The Spanish Resource Grammar
The Tibidabo treebank

Montserrat Marimon
Universitat de Barcelona

Spanish Resource Grammar

Two components (integrated in the logon tree)

1- External component: the FreeLing Tool

2- The LKB Spanish grammar (based on the grammar Matrix)

The Spanish Resource Grammar

The FreeLing Tool (developed at the UPC)

- open-source language analysis tool kit performing shallow processing functionalities.
- Integration of 2 functionalities
  - Morphological analysis (lexical look-up, *gilcUB* dictionary)
  - Named Entities (proper names, percentages, numbers,...) detection and classification-

  - By means of the SPPP protocol and using the inflectional rules to associate the PoS tags delivered by FL to morpho-syntactic features

    aq0ms0 :=

    %suffix (aq0ms0)

    [SYNSEM.LOCAL [CAT .HEAD adj,

    AGR [PN 3sg, GEM masc]] .

The Spanish Resource Grammar

(July-2009)

- 188 phrase structure rules
- lexical coverage
  - 500 types
  - 46,326 words
    - Verbs: 4318 8427
    - Nouns: 27646 27986
    - Adjectives: 10195 10670
    - Adverbs: 3947 4079
    - close classes: 220

  + 65 lexical rules

The Spanish Resource Grammar

Current & future tasks

- To build up a Treebank
- To evaluate the SRG
- To evaluate research work in lexical acquisition (Núria Bel)

The Tibidabo Treebank

- We use the AnCora and the SENSEM corpora, both mainly composed of newspaper texts
- AnCora
  - about 500,000 words (18,000 sentences)
  - used in International Competitions (e.g. CoNLL).
- SENSEM
  - about 700,000 words (25,000 sentences)
  - 100 examples including of the 250 most frequent verbs (M. Davis *A Frequency Dictionary of Spanish*, 2005).

13,170 cases from AnCora...

0-10	10.40%
10-20	22.36%
20-30	23.22%
30-40	21.06%
40-50	14.49%
50-60	5.93%
60-80	2.24%
80-100	0.22%
100-145	0.08%

- First analysis: only about 7.5% covered!!!
  - Gaps in the grammar coverage: syntactic structures & lexicon (entries & readings)
  - Missing words in the FL lexicon
  - Bugs in the SRG and the *integration of the PoS tags*
  - Lexical semantic restrictions on verbal complements

  - syntactic errors in the corpus
  - misspelled words, foreign words (and misspelled foreign words)

Ancora11: 1000 cases, 39.1%

0-10	68.3%
10-20	48.3
20-30	46.8
30-40	12.4
40-50	5.1
50-60	0%
60-80	0%
80-100	0%
100-145	--

- Still gaps in the grammar coverage: syntactic structures & lexicon
- Still missing words in the FL lexicon (added a module to deal with prefixes).
- Still bugs in the SRG and the integration of the PoS tags
- And still syntactic errors in the corpus, misspelled words, foreign words