motivation
How can we predict the cost of developing a morphosyntactic lexicon for a new language?

Goals
- Evaluate if we can benefit from unsupervised learning of morphology
- Input: Bible parallel corpus, tool Linguistica (Goldsmith 2001, 2006)
State of the Art: Induction of morphology

Objective
- induce morphological information from raw data

Affix inventory
- Brent et al. 1995; Kazakov, 1997
- MDL (Rissanen, 1998)

Cluster of stems and affixes
- Schone and Jurafsky 2001
- Yarowsky and Wicentowski 2001
State of the Art II

Using linguistic knowledge or not

**Lexicon**
- Nakov et al (2003); Oliver (2005)
- Learn all possible endings of an unknown word
- Apply Maximum Likelihood Estimation (Mikheev)

**Inflection Rules**
- Clément et al. (2004)
- Fosbert et al (2006); Loupy et al. (2008)
- **Pos-tagger** Zanchetta and Baroni (2005)
Linguistica: How it works I

• Knowledge-free
• Input: raw corpus
• Heuristics to generate a probabilistic morphological grammar
• MDL (minimum length description) & EM (expectation-maximization algorithm) to filter out inappropriate analysis
Linguistica: How it works II

Signatures
Paradigm-like clusters with words sharing the same affixes
→ could help to build a morphological grammar

The algorithm:
- Splits a word into stem and affix
- For each stem, list of affixes
- Cluster of stems sharing the same affixes
Linguistica: How it works III

Signatures

```
NULL.ed.ing.s 68 7889
gather abound account ascend ask belong boil chasten concern confirm consider delay doubt encamp enter exceed explain fail fasten fold gain gather glean greet groan guard hang happen harden insult journey knock lack leap lift listen look minister number obey offer overflow
```
Main hurdles

1) **Allomorphy**

ES \textit{colgar} -> colg, cuelg
FR \textit{acheter} -> achet, achèt

2) **Incomplete paradigms** due to bad segmentation

Spanish verb \textit{anunciar}:
\textit{anunci}(o, en, etc.), \textit{anunci}ab\textit{a}(a)

3) No distinction between inflectional and derivational suffixes
Experiments and Results I

number of suffixes generated by Linguistica
Experiments and Results II
Number of paradigmes and number of suffixes
Experiments and Results III

Max nb forms per signature (Linguistica)
# Experiments and Results IV

Knowledge-free vs. Knowledge based

| Max nb forms per signature (Linguistica) | Max nb forms per paradigm (Multext) |
|----------------------------------------|----------------------------------|
| es                                     | it                               |
| 31                                     | 63                               |
| it                                     | fr                               |
| 28                                     | 62                               |
| fr                                     | es                               |
| 24                                     | 55                               |
| de                                     | de                               |
| 14                                     | 29                               |
| en                                     | en                               |
| 9                                      | 14                               |
# Experiments and Results V

Longest signatures suggested by Linguistica for a stem

| Affix | Stem | signature |
|-------|------|-----------|
| pl    | 39   | da        |
|       |      | NULL.ch.cie.dzą.j.je.jmy.jmyż.jąc.li.liście.liśmy.m.my.na.ne.nej.ni.nie.niu.no.ny.ną.rze.sz.wa.wał.wszy.ć.t.ła.łby.łbyś.łem.łeś.ło.ły.ń |
| es    | 31   | annunci   |
|       |      | a.ad.ada.adas.adlo.ado.amos.an.ando.ar.arararl.es.aron.aros.arte.ará.arán.arás.aré.as.ase.asen.e.emos.en.es.o.áis.é.éis.ó |
| de    | 14   | heil      |
|       |      | NULL.e.en.et.ig.los.lose.loser.sam.same.sames.t.te.ten |
| en    | 9    | light     |
|       |      | NULL.ed.en.er.ing.ly.ness.ning.s |
Experiments and Results VI
List of most frequent prefixes for German

| Prefix | Nb occ. | Prefix | Nb occ. | Prefix | Nb occ. |
|--------|---------|--------|---------|--------|---------|
| ge     | 40      | her    | 13      | er     | 8       |
| aus    | 30      | un     | 13      | *nied  | 7       |
| ver    | 21      | weg    | 11      | bei    | 6       |
| hin    | 20      | be     | 10      | heim   | 6       |
| auf    | 19      | zu     | 10      | über   | 5       |
| ab     | 19      | *üb    | 9       | durch  | 5       |
| ein    | 16      | an     | 9       | ent    | 4       |
Conclusions and Further Work

- Useful information to evaluate the richness and complexity of the morphology of a language
- Unsupervised techniques should be improved with human input: handwritten-rules are necessary for dealing with allomorphy and correct bad segmentation (Karasimos & Petropoulo 2010)
- Complete paradigms using the web (Oliver 2005) or
- Output quality is language-dependent, English better results than other languages (complete verbal paradigms)
Thank you
Grazzi