Extending the tool, or how to annotate historical language varieties

Cristina Sánchez-Marco
Universitat Pompeu Fabra

Gemma Boleda, Lluís Padró
Universitat Politècnica de Catalunya

Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011), ACL workshop, Portland, Oregon, June 24 2011
Extending the tool, or how to annotate historical language varieties
Introduction
The analyzer
Data
Method
Evaluation
Conclusion & Future work

Extending the tool, or how to annotate historical language varieties

Cristina Sánchez-Marco
Universitat Pompeu Fabra
Gemma Boleda
Universitat Politècnica de Catalunya
Extending the tool, or how to annotate historical language varieties
Extending the tool, or how to annotate historical language varieties
An example

[fol. 32v]

{CB1.
Ca delo q<ue> mas amaua yal viene el mandado
Dozi[en]tos cauall<er>os mando exir p<r><<i>>uado
Q<ue> Rec`iban a myanaya alas duenas fijas dalgo
El sedie en valenc`ia curiendo guardando
Ca bie<n> sabe q<ue> albarfanez t<r><<a>>he todo Recabdo
An example: Paleographic symbols

[fol. 32v]

{CB1.

Ca delo q<ue> mas amaua yal viene el mandado
Dozi[en]tos caull<er>os mando exir p<r><i>uado
Q<ue> Rec’iban a myanaya alas duenas fijas dalgo
El sedie en valenc’ia curiando guardando
Ca bie<n> sabe q<ue> albarfanez t<r><a>he todo Recabdo
An example: Non-standard spelling

[fol. 32v]
{CB1.
Ca delo de lo que mas amaua amaba yal ya le viene el mandado
Dozientos Doscientos cavalleros caballeros mando mandó exir priuado privado
Que Rec’iban reciban a myanaya alas a las duenas dueñas fijas dalgo hidalgas
El Él sedie seria en valenc’ia valencia curiendo curando & guardando
Ca bien sabe que albarfanez trahe trae todo Recabdo Recaudo
An example: Capital letters and punctuation

[fol. 32v]

{CB1.
Ca delo que mas amaua yal viene el mandado,
Dozientos caualleros mando exir priuado,
Que Rec’iban a myanaya Myanaya alas duenas fijas dalgo,
El sedie en valenc’ia Valencia curiando guardando,
Ca bien sabe que albarfanez Albarfanez trahe todo Recabdo recaudo.
An example: Non-standard word order

[fol. 32v]
{CB1.

Ca delo que mas amaua yal viene el mandado
Dozientos caualleros mando exir priuado
Que Rec’iban a myanaya alas duenas fijas dalgo
El sedie en valenc’ia curiando guardando
Ca bien sabe que albarfanez trahe todo Recabdo
Solutions to this challenge

1. Use a tool for a standard variety
   - Advantages: resource saving
   - Disadvantages: non-acceptable accuracy, manual correction

2. Adapt the tool
   - Advantages: reusable, sustainable, relatively easy to adapt, resource saving, extensible to other language varieties
Our specific case study and proposal
Our specific case study and proposal

Solution 2
Adapt the tool
Our specific case study and proposal

Solution 2
Adapt the tool

- Tool: Freeling [http://nlp.lsi.upc.edu/freeling](http://nlp.lsi.upc.edu/freeling)
- Language: Old Spanish
Our specific case study and proposal

Solution 2
Adapt the tool

- Tool: Freeling http://nlp.lsi.upc.edu/freeling
- Language: Old Spanish

Specific advantages of adapting Freeling
Our specific case study and proposal

Solution 2
Adapt the tool

- Tool: Freeling http://nlp.lsi.upc.edu/freeling
- Language: Old Spanish

Specific advantages of adapting Freeling
- open-source
- well documented and actively maintained
- modular, relatively easy to adapt
The analyzer

raw text

ANALYZER

tokenizer

morphological analysis

dictionary

affixation

probabilities

TAGGER

probabilities

tagged corpus

Cristina Sánchez-Marco
Universitat Pompeu Fabra

Gemma Boleda, Lluís Padró
Universitat Politècnica de Catalunya

Extending the tool, or how to annotate historical language varieties
Data

1. Old Spanish Corpus
2. Gold Standard Corpus
3. Standard Spanish Corpus
1. Old Spanish Corpus

- Electronic editions by the Hispanic Seminary of Medieval Studies
- Critical editions of the original manuscripts
- 12th to 16th century Spanish
- Representative corpus:
  - more than 20 million tokens, 470 thousand types
  - variety of genres (fiction and non-fiction)
2. Gold Standard Corpus

- 30,000 tokens from the Old Spanish Corpus
- to retrain the tagger and carry out the evaluation and error analysis
- mirroring the Old Spanish Corpus in size and text-type distribution
3. Standard Spanish Corpus

- baseline performance for the tagger
- Corpus *LexEsp* (Sebastián et al 2000)
- from 1975 to 1995
- more than 5 million words
- variety of genres
Method

Using the existing Standard Spanish tool as a basis to create an Old Spanish processor

- Expansion of the dictionary
- Modification of other modules: tokenization, affixation
- Retraining of the tagger
Dictionary expansion: Data

- 32,015 Old Spanish words (55,000 lemma-tag pairs) + 556,210 Standard Spanish words = 588,225 word forms
Dictionary expansion: Data

- 32,015 Old Spanish words (55,000 lemma-tag pairs) +
  556,210 Standard Spanish words = 588,225 word forms

Distribution of words added to the dictionary

| Category     | Percentage |
|--------------|------------|
| Verbs        | 48.8%      |
| Nouns        | 20.8%      |
| Adjectives   | 7.0%       |
| Pronouns     | 0.6%       |
| Prepositions | 0.5%       |
| Adverbs      | 0.4%       |
| Determiners  | 0.3%       |
| Conjunctions | 0.3%       |
| Interjections| 0.2%       |
| Numbers      | 0.2%       |
| Punctuation  | 0.01%      |
Dictionary expansion: Method

- Mapping rules:
  - **Substring rules** (54 sequences of characters): 81.4% of the words added

Old Modern Example

- euo → nuevo
- uio → vió
- ujd → vida
- -f -ube → nube
- sp- esp- → espera

Word rules: 18.5% of the words added

Example

- consul → c´ onsul
- catholica → cat´ olica
Dictionary expansion: Method

- Mapping rules:
  - Substring rules (54 sequences of characters): 81.4% of the words added

| Old  | Modern | Example    |
|------|--------|------------|
| euo  | evo    | nuevo → nuevo 'new' |
| uio  | vio    | uio → vio 'saw' |
| ujd  | vid    | ujda → vida 'life' |
| -f   | -ube   | nuf → nube 'cloud' |
| sp-  | esp-   | spera → espera 'wait' |
Dictionary expansion: Method

- **Mapping rules:**
  - **Substring rules** (54 sequences of characters): 81,4% of the words added

| Old  | Modern | Example            |
|------|--------|--------------------|
| euo  | evo    | nuevo → nuevo 'new'|
| uio  | vio    | uio → vio 'saw'    |
| ujd  | vid    | ujda → vida 'life' |
| -f   | -ube   | nuf → nube 'cloud' |
| sp-  | esp-   | spera → espera 'wait' |

- **Word rules:** 18,5% of the words added
Dictionary expansion: Method

- **Mapping rules:**
  - **Substring rules** (54 sequences of characters): 81.4% of the words added

| Old  | Modern | Example            |
|------|--------|--------------------|
| euo  | evo    | nuevo \(\rightarrow\) nuevo 'new' |
| uio  | vio    | uio \(\rightarrow\) vio 'saw'     |
| ujd  | vid    | ujda \(\rightarrow\) vida 'life'   |
| -f   | -ube   | nuf \(\rightarrow\) nube 'cloud'         |
| sp-  | esp-   | spera \(\rightarrow\) espera 'wait'     |

- **Word rules:** 18.5% of the words added

| Example         |
|-----------------|
| consul \(\rightarrow\) cónsul 'consul' |
| catholica \(\rightarrow\) católica 'catholic' |
Modification of other modules

- **Tokenization:** special symbols
  - $c'$ (yac’e ’lay’) (standard: $\varsigma$)
  - $n^\sim$ (cin$\sim$o ’adhered’) (standard: $\tilde{n}$)

- **Affixation:** non standard suffixes
  - adverbs in -miente or -mjentre (standard: -mente)
  - clitics -gela, -li (standard: -sela, -le)
Retraining of the tagger

- Use of the Gold Standard Corpus
- Two taggers:
  - Hybrid (*relax*)
  - Hidden Markov Model (*hmm*)
Retraining of the tagger

- Use of the Gold Standard Corpus
- Two taggers:
  - Hybrid \textit{(relax)}
  - \textbf{Hidden Markov Model (hmm)}
Evaluation

- Dictionary:
  - Ambiguity
  - Coverage
  - Precision
  - Recall
Evaluation

- Dictionary:
  - Ambiguity
  - Coverage
  - Precision
  - Recall

- Tagging
  - Accuracy
Evaluation of the dictionary

| Ambiguity       | Type-based | Token-based |
|-----------------|------------|-------------|
| Old Spanish     | 1.21       | **1.85**    |
| Standard Spanish| 1.20       | **1.68**    |
### Evaluation of the dictionary

| Ambiguity         | Type-based | Token-based |
|-------------------|------------|-------------|
| Old Spanish       | 1.21       | **1.85**    |
| Standard Spanish  | 1.20       | **1.68**    |

|                      | Precision | Recall | Coverage |
|----------------------|-----------|--------|----------|
|                      | Lemmas    | PoS    |          |
| Old Spanish          | 99.2%     | 98.6%  | 92.6%    |
| Standard Spanish     |           |        | 99.4%    |
Accuracy in the tagging

- C0: original tool (standard Spanish)
- C1: expanded dict. + modules + standard Spanish tagger
- C2: expanded dict. + modules + trained tagger (10,000-token)
- C3: expanded dict. + modules + trained tagger (20,000-token)
- C4: expanded dict. + modules + trained tagger (30,000-token)
Accuracy in the tagging

- C0: original tool (standard Spanish)
- C1: expanded dict. + modules + standard Spanish tagger
- C2: expanded dict. + modules + trained tagger (10,000-token)
- C3: expanded dict. + modules + trained tagger (20,000-token)
- C4: expanded dict. + modules + trained tagger (30,000-token)

|       | Lemma | PoS-1 | PoS-2 |
|-------|-------|-------|-------|
| C0    | 72.4  | 70.9  | 77.4  |
| C1    | 90.7  | 86.0  | 91.0  |
| C2    | 91.2  | 87.5  | 91.9  |
| C3    | 92.3  | 89.5  | 93.7  |
| C4    | **92.6** | **89.9** | **94.5** |
| SS    | **99.1** | 94    | **97.6** |
Error analysis

- 100 most frequent errors in the tagging
- 90% errors are due to ambiguity
  - determiner vs clitic readings of *la, las 'the/it’*
  - first vs third person singular (*quería 'I/he wanted’*)
  - accentuation (*llegó 'arrive/arrived’*) (standard: *llego/llegó*)
- 10% errors are words out of the dictionary
  - proper nouns (*pierres, antolinez*)
  - words not covered by any mapping rule (*coita 'wish’*)
    (standard: *cuita*)
Conclusion

- Simple and general method to adapt an existing tool for Modern standard Spanish in order to annotate Old Spanish
- Advantages: resource-saving, sustainable, relatively easy
- Benefit from:
  - similarities between historical and modern standard language varieties
  - existing tools
  - philological studies
- The quality of the tagging is quite good
- The greatest improvement is obtained when the lexicon is expanded with non-standard variants
- Method extensible to other non-standard language varieties
Future work

- Bigger training corpus
- Incorporate rules to the grammar
- Extend the tool for other non-standard language varieties
Thanks!
cristina.sanchezm@upf.edu
gboleda@lsi.upc.edu
padro@lsi.upc.edu

Acknowledgements
contracts AP2006-03547 (FPU), FFI2010-15006 (OntoSem 2), TIN2009-14715-C04-04 (KNOW-II), JCI2007-57-1479, EU PASCAL2 Network of Excellence (FP7-ICT-216886)