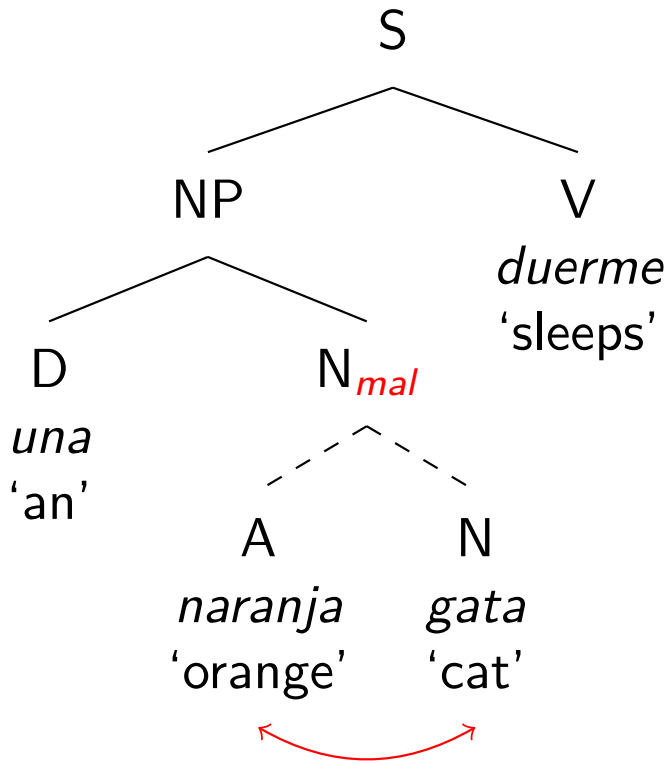
# Spanish Resource Grammar updates

## for the DELPH-IN summit

Olga Zamaraeva, Lorena Suárez Allegue,
Carlos Gómez-Rodríguez, Margarita Alonso Ramos
Department of Informatics/CITIC, Department of Philology
Universidade da Coruña

June 26 2023

# One year into my MSCA grant




- Two years of funding (one year left)
  - ...thanks to all the advice I got from you!
- Goal: Grammar coaching with SRG + mal-rules

# Original scope

1. Developing the SRG for coverage
2. Increasing parsing speed
3. Designing and integrating mal-rules
4. Designing feedback
5. Developing and serving the application
6. Testing and supporting the application

# Reality so far

0 **Reverse-engineering Freeling-SRG interface and updating to Freeling 4.1**
   - Still not finished...

# Reality so far

0. **Reverse-engineering Freeling-SRG interface and updating to Freeling 4.1**
   - Still not finished...
1. Developing the SRG for coverage (not started)

# Reality so far

0. **Reverse-engineering Freeling-SRG interface and updating to Freeling 4.1**
   - Still not finished...
1. Developing the SRG for coverage (not started)
2. Increasing parsing speed (started)

# Reality so far

0. **Reverse-engineering Freeling-SRG interface and updating to Freeling 4.1**
   - Still not finished...
1. Developing the SRG for coverage (not started)
2. Increasing parsing speed (started)
3. Designing and integrating mal-rules
   - started working with the learner corpus
   - will have to focus on just gender agreement

# Reality so far



0. **Reverse-engineering Freeling-SRG interface and updating to Freeling 4.1**
   - Still not finished...
1. Developing the SRG for coverage (not started)
2. Increasing parsing speed (started)
3. Designing and integrating mal-rules
   - started working with the learner corpus
   - will have to focus on just gender agreement
4. Designing feedback (not started)
5. Developing and serving the application (not started)
6. Testing and supporting the application (not started)

# Freeling - SRG interface

- Freeling provides morphophonological analysis
- Freeling tags are mapped to lexical rules in the grammar

# Freeling - SRG interface

- Freeling provides morphophonological analysis
- Freeling tags are mapped to lexical rules in the grammar
- Orthography not handled in the grammar directly

# Freeling - SRG interface

- Freeling provides morphophonological analysis
- Freeling tags are mapped to lexical rules in the grammar
- Orthography not handled in the grammar directly
- Freeling 3.0 cannot be easily configured on modern OS
  - the precompiled version from LOGON can be used with the LKB but that is outdated and inflexible

# Freeling - SRG interface



- Freeling provides morphophonological analysis
- Freeling tags are mapped to lexical rules in the grammar
- Orthography not handled in the grammar directly
- Freeling 3.0 cannot be easily configured on modern OS
  - the precompiled version from LOGON can be used with the LKB but that is outdated and inflexible
- It was decided to update to Freeling 4.1

# Freeling - SRG interface

- The plan was:
  - Locate a list of tags that changed between 3.0 and 4.1
  - Replace-all the affected tag names in the grammar
  - **update the existing treebanks**

# Reality

- Old SRG relied on a special interface with Freeling
- a C-program which had to be reverse-engineered for the updated version
  - Practically undocumented interface
  - Too many docs for Freeling itself
- This reverse engineering took months in practice
- Updating the treebanks requires looking at every tree
  - ...or manipulating old treebanks in non-obvious ways

# SRG treebanks: A difficult comparison



- The old treebanks were never published or finished
  - Marimon 2010 reports only raw coverage
- They are only partially treebanked
- Some decisions should probably not be kept
  - e.g. questions should not be accepted as propositions (etc.)
- Only the beginning of the length distribution is represented (1-19 words)
  - the AnCora corpus (where the tbdb data comes from) goes up to length 129

# SRG treebanks: A difficult comparison

| Testsuite | Items | Cov | Old Cov | Acc | Old Acc* | RAM limit* |
|---|---|---|---|---|---|---|
| MRS | 106* | 0.97 | 1 | 0.79 | 0.84 | |
| tbdb01 | 65 | 1 | 1 | 1 | 1 | |
| tbdb02 | 177 | 0.92 | 0.98 | 0.90 | 0.92 | |
| tbdb03 | 181 | 0.90 | 0.91 | 0.87 | 0.84 | |
| tbdb04 | 219 | 0.90 | 0.92 | 0.86 | 0.81 | |
| tbdb05 | 229 | 0.89 | 0.93 | 0.82 | 0.80 | |
| tbdb06 | 211 | 0.89 | 0.92 | 0.80 | 0.80 | |
| tbdb07 | 246 | 0.89 | 0.91 | 0.76 | 0.82 | |
| tbdb08 | 278 | 0.90 | 0.93 | 0.81 | 0.82 | |
| tbdb09 | 326 | 0.88 | 0.93 | – | 0.80 | 5 |
| tbdb10 | 359 | 0.89 | 0.92 | – | 0.81 | 3 |
| tbdb11 | 352 | 0.88 | 0.89 | – | 0.75 | 3 |
| tbdb12 | 399 | 0.83 | 0.83 | – | 0.68 | 14 |
| tbdb13 | 357 | 0.82 | 0.80 | – | 0.65 | 15 |
| tbdb14 | 388 | 0.81 | 0.74 | – | 0.78 | 18 |
| tbdb15 | 378 | 0.68 | 0.78 | – | 0.60 | 29 |
| tbdb16 | 383 | 0.71 | 0.74 | – | 0.53 | 48 |
| tbdb17 | 408 | 0.65 | 0.71 | – | 0.51 | 80 |
| tbdb18 | 389 | 0.61 | 0.71 | – | 0.47 | 93 |
| tbdb19 | 443 | 0.55 | 0.64 | – | 0.28 | 127 |

# Remaining reconciliation issues

- Multiword expressions
  - including superfrequent ones e.g. *por qué*
- More ambiguity may be needed
  - including for superfrequent things like *ser* (to be)
- Some treebanking mysteries
  - maybe fftb doesn't render something unexpected

# The AnCora distribution

| sentence length | number of sentences | ratio |
|---|---|---|
| 0-4 | 644 | 3.7% |
| 5-9 | 1290 | 11.1% |
| 10-14 | 1858 | 21.8% |
| 15-19 | 2001 | 33.4% |
| 20-24 | 2096 | 45.4% |
| 25-29 | 1952 | 56.7% |
| 30-34 | 1949 | 67.9% |
| 35-39 | 1707 | 77.7% |
| 40-44 | 1401 | 85.8% |
| 45-49 | 1059 | 91.9% |
| 50-54 | 615 | 95.4% |
| 55-59 | 357 | 97.5% |
| 60-64 | 206 | 98.7% |
| 65-69 | 112 | 99.3% |
| 70-74 | 52 | 99.6% |
| 75-79 | 22 | 99.8% |
| 80-84 | 11 | 99.8% |
| 85-89 | 9 | 99.9% |
| 90-94 | 4 | 99.9% |
| 95-99 | 5 | 99.9% |
| 100-104 | 2 | 99.9% |
| 105-109 | 4 | 100% |
| 110-114 | 4 | 100% |
| 120-124 | 2 | 100% |
| 125-129 | 1 | 100% |
| 130-134 | 1 | 100% |
| 125-129 | 1 | 100% |

- Items up to length 19 represent 33% of the corpus
- We also only have a few hundred items from each portion
- would be good to parse and treebank up to e.g. length 45 but that's probably months of work
- However, let's refocus on the MSCA project objective...

# Learner Treebank for grammar checking

- The main goal of the project is grammar checking

# Learner Treebank for grammar checking

- The main goal of the project is grammar checking
- We will focus on gender agreement
  - have an annotated corpus

# Learner Treebank for grammar checking

- The main goal of the project is grammar checking
- We will focus on gender agreement
  - have an annotated corpus
- We created a subcorpus of a learner corpus
- ... and parsed it with the SRG

# COWSL2H learner corpus



- Written Spanish of L2 and Heritage Speakers
- Developed by UC Davis (Yamada et al. 2020)
- 900K words, 100K sentences, 2K authors
  - 50% of data and authors in 1st year (Intro)
- 7K sentences annotated for gender agreement errors
  - currently 50% coverage but there are errors in the sentences
  - 736 RAM-limit failures (default ACE settings)

# Plan for work on COWSL2H subcorpus

- Run the grammar on corrected and uncorrected sentences
  - Mistakes other than gender to be corrected manually, or sentences to be excluded
  - *El está muy cómica tambien*
  - (Intended: He is also very funny)
  - Él está muy [comica]{cómico}<ga:mf:adj:an> tambien.

# Plan for work on COWSL2H subcorpus

- Run the grammar on corrected and uncorrected sentences
  - Mistakes other than gender to be corrected manually, or sentences to be excluded
  - *El está muy cómica tambien*
  - (Intended: He is also very funny)
  - Él está muy [comica]{cómico}<ga:mf:adj:an> tambien.
- Determine coverage/overgeneration

# Plan for work on COWSL2H subcorpus

- Run the grammar on corrected and uncorrected sentences
  - Mistakes other than gender to be corrected manually, or sentences to be excluded
  - *El está muy cómica tambien*
  - (Intended: He is also very funny)
  - Él está muy [comica]{cómico}<ga:mf:adj:an> tambien.
- Determine coverage/overgeneration
- Treebank a few hundred sentences
- Improve the grammar accordingly

# Plan for work on COWSL2H subcorpus



- Run the grammar on corrected and uncorrected sentences
  - Mistakes other than gender to be corrected manually, or sentences to be excluded
  - *El está muy cómica tambien*
  - (Intended: He is also very funny)
  - Él está muy [comica]{cómico}<ga:mf:adj:an> tambien.
- Determine coverage/overgeneration
- Treebank a few hundred sentences
- Improve the grammar accordingly
- Finally, add mal-rules and develop the grammar checking system prototype

# Summary

- Very exciting to revive such a large grammar
- The documentation in the grammar (particularly examples) helps **a lot**
- Having even small treebanks as a baseline is great

# Summary

- Grammar engineering continues to be time consuming
- Freeling dependency proves problematic, despite the convenience of YY-input
  - Passing a grammar from one developer to another is hard, but with a Freeling interface it is harder
  - Updating treebanks is very time consuming even with FFTB autoupdate
  - Would better documentation solve this?
    - Maybe, but a lot of documentation is no better than no documentation