“A mas novas vos torn / Now I take you back to my tale”

The Romance of Flamenca

Olga Scrivner, E.D. Blodgett*, Sandra Kübler, Michael McGuire

Indiana University
*University of Alberta

June 2013
In the past, historical documents and manuscripts were studied exclusively by using a manual paper-based approach.

Recent achievements of corpus linguistics have introduced state-of-art methods and tools for digitization, semi-automatic annotation, and visualization of such resources.
Linguistic Annotation

“By accessing linguistic annotation, we can extend the range of phenomena that can be found with high precision” (Kübeler and Zinsmeister, 2014)

1. Morphological annotation - collocations, spelling variation
2. Syntactic annotation - sentence structure in narratives vs. dialogues, prose vs. verse
3. Discourse annotation - analysis of scenes and characters (Female vs. male speaker, King vs. servants)
In recent years, a number of annotated corpora have been developed for Medieval Romance languages:

- Corpora of Old Spanish (Davies, 2002)
- Old Portuguese (Davies and Ferreira, 2006)
- Old French (Stein, 2008; Martineau et al., 2010)
In recent years, a number of annotated corpora have been developed for Medieval Romance languages:

- Corpora of Old Spanish (Davies, 2002)
- Old Portuguese (Davies and Ferreira, 2006)
- Old French (Stein, 2008; Martineau et al., 2010)

There exist (to our knowledge) two electronic databases:

1. “The Concordance of Medieval Occitan” (Ricketts and Reed, 2005)
2. “Provençal poetry” (ARTFL Project, 1998)

Users of those corpora are limited to lexical search.
Le Roman de Flamenca, a “universally acknowledged masterpiece of Old Occitan narrative” (Fleischmann, 1995).

“Flamenca est la création d’un homme d’esprit qui a voulu faire une oeuvre agréable où fût représentée dans ce qu’elle avait de plus brillant la vie des cours au XII siècle. C’était un roman de moeurs contemporaines” (Meyer, 1865)

“Flamenca is the creation of a man of talent who wished to write an agreeable work representing the most brilliant aspects of courtly life in the twelfth century. It is a novel of manners” (Bradley, 1922)
This romance presents a very intriguing love story between the beautiful **Flamenca**, who is imprisoned in a tower by her jealous husband **Archambaut**, and the sharp-witted knight **Guillem**.

The photo of the tapestry is used by permission of FreeLargePhotos.com
The anonymous manuscript of *Le Roman de Flamenca* was accidentally discovered in Carcassonne (France) by Raynouard and was first fully edited P. Meyer in 1865.

This romance is unique in genre (“the first modern novel”), its use of setting, adventures, and character portrayal (Blodgett, 1995; Bradley, 1922; Meyer, 1865).

The potential value of this historical resource, however, is limited by the lack of an accessible digital format and linguistic annotation.
Goals

Our corpus is intended not only as material for linguistic research, but also to aid in broader studies:

- Interactive online database with access to a glossary, to translations of verses, and to comments (Meyer, 1901) http://nlp.indiana.edu/~obscrivn/Introduction.html

- Multiple-level annotation - morphological, syntactic and pragmatic (Scrivner et al., 2013)

- Parallel English-Occitan corpus (Blodgett, 1995)
**What is Parallel Corpus**

- **Parallel corpus** is “an association between two texts (written or spoken) in different languages that represent translations of each other” (Tufis, 2006).

- **Parallel alignment** is reciprocal translation units that encode valuable lexical and syntactic knowledge.
Alignment types

- One-to-one: one word from a source language corresponds to only one word in a target language.

Example:

Eu: ai ric
I have a shady grove
Alignment types

- **One-to-one**: one word from a source language corresponds to only one word in a target language
  
  ![Example of one-to-one alignment]

- **One-to-many**:
  
  ![Example of one-to-many alignment]
Given that parallel words have the same content, we can identify forms that have not been studied (Koolen et al., 2006; Enrique-Arias, 2012):

- Spelling and lexical variation
- Morphosyntactic variation
- Null occurrences
A mas novas Ø vos torn Now I take you back to my tale
Null Occurrences

A mas novas Ø vos torn Now I take you back to my tale

I take you back to my tale

explicit

implicit
Automatic Alignment

- English and Occitan texts are aligned by lines of verses
- Bilingual lexicon is generated by NATools\(^1\) (Matrix of word-to-word probabilities)
- Automatic alignment via Berkeley parser (Liang et al., 2006)
- Manual correction of alignment

\(^1\)http://linguateca.di.uminho.pt/natools/
<table>
<thead>
<tr>
<th>Tag</th>
<th>Definition</th>
<th>Tag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>adjective</td>
<td>MDPP</td>
<td>past participle of modal verb</td>
</tr>
<tr>
<td>ADJR</td>
<td>comparative form of adjective</td>
<td>MDX</td>
<td>infinitive of modal verb</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>NCPL</td>
<td>noun common plural</td>
</tr>
<tr>
<td>ADVR</td>
<td>comparative form of adverb</td>
<td>NCS</td>
<td>noun common singular</td>
</tr>
<tr>
<td>AG</td>
<td>gerundive of auxiliary 'to have'</td>
<td>NEG</td>
<td>negation</td>
</tr>
<tr>
<td>AJ</td>
<td>present of auxiliary 'to have'</td>
<td>NPRPL</td>
<td>noun proper plural</td>
</tr>
<tr>
<td>APP</td>
<td>past participle of auxiliary 'to have'</td>
<td>NPRS</td>
<td>noun proper singular</td>
</tr>
<tr>
<td>AX</td>
<td>infinitive of auxiliary 'to have'</td>
<td>NUM</td>
<td>numeral</td>
</tr>
<tr>
<td>CONJO</td>
<td>coordinative conjunction</td>
<td>P</td>
<td>preposition</td>
</tr>
<tr>
<td>CONJS</td>
<td>subordinate conjunction</td>
<td>PON</td>
<td>punctuation inside the clause</td>
</tr>
<tr>
<td>COMP</td>
<td>comparative adverb</td>
<td>PONFP</td>
<td>the end of the sentence</td>
</tr>
<tr>
<td>D</td>
<td>determiner (indefinite, definite, demonstrative)</td>
<td>PRO</td>
<td>pronoun</td>
</tr>
<tr>
<td>DAT</td>
<td>dative</td>
<td>Q</td>
<td>quantifier</td>
</tr>
<tr>
<td>DZ</td>
<td>possessive determiner</td>
<td>VG</td>
<td>gerundive of the main verb</td>
</tr>
<tr>
<td>EG</td>
<td>gerundive of auxiliary 'to be'</td>
<td>VJ</td>
<td>present of the main verb</td>
</tr>
<tr>
<td>EJ</td>
<td>present of auxiliary 'to be'</td>
<td>VPP</td>
<td>past participle of the main verb</td>
</tr>
<tr>
<td>EPP</td>
<td>past participle of auxiliary 'to be'</td>
<td>VX</td>
<td>infinitive of the main verb</td>
</tr>
<tr>
<td>EX</td>
<td>infinitive of auxiliary 'to be'</td>
<td>WADV</td>
<td>interr., rel. or excl. adverb</td>
</tr>
<tr>
<td>ITJ</td>
<td>interjection</td>
<td>WD</td>
<td>interr., rel. or excl. determiner</td>
</tr>
<tr>
<td>MDG</td>
<td>gerundive of modal verb</td>
<td>WPRO</td>
<td>interr., rel. or excl. pronoun</td>
</tr>
<tr>
<td>MDJ</td>
<td>present of modal verb</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
flamenca non i volc giquir ;
NPRS NEG PRO MDJ VX PONFP
```
### Syntactic Annotation - Berkeley Parser

<table>
<thead>
<tr>
<th>Labels</th>
<th>Definition</th>
<th>Labels</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJP</td>
<td>Adjectival Phrase</td>
<td>CP-THT</td>
<td>Complement Clause</td>
</tr>
<tr>
<td>ADVP</td>
<td>Adverbial Phrase</td>
<td>INTJ</td>
<td>Interjection</td>
</tr>
<tr>
<td>ADV-P-LOC</td>
<td>Adverbial Locative Phrase</td>
<td>IP-IMP</td>
<td>Imperative Proposition</td>
</tr>
<tr>
<td>ADV-P-TMP</td>
<td>Adverbial Temporal Phrase</td>
<td>IP-INF</td>
<td>Infinitival Proposition</td>
</tr>
<tr>
<td>CONJP</td>
<td>Conjunction</td>
<td>IP-MAT</td>
<td>Main Proposition</td>
</tr>
<tr>
<td>CP-ADV</td>
<td>Adverbial Clause</td>
<td>IP-PPL</td>
<td>Participial Proposition</td>
</tr>
<tr>
<td>CP-ADV-TMP</td>
<td>Temporal Clause</td>
<td>IP-SUB</td>
<td>Subordinate Proposition</td>
</tr>
<tr>
<td>CP-CAR</td>
<td>Prepositional Clause</td>
<td>NP-ACC</td>
<td>Direct Object</td>
</tr>
<tr>
<td>CP-CMP</td>
<td>Comparative Clause</td>
<td>NP-DTV</td>
<td>Indirect Object</td>
</tr>
<tr>
<td>CP-DEG</td>
<td>Degree Clause</td>
<td>NP-PRD</td>
<td>Predicative NP</td>
</tr>
<tr>
<td>CP-EXL</td>
<td>Exclamative Clause</td>
<td>NP-RFL</td>
<td>Reflexive NP</td>
</tr>
<tr>
<td>CP-FRL</td>
<td>Small Clause</td>
<td>NP-SBJ</td>
<td>Subject NP</td>
</tr>
<tr>
<td>CP-OPT</td>
<td>Optative Clause</td>
<td>NP-TMP</td>
<td>Temporal NP</td>
</tr>
<tr>
<td>CP-QUE</td>
<td>Interrogative Clause</td>
<td>PP</td>
<td>Prepositional</td>
</tr>
<tr>
<td>CP-REL</td>
<td>Relative Clause</td>
<td>PP-DIR</td>
<td>Prepositional Directional Phrase</td>
</tr>
<tr>
<td>PP-LOC</td>
<td>Prepositional Locative Phrase</td>
<td>QR</td>
<td>Quantifier Phrase</td>
</tr>
<tr>
<td>V</td>
<td>Verb</td>
<td>-PRN</td>
<td>Adjunct</td>
</tr>
<tr>
<td>-LFD</td>
<td>Left Dislocated Phrase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-SPE</td>
<td>Direct Speech</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“...nor did he want to omit Flamenca”
Discourse Annotation - Speakers

The labels correspond to the main characters names, namely Flamenca, Archambaut, Guillem, Father, King, Queen.

<table>
<thead>
<tr>
<th>narration</th>
<th>Father</th>
</tr>
</thead>
<tbody>
<tr>
<td>dis</td>
<td>vostre</td>
</tr>
<tr>
<td>tot</td>
<td>cor</td>
</tr>
<tr>
<td>en</td>
<td>no</td>
</tr>
<tr>
<td>apert</td>
<td></td>
</tr>
</tbody>
</table>

Less important characters are marked as FemaleSpeakers and MaleSpeakers.
Parallel Alignment Annotation

joy as mine for mentioning flamenca by name.

es le mieus quan vos flamenca mentaves.

VJ D ADJ CONJS PRO NPRS VJ
Since we are targeting two different types of users, linguists and non-linguistics, with different needs, the corpus is made available in two different modes:

- **Web Interface**: Users can mainly browse the text and look up translations, glosses, and comments.
- **Query Search**: Users interested in the linguistic annotation can query the corpus on-line.
1. Web Database

Glossary definitions, comments, and footnotes are linked to tokens and are made visible when the user hovers over a marked word.

183. Que lail des lo venres ol sapte :
184. Si per compra ni per acapte
185. acapte 184. acquisition par voie d'acapte, distinct de la compra, qui désigne un genre d'acquisition plus complet
186.
187. L'endeman de [la] Pantecosta (French)
188. Dreg a Nemurs li cors s'ajosta
189. Bela e rica e pleniera .
190. Anc [mais] negus\textsuperscript{190} hom non vi fiera ,
191. Ni a Liniec ni a Proîs ,
192. Que i agues tant e var\textsuperscript{192} e gris
193. E drap de seda e de lana .
Our web search based on ANNIS allows for basic queries, to search for a word or phrase, and more complex queries for syntactic, morphosyntactic, discourse and alignment annotation.
Modern Occitan varieties are null subject languages (Hinzelin and Kaiser, 2012)

(1) Ø Era pertot, dintrava pertot
    was everywhere, entered everywhere
    ‘the light was everywhere and it was coming from everywhere’ (Lo bèn de la nuòch, Max Rouquette)
Overt subjects - disambiguation or “mise en relief”

(2) Femna que ieu ame illuminada de non rën
‘Woman who I love illuminated from nothingness’
(Saume dins lo vent, Serge Bec)

- Person - more frequent with 1st person (Vance, 2009)
- Genre - more frequent in prose
- No difference by clause types (Sitaridou, 2005)
- Subjunctive clause - preference for null subject (Vance, 2009)
Search by Genre

Discourse annotation: Flamenca, King etc

ex. `speaker="Flamenca"`

- Narrative vs. Dialogues
- Male vs. Female
- High social rank vs. low social rank
Token annotation: I, you, it etc

ex. \texttt{tok=“it”}

- Personal pronouns
- Impersonal pronouns
Search by Clause Type

Syntactic annotation:

ex. matrix \[\text{cat}="\text{IP}\] >[\text{func}="\text{MAT}\]

embedded \[\text{cat}="\text{IP}\] >[\text{func}="\text{SUB}\]

- Main clause
- Embedded clause
Overt vs. Null Subjects

Searching for null subjects:

- **Occitan node**
  - align

- **English node**
  - equal

- **Occitan verb**
  - cat = V

- **English pronoun**
  - epos = PRP
Searching for overt subjects:
## Overt vs. Null Subjects - 1000 lines

<table>
<thead>
<tr>
<th>Factor</th>
<th>Null (%</th>
<th>Overt (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>308 (87)</td>
<td>45 (13)</td>
</tr>
<tr>
<td>Matrix Clause</td>
<td>34 (87)</td>
<td>5 (13)</td>
</tr>
<tr>
<td>Embedded Clause</td>
<td>107 (84)</td>
<td>21 (16)</td>
</tr>
<tr>
<td>Impersonal pronouns</td>
<td>23 (92)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>1st person</td>
<td>32 (70)</td>
<td>14 (30)</td>
</tr>
<tr>
<td>2nd person</td>
<td>28 (88)</td>
<td>4 (12)</td>
</tr>
<tr>
<td>3rd person</td>
<td>200 (91)</td>
<td>19 (9)</td>
</tr>
<tr>
<td>Narration</td>
<td>187 (91)</td>
<td>19 (9)</td>
</tr>
<tr>
<td>Discourse</td>
<td>121 (88)</td>
<td>26 (12)</td>
</tr>
</tbody>
</table>
Explicite Impersonal Pronouns

Only tonic pronouns

1. mais aisso -m par causa tro brava
   but it seems to me hard

2. mais so fon sos meillors thesaurs
   and it was her greatest treasure
## Social Variation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Null (%)</th>
<th>Overt (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male speakers</td>
<td>46 (87)</td>
<td>7 (13)</td>
</tr>
<tr>
<td>Female speaker</td>
<td>29 (78)</td>
<td>8 (12)</td>
</tr>
<tr>
<td>High social status</td>
<td>54 (86)</td>
<td>9 (13)</td>
</tr>
<tr>
<td>Low social status</td>
<td>29 (83)</td>
<td>6 (17)</td>
</tr>
<tr>
<td>Author</td>
<td>11 (92)</td>
<td>1 (8)</td>
</tr>
</tbody>
</table>
In contrast to traditional corpora, this corpus is structured to fulfill two objectives:

- The web design facilitates the reading and understanding of *The Romance of Flamenca*. Words are interactively linked to the glossary, comments, and translations.

- The corpus search design via its ANNIS interface allows for a visualization and for complex queries of the (morpho-)syntactic, discourse and parallel word aligned annotations.
Future Directions

- Culmination of manual correction
- Preservation and annotation of other Old Occitan manuscripts
- Building a collaborative effort to continue with this project
http://artfl-project.uchicago.edu/content/proven%5Cc%7Bc%7Dal.


