Le Roman de Flamenca: An Annotated Corpus of Old Occitan

Olga Scrivner, Sandra Kübler, Barbara Vance, Eric Beuerlein

Indiana University
{obscrivn,skuebler,bvance,ebeuerle}@indiana.edu

Abstract

This paper describes an ongoing effort to digitize and annotate the corpus of *Le Roman de Flamenca*, a 13th-century romance written in Old Occitan. The goal of this project is twofold: The first objective is to digitize one of the earliest editions of the text and to create an interactive online database that will allow parallel access to a glossary, to translations of verses, and to comments from Paul Meyer's edition. The second objective is to lemmatize and syntactically annotate the corpus and make it accessible using the ANNIS online-search engine.

1 Introduction

Le Roman de Flamenca holds a unique position in Provençal literature. "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[I] siècle. C'était un roman de moeurs contemporaines[1]" [14]. In the past, the 13th-century manuscript of Flamenca has been extensively studied in its raw text format. The potential value of this historical resource, however, is limited by the lack of an accessible digital format and linguistic annotation.

This paper focuses on the creation of an annotated corpus of Old Occitan that preserves one of the earliest editions of the manuscript [15]. While it was succeeded by many editions and translations, "no student of the manuscript can afford to overlook Meyer's editions" [2]. Thus, the purpose of this corpus is twofold. It is intended not only as material for linguistic research, but also to aid in broader studies. The first objective is to provide access to *Le Roman de Flamenca* through an interactive online database that will allow parallel access to a glossary, to translations of verses, and to comments from Paul Meyer's edition [15]. The second

[1] "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" [3]. Note that elsewhere Meyer places Flamenca in the 13th century, a date which is also universally accepted today, and so the reference here to the 12th century must be in error.

objective is to lemmatize and syntactically annotate the corpus and to make it accessible using the ANNIS web-search engine.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the language Old Occitan and the romance of Flamenca. Section 3 outlines the structure, content, and annotation process of the corpus. Section 4 describes a range of applications of the corpus interfaces. Section 5 concludes and presents directions for future work.

2 Old Occitan and *Le Roman de Flamenca*

Old Occitan, formerly referred to as Old Provençal after one of its major dialects, is the ancestor of the endangered language spoken today in southern France by an undetermined number of bilingual individuals. This language constitutes an important element of the literary, linguistic, and cultural heritage of the Romance languages; it was known throughout the western medieval world through the lyric poetry of the Troubadours. While the historical importance of this language is indisputable, Occitan, as a language, remains linguistically understudied. In the past decade, a number of annotated corpora have been developed for other Medieval Romance languages, for example, Old Spanish [5], Old Portuguese [6], and Old French [13, 19]. However, annotated data for Old Occitan are still sparse. There exist (to our knowledge) two electronic databases, "The Concordance of Medieval Occitan"[2] [17] and "Provençal poetry"[3] [1], but users of those corpora are limited to lexical search.

This project focuses on the 13th-century Old Occitan romance *Le Roman de Flamenca*. The anonymous manuscript of *Le Roman de Flamenca* was accidentally discovered in Carcassonne (France) by Raynouard and was first fully edited and translated by P. Meyer in 1865. This romance has been variously characterized as a comedy of manners, "the first modern novel", and a psychological romance, among other characterizations [2, 3, 14]. This prose in verse played an influential role in the development of French literature [11]. Apart from a very intriguing love story between beautiful Flamenca, who is imprisoned in a tower by her jealous husband Archambaut, and the sharp-witted knight Guillem, this 8095-line story is a very interesting linguistic document and is the "universally acknowledged masterpiece of Old Occitan narrative" [8]. This romance features multiple literary styles, such as internal monologues, dialogues, and narratives, and offers a rich lexical, morphological, and syntactic representation of the language spoken in medieval southern France.

[2]The database includes Gschwind's edition [9] of *Le Roman de Flamenca*

[3]http://artfl-project.uchicago.edu/content/provençal

3 Structure of the Corpus *Le Roman de Flamenca*

3.1 Data

We are convinced that a digitized and annotated corpus of Old Occitan will be a valuable resource, not only for corpus linguistics studies, but also for a more general audience that wishes to become acquainted with Occitan literature. Therefore, one important goal is public accessibility. As a consequence, the corpus comprises only the editions that are not subject to copyright restrictions. Thus, we have selected the second edition of the manuscript by Meyer [15], supplemented with a glossary, translation of the manuscript into French, and a plot summary from the first edition by Meyer [14].

3.2 Workflow

The Romance of Flamenca [14, 15] is available in a scanned format, digitized by Google[4]. The manuscript consists of 8095 lines, footnotes, and 110 pages of glossary. As an initial step of text processing, digital images of the book are first sent through OCR and then manually corrected, using TESSERACT.[5] As a baseline, TESSERACT's Old French language model was used. With Old French, the OCR engine was then trained using sample images from *Flamenca*. However, it quickly became apparent that the glossary and character selection of Old French, although the closest match, is not ideal. Specifically, the Old French model contains numerous diacritics (e.g., é, û, ë) that are not used in Old Occitan. For better initial OCR results, these characters were disallowed during text recognition. In addition, some individual text strings were disallowed as they caused systematic problems. For example, without additional commands, the Old French search engine regularly recognized the Old Occitan word "anc" as "ane". Therefore, we disallowed "ane" and at the same time added "anc" to a list of additional words.

Figure 1 shows a scanned page from the Meyer's edited manuscript, which was then converted into a text format. We have created a TEI XML document, preserving verse lines, footnotes, paging, glossary, and comments by the author. We have also included the French translation provided by Paul Meyer in his first edition of the manuscript [14].

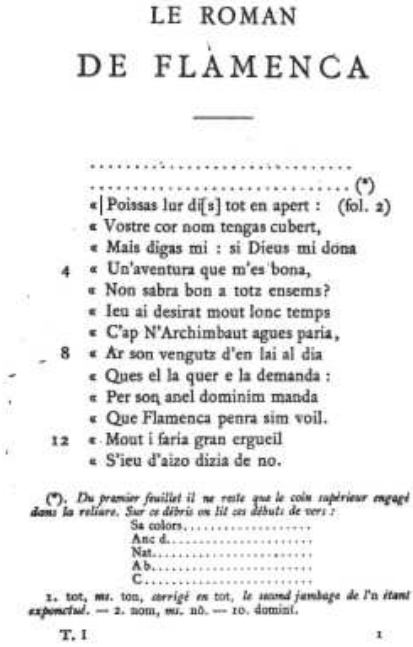
The second step of processing includes the following steps: 1) text tokenization, 2) POS tagging, 3) lemmatizing, 4) parsing, 5) converting the text into the format used by ANNIS [6], and 6) uploading it to the ANNIS database, for use with its graphical interface [20].

First, we have segmented our corpus into 52 200 lexically relevant tokens. In cases where the Meyer edition had a pronoun attached to a verb, negation, or other pronoun, we have detached them. For example, in (1a) the clitic *m* 'me' is attached

[4] http://www.archive.org/details/leromandeflamen00meyegoog

[5] http://tesseract-ocr.googlecode.com/svn/trunk/doc/tesseract.1.html

[6] http://korpling.german.hu-berlin.de/saltnpepper/



LE ROMAN

DE FLAMENCA

.....

..... (*)

« Poissas lur di[s] tot en apert : (fol. 2)
« Vostre cor nom tengas cubert,
« Mais digas mi : si Dieus mi dona
« Un'aventura que m'es bona,
« Non sabra bon a totz ensems?
« Ieu ai desirat mout lonc temps
« C'ap N'Archimbaut agues paria,
« Ar son vengutz d'en lai al dia
« Ques el la quer e la demanda :
« Per son anel dominim manda
« Que Flamenca penra sim voil.
« Mout i faria gran ergueil
« S'ieu d'aizo dizia de no.

(*). *Du premier feuillet il ne reste que le coin supérieur engagé dans la reliure. Sur ce débris on lit ces débuts de vers :*
Sa colors.....
Anc d.....
Nat.....
Ab.....
C.....

1. tot, *ms.* ton, *corrigé en* tot, *le second jambage de l'n étant exponctué.* — 2. nom, *ms.* nō. — 10. domini.

T. I 1

Figure 1: Scanned image of the Meyer's edition of *Flamenca*

to the preceding noun *domini* 'lordship' and to the conjunction *si* 'if'. The detached clitics are shown in (1b). In our syntactic interpretation of such cases, we follow more recent editions of the manuscript [2, 9, 10].

(1) a. *Per son anel domini**m*** *manda Que Flamenca penra si**m***
for his ring lordship-me sends that Flamenca takes if-me
voil.
want
'With his own ring he has let me know that he wants to marry Flamenca if I permit.' (line 10, [2])

b. Per son anel domini **-m** manda Que Flamenca penra si **-m** voil .

Since creating linguistic corpus annotation can be a labor-intensive and time-consuming process, we have addressed this issue with a resource-light method that exploits existing resources. Several studies have shown that linguistic information available for a resource-rich language can be transferred to a closely related resource-poor language [7, 18]. Given that Old Occitan and Old French share many lexical, morphological, and syntactic characteristics, we have selected the MCVF

Tag	Definition	Tag	Definition
ADJ	adjective	MDJ	present of modal verb
ADJNUM	cardinal adjective	MDPP	past participle of modal verb
ADJR	comparative form of adjective	MDX	infinitive of modal verb
ADV	adverb	NCPL	noun common plural
ADVR	comparative form of adverb	NCS	noun common singular
AG	gerundive of auxiliary 'to have'	NEG	negation
AJ	present of auxiliary 'to have'	NPRPL	noun proper plural
APP	past participle of auxiliary 'to have'	NPRS	noun proper singular
AX	infinitive of auxiliary 'to have'	NUM	numeral
CMP	prep. comparison	P	preposition
CONJO	coordinative conjunction	PON	punctuation inside the clause
CONJS	subordinate conjunction	PONFP	the end of the sentence
COMP	comparative adverb	PRO	pronoun
D	determiner (indefinite, definite, demonstrative)	Q	quantifier
DAT	dative	QR	quantifier (more, less)
DF	partitive article	VG	gerundive of the main verb
DZ	possessive determiner	VJ	present of the main verb
EG	gerundive of auxiliary 'to be'	VPP	past participle of the main verb
EJ	present of auxiliary 'to be'	VX	infinitive of the main verb
EPP	past participle of auxiliary 'to be'	WADV	interr., rel. or excl. adverb
EX	infinitive of auxiliary 'to be'	WD	interr., rel. or excl. determiner
ITJ	interjection	WPRO	interr., rel. or excl. pronoun
MDG	gerundive of modal verb		

Table 1: Occitan Part-of-Speech tagset (adapted from Martineau et al. [12])

corpus of Old French [12] annotated with part-of-speech (POS) tags and syntactic constituency labels. We have adopted the POS tagset from the MCVF (see Table 1) and have trained the TnT tagger [4] on 28 265 sentences from the Medieval French section of the MCVF. This trained model was used to POS tag Old Occitan. An evaluation of the accuracy of this approach can be found in our previous work [18]. The tagger output was further manually corrected. In addition, we have augmented tokens with lemmas from the Occitan dictionary. This dictionary is based on the glossary to *Le Roman de Flamenca* [15] and consists of 2 800 entries.

For syntactic parsing, we have trained the Berkeley parser [16] using a constituency treebank from the MCVF Medieval corpus [12]. The trained model was used to parse our corpus. We have adopted syntactic labeling from the MCVF corpus. However, we have added an additional label V for verbs in order to facilitate queries. Table 2 lists the syntactic labels used in the annotation of the Old Occitan corpus.

The parsed trees were manually corrected. As a reference guide we have used the MCVF manual[7]. However, for lack of resources, and in contrast to the MCVF corpus, we did not manually add traces and empty categories. An example of the syntactic structure in our corpus is shown in Figure 2.

In addition we manually added a discourse layer that identifies different speakers in the dialogues. The labels correspond to the main characters names, namely

[7] `http://gtrc.voies.uottawa.ca/manuel/syntax-manual-fr/index.htm`

Labels	Definition	Label	Definition
ADJP	Adjectival Phrase	IP-IMP	Imperative Proposition
ADVP	Adverbial Phrase	IP-INF	Infinitival Proposition
ADVP-LOC	Adverbial Locative Phrase	IP-MAT	Main Proposition
ADVP-TMP	Adverbial Temporal Phrase	IP-PPL	Participial Proposition
CONJP	Conjunction	IP-SUB	Subordinate Proposition
CP-ADV	Adverbial Clause	NP-ACC	Direct Object
CP-ADV-TMP	Temporal Clause	NP-COM	NP Complement
CP-CAR	Prepositional Clause	NP-DTV	Indirect Object
CP-CMP	Comparative Clause	NP-PRD	Predicative NP
CP-DEG	Degree Clause	NP-RFL	Reflexive NP
CP-EXL	Exclamative Clause	NP-SBJ	Subject NP
CP-FRL	Small Clause	NP-TMP	Temporal NP
CP-OPT	Optative Clause	PP	Prepositional Phrase
CP-QUE	Interrogative Clause	PP-DIR	Prepositional Directional Phrase
CP-REL	Relative Clause	PP-LOC	Prepositional Locative Phrase
CP-THT	Complement Clause	QP	Quantifier Phrase
INTJ	Interjection	V	Verb
-LFD	Left Dislocated Phrase	-PRN	Adjunct
-SPE	Direct Speech		

Table 2: Occitan Syntactic Labels (adapted from Martineau et al. [12])

Flamenca, Archambaut, Guillem, Father. Less important characters are marked as FemaleSpeakers and MaleSpeakers. This additional information could enhance studies focusing on social variation.

Finally, the tagged and parsed texts were merged and converted to the ANNIS[8] format. ANNIS is a web-based corpus application that allows for visualization and querying of the corpus at multiple levels [21].

4 Corpus Applications

Since we are targeting two different types of users, linguists and non-linguists, with different needs, the corpus is made available in two different modes. In the first mode, the user can mainly browse the text and look up translations and glosses, in an intuitive interface (see section 4.1). In the second mode, users interested in the linguistic annotation can query the corpus for linguistic phenomena. This requires a more complex query language and thus a more complex query tool. We show such queries in section 4.2.

4.1 Textual Representation

In order to provide access to a general audience, the TEI XML data are transformed into an interactive web database[9], which allows the user to read the romance without looking at the annotations, but with access to supplemental information in-

[8] http://www.sfb632.uni-potsdam.de/annis/

[9] http://nlp.indiana.edu/~obscrivn/Introduction.html

```
((IP-MAT-SPE
   (QP (Q assaz))
   (V (MDJ podes))
   (IP-INF
      (V
          (V (VX donar))
          (CONJO e)
          (V (VX metre))))
   (PONFP ;)))
```

Eng.: “you can bestow and spend a great deal ” (line 117, [2])

Figure 2: An example for the syntactic annotation.

tended to facilitate the access to the text. The corpus is divided into Meyer’s introduction, sections of the text, glossary, and the annotated text of the romance itself. Each section of the text provides access to its French translation by Meyer.[10] Glossary definitions, comments, and footnotes are linked to tokens and are made visible when the user hovers over a marked word, as shown in Figure 3 for a glossary entry for the word *acapte*.

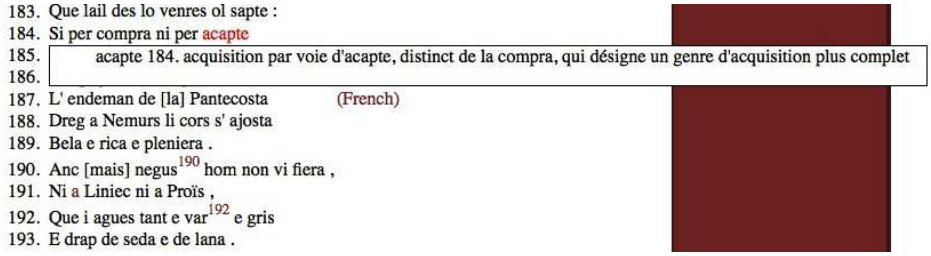


Figure 3: Glossary definition for the word *acapte*

4.2 Querying the Annotations

In order to allow queries that go beyond searching for (sequences of) words and to allow access to the annotations, we imported all the annotations into ANNIS.[11] Our web search based on ANNIS allows for basic queries, to search for a word or

[10] Although this translation (from Meyer’s first edition [14]) is sufficient to give a good idea of the content of the text, corrections to the reading of the manuscript made in later editions (including Meyer [15]) will obviously not be reflected.

[11] http://nlp.indiana.edu:8080/annis-gui-3.0.0/

phrase, and more complex queries for syntactic and morphosyntactic annotation. For example, to find all the occurrences of the word *cor* 'heart', the user can submit the query "cor" in the Search Window. The total number of occurrences will be shown in the Status Window, and the results will be displayed in the Query Result window, as shown in Figure 4.

Figure 4: Results for lexical query of the word *cor* 'heart'

The default view is in KWIC format, which displays only tokens and POS tags per line. But it is also possible to access a grid view with lemmas or a constituency tree, as shown in Figure 5.

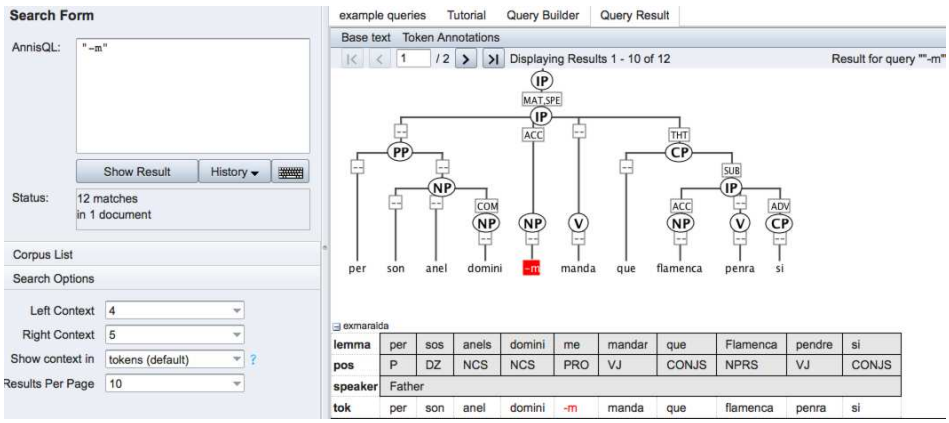


Figure 5: Grid and constituency tree visualization

ANNIS can also handle more complex queries. For instance, the query illustrated in Figure 6 shows how to search for all the cases of subjects that directly precede verbs. In this query, we describe the partial tree structure that corresponds to the phenomenon in which we are interested: We select an IP node (cat = IP) and specify a grammatical relation subject (func = SBJ) between the IP node and its NP daughter (cat = NP). To define a precedence relation between the NP and a verb (cat = V), we connect the two nodes by means of operator "." (direct precedence).

The results of queries can be exported and downloaded in plain text format with or without POS tags. The example of a result without POS tags is shown in (2a), and the example with POS tags is shown in (2b):

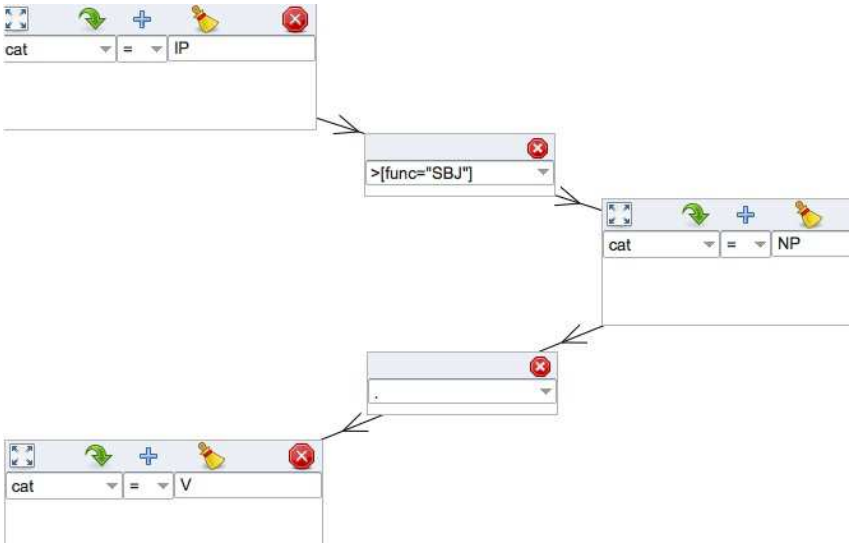


Figure 6: Example of syntactic query for a subject (NP-SBJ) preceding a verb (V)

(2) a. *le reis a dih a totz em bala :*
the king has said to all in assembly :
'The king said to the whole assembly' (line 714 [2])

b. le/D reis/NCS a/AJ dih/VPP a/DAT totz/Q em/P bala/NCS :/PONFP

In the following query, we show that it is possible to integrate different types of annotation into one query. Thus we restrict the previous search to only subject pronouns by adding a new condition, namely POS tag information. In this case, the syntactic category NP is connected to the POS tag PRO (pronoun) by means of the equality operator "_=_". The query is illustrated in Figure 7.

As a result, we find only 10 subject pronouns followed by a verb in the lines 1-798, compared to 80 in the previous query.

5 Conclusion

This paper describes an ongoing effort to digitize and annotate the corpus of *Le Roman de Flamenca*, a 13th-century romance, written in Old Occitan. In contrast to traditional corpora, this corpus is structured to fulfill two objectives. First, the web design facilitates the reading and understanding of *The Romance of Flamenca*. Words are interactively linked to the glossary, comments, and translations. Second, the corpus search design via its ANNIS interface allows for a visualization and for complex queries of the morpho-syntactic and syntactic annotations. Finally, the

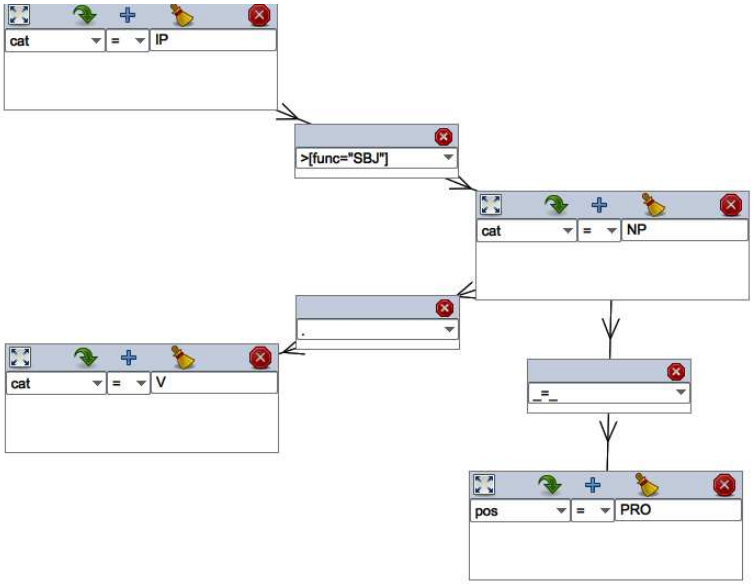


Figure 7: Example of syntactic query for a pronominal subject preceding a verb

manual correction of automatically processed text guarantees the high accuracy of the results.

In the future, we plan to augment our project with a parallel English translation [2] and to build an Old Occitan-English dictionary. In addition, we plan to add empty categories and traces to our constituency treebank, following the guidelines from the MCVF corpus [12].

6 Acknowledgements

The authors would like to thank Professor France Martineau for permission to use the MCVF corpus as a training model for our tagger and parser. Also we would like to thank Thomas Krause and Amir Zeldes for helping with the installation and configuration of ANNIS, and Michael McGuire for providing help with OCR.

References

[1] ARTFL Project. *Provençal Poetry database (American and French Research on the Treasury of the French Language), Robert Morrissey, director, with F.R. Akehurst*, 1998.

[2] E.D. Blodgett. *The Romance of Flamenca*. Garland, New York, 1995.

[3] W.A. Bradley. *The Story of Flamenca*. Harcourt Brace, New York, 1922.

[4] Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing (ANLP/NAACL)*, pages 224–231, Seattle, WA, 2000.

[5] Mark Davies. Corpus del Español: 100 million words, 1200s-1900s. Available online at `http://www.corpusdelespanol.org`, 2002.

[6] Mark Davies and Michael Ferreira. Corpus do Portugues: 45 million words, 1300s-1900s. Available online at `http://www.corpusdoportugues.org`, 2006.

[7] Anna Feldman and Jirka Hana. *A Resource-Light Approach to Morpho-Syntactic Tagging*. Rodopi, 2010.

[8] Suzanne Fleischmann. The non-lyric texts. In F.R.P. Akehurst and Judith M. Davis, editors, *A Handbook of the Troubadours*, pages 176–184. University of California Press, 1995.

[9] Ulrich Gschwind. *Le Roman de Flamenca. Nouvelle occitane du 13e siècle*, volume 2. Francke, Berne, 1976.

[10] Jean-Charles Huchet. *Flamenca. Roman Occitan du XIII siècle.* Union Générale d'Editions, Paris, 1988.

[11] René Lavaud and René Nelli. *Les Troubadours*. Paris: Desclée de Brouwer, 1960.

[12] France Martineau, Constanta Diaconescu, and Paul Hirschbühler. Le corpus 'voies du français': De l'élaboration à l'annotation. In Pierre Kunstmann and Achim Stein, editors, *Le Nouveau Corpus d'Amsterdam*, pages 121–142. Steiner, 2007.

[13] France Martineau, Paul Hirschbühler, Anthony Kroch, and Yves Charles Morin. Corpus MCVF (parsed corpus), modéliser le changement: les voies du français, Départment de Français, University of Ottawa. CD-ROM, first edition, `http://www.arts.uottawa.ca/voies/voies_fr.html`, 2010.

[14] Paul Meyer. *Le Roman de Flamenca*. Béziers, 1865.

[15] Paul Meyer. *Le Roman de Flamenca*. Librairie Emile Bouillon, 2nd edition, 1901.

[16] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual*

Meeting of the Association for Computational Linguistics, pages 433–440, Sydney, Australia, 2006.

[17] Peter T. Ricketts and Alan Reed. *Concordance de l'Occitan Médiéval. COM 2: Les Troubadours, Les Textes Narratifs en vers*. Brepols, Turnhout, 2005.

[18] Olga Scrivner and Sandra Kübler. Building an Old Occitan corpus via cross-language transfer. In *Proceedings of the First International Workshop on Language Technology for Historical Text(s)*, Vienna, Austria, 2012.

[19] Achim Stein. Syntactic annotation of Old French text corpora. *Corpus*, 7:157–161, 2008.

[20] Amir Zeldes. *ANNIS: User Guide - Version 3.0.0*. SFB 632 Information Structure / D1 Linguistic Database, Humboldt-Universität zu Berlin & Universität Potsdam, June 2013.

[21] Amir Zeldes, J. Ritz, Anke Lüdeling, and Christian Chiarcos. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics*, Liverpool, UK, 2009.