Building An Old Occitan Corpus via Cross-Language Transfer

Olga Scrivner
Indiana University
Bloomington, IN, USA
obscrivn@indiana.edu

Sandra Kübler
Indiana University
Bloomington, IN, USA
skuebler@indiana.edu

Abstract

This paper describes the implementation of a resource-light approach, cross-language transfer, to build and annotate a historical corpus for Old Occitan. Our approach transfers morpho-syntactic and syntactic annotation from resource-rich source languages, Old French and Catalan, to a genetically related target language, Old Occitan. The present corpus consists of three sub-corpora in XML format: 1) raw text; 2) part-of-speech tagged text; and 3) syntactically annotated text.

1 Introduction

In the past decade, a number of annotated corpora have been developed for Medieval Romance languages, namely corpora of Old Spanish (Davies, 2002), Old Portuguese (Davies and Ferreira, 2006), and Old French (Stein, 2008; Martineau, 2010). However, annotated data are still sparse for less-common languages, such as Old Occitan. For example, the only available electronic database "The Concordance of Medieval Occitan"[1], published in 2001, is not free and is limited to lexical search.

While the majority of historical corpora are built by means of specific tools developed for each language and project, such as TWIC for NCA (Nouveau Corpus d'Amsterdam) (Stein, 2008), or a probabilistic parser for the GTRC project (MCVF Corpus) (Martineau et al., 2007), the goal of this project is to implement a resource-light approach, exploiting existing resources and common characteristics shared by Romance languages.

It is well known that Romance languages share many lexical and syntactic properties. The following example illustrates the similarity in word order and lexicon of Old Occitan, Old Catalan, and Old French[2]:

(1) *Oc:* *Dedins* *la* *cambra* *son*
Cat: Dins la cambra són
Fr: Dans la chambre elles sont
vengudas, *dejosta* *lui* *son*
vingudas, davant ell són
entrées, devant lui elles se sont
assegudas.
assegudas.
assises.
'They came into the room, they sat down next to him.'

Several recent experiments have demonstrated that genetically related languages can share their knowledge through cross-language transfer (Hana et al., 2006; Feldman and Hana, 2010) between closely related languages. That is, a resource-rich language can be used to process unannotated data in other genetically related languages. For example, Hana et al. (2006) have used Spanish morphological and lexical data for automatic tagging of Brazilian Portuguese. While the idea of cross-language transfer is not new and

[1] http://www.digento.de/titel/100553.html

[2] Catalan and French translation is ours. We approximated Old Catalan and Old French word order as far as possible.

is mainly used with parallel corpora and large bilingual lexicons (Yarowsky and Ngai, 2001; Hwa et al., 2005), experiments by Hana et al. (2006) have demonstrated the usability of this method in situations where there are no parallel corpora but instead resources for a closely related language.

Our goal is to build a corpus of Old Occitan in a resource-light manner, by using cross-language transfer. This approach will be used not only for part-of-speech tagging, but also for syntactic annotation.

The organization of the remainder of the paper is as follows: Section 2 provides a brief description of Old Occitan. Section 3 reviews the concept of a resource-light approach in corpus linguistics. Section 4 provides details on corpus pre-processing. The methods for cross-linguistic part-of-speech (POS) tagging and cross-linguistic parsing are described in Sections 5 and 6. Finally, the conclusions and directions for further work are presented in Section 7.

2 Old Occitan (Provençal)

Occitan, often referred to as Provençal, constitutes an important element of the literary, linguistic, and cultural heritage in the history of Romance languages. Provençal (Occitan) poetry was a predecessor of French lyrics. Moreover, Occitan was the only administrative language in Medieval France, besides Latin (Belasco, 1990). While the historical importance of this language is indisputable, Occitan, as a language, remains linguistically understudied. Compared to Old French, Provençal is still lacking digitized copies of scanned manuscripts, as well as annotated corpora for morpho-syntactic or syntactic research.

Typologically, Old Occitan is classified as one of the Gallo-Roman languages, together with French and Catalan (Bec, 1973). If one examines Old Occitan, Old French, and Old Catalan, on the one hand, it is striking how many lexical and morphological characteristics these languages share. For example, French and Occitan have rich verbal inflection and a two-case nominal system (nominative and accusative), illustrated with an example of the word 'wall' in (2):

Figure 1: Linguistic map of France, from (Bec, 1973)

	Case	Old Occitan	Old French
(2)	Nominative	lo murs	li murs
	Accusative	lo mur	lo mur

On the other hand, Occitan has syntactic traits similar to Catalan, such as a relatively free word order and null subjects, illustrated in (3) and (4).

(3) *Gran honor nos fai*
great honor us$_{Dat}$ does
'He grants us a great honor' (Old Occitan - Flamenca)

(4) *molt he gran desig*
much have big desire
'I have a lot of great desire...' (Old Catalan - Ramon Llull)[3]

The close relationship between these three languages is also marked geographically. The northern border of the Occitan-speaking area is adjacent to the French linguistic domain, whereas in the south, Occitan borders on the Catalan-speaking area, as shown in Figure 1.

This project focuses on the 13th century Old Occitan romance "Flamenca", from the edition by Meyer (1901). Apart from a very intriguing fable of beautiful Flamenca imprisoned in a tower by her jealous husband, this story presents a

[3] http://orbita.bib.ub.edu/ramon/velec.asp

very interesting linguistic document consisting of 8097 lines of the "universally acknowledged masterpiece of Old Occitan narrative" (Fleischmann, 1995). Multiple styles, such as internal monologues, dialogues and narratives, provide a rich lexical, morphological and syntactic database of a language spoken in southern France.

3 Linguistic Annotation via Cross-Language Transfer

Corpus-based approaches often require a great amount of parallel data or manual labor. In contrast, the cross-language transfer, as proposed by Hana et al. (2006), is a resource-light approach. That is, this method does not involve any resources in the target language, neither training data, a large lexicon, nor time-consuming manual annotation.

While cross-language transfer has been previously applied to languages with parallel corpora and bilingual lexica (Yarowsky and Ngai, 2001; Hwa et al., 2005), Hana et al. (2006) introduced a method in the area where these additional resources are not available. Feldman and Hana (2010) performed several experiments with Romance and Slavic languages. The only resources they used were i) POS tagged data of the source language, ii) raw data in the target language, and iii) a resource-light morphological analyzer for the target language. For POS tagging, the Markov model tagger TnT (Brants, 2000) was trained on a source language, namely Spanish and Czech, in order to obtain transition probabilities. The reasoning is that the word order patterns of source and target languages are very similar so that given the same tagset, the transition probabilities should be similar, too. Since the languages differ in their morphological characteristics, a direct transfer of the lexical probabilities was not possible. Instead, a shallow morphological analyzer was developed for the target languages, using cognate information, among other similarities. The trained models were then applied to the target languages, Portuguese, Catalan, and Russian. Tagging accuracies for Catalan, Portuguese, and Russian yielded 70.7%, 77.2%, and 78.6% respectively.

In contrast, syntactic transfer is mainly used in machine translation. This approach requires a bilingual corpus aligned on a sentence level. That is, words in a source language are mapped to words in a target language. Dien et al. (2004) used this method on English-Vietnamese corpus. They extracted a syntactic tree set from English and transferred it into the target language. Obtained two sets of parsed trees, English and Vietnamese, were further used as training data to extract transfer rules. In contrast, Hanneman et al. (2009) extracted unique grammar rules from English-French parallel parsed corpus and selected high-frequency rules to reorder position of constituents. Hwa et al. (2005) describe an approach that focuses on syntax projection per se, but their approach also relies on word alignment in a parallel corpus. They show that the approach works better for closely related languages (English to Spanish) than for languages as different as English and Chinese. It is widely agreed that word alignment and thus, syntactic transfer, is best applied in similar languages due to their word order pattern (Watanabe and Sumita, 2003). Therefore, genetically related Romance languages should be well suited for syntactic cross language transfer. McDonald et al. (2011) and Naseem et al. (2012) describe novel approaches that use more than one source language, reaching results similar to those of a supervised parser for the source language.

While cross-language transfer has been applied successfully to modern languages, we decided to use it to transfer linguistic annotation to a historical corpus. The choice of source languages was based on the availability of annotated resources and the similarity of language characteristics. Thus, Old French corpus (Martineau et al., 2007) was selected as a source for the morpho-syntactic annotation of Occitan. However, to transfer syntactic information, we used the Catalan dependency treebank (Civit et al., 2006) since modern Catalan displays a pro-drop feature and a relatively free word order, similarly to Old Occitan.

4 Corpus Pre-Processing

The romance 'Flamenca' is available in scanned images format, therefore, the initial step included conversion to an electronic version via OCR and manual correction. Figure 2 shows a sample of the manuscript.

Corrections by the editor were omitted. As can be seen in the first line of the document (see Figure 2), the editor (Meyer, 1901) enclosed a silent letter 's' in brackets which we have excluded from our pre-processed text. In addition, we detached the clitic pronouns that are joined to the verbs, as in (5), as shown in (6).

(5) *Per son anel domini**m*** *manda Que*
for his ring coat of arms sends that
*Flamenca penra si**m*** *voil.*
Flamenca takes if me want
'He is sending his family ring as a guarantee that, if I want, he will marry Flamenca'

(6) Per son anel domini **m** manda Que
Flamenca penra si **m** voil.

« Poissas lur di[s] tot en apert : (fol. 2)
« Vostre cor nom tengas cubert,
« Mais digas mi : si Dieus mi dona
4 « Un'aventura que m'es bona,
« Non sabra bon a totz ensems?
« Ieu ai desirat mout lonc temps
« C'ap N'Archimbaut agues paria,
8 « 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.

Figure 2: Sample of page from 'Flamenca'

Currently, we have pre-processed and formatted 3 095 lines, which corresponds to 12 573 tokens. The text file was then converted to XML format using EXMARaLDA[4]. While EXMARaLDA is mostly used for transcriptions, it also imports files from several formats, such as plain text or tab format, and exports them as EXMARaLDA XML files. This XML is timeline-based and supports the annotation of different linguistic levels in different tiers.

5 Cross-Linguistic POS Tagging

Since Old French and Old Occitan share many morphological features, we have adopted the POS tagset from the MCVF corpus of Old French (Martineau et al., 2007). The MCVF tagset is based on the annotation scheme of the Penn-Helsinki Parsed Corpus of Middle English (PPCME) (Kroch and Taylor, 2000), which was modified to represent French morpho-syntactically, as illustrated in (7).

(7) *Les/D* *petites/ADJ* *filles/NCPL*
the/Det. little/Adj. girls/Noun-Cmn-pl.
'the little girls'

While the MCVF tagset consists of 55 tags, we have decreased the tagset to 39 tags for our corpus. The Occitan tagset is shown in Table 1. The simplification included joining certain subclasses into one class. The reason for this modification lies in the particularities of Occitan. First, as a pro-drop language, Occitan omits an impersonal pronoun (8), in contrast to Old French (9).

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

Table 1: Occitan tagset

[4] www.exmaralda.org

(8) *No m' o cal dir*
Not me it$_{Acc}$ must say
'it is not necessary for me to tell this'

(9) *Que te faut il en ce*
what you$_{Acc}$ must it$_{impersonal}$ in this
pais?
country
'What do you need in this country?'
(MCVF corpus)

Furthermore, the grammar of Old Occitan (Anglade, 1921) does not use "near future" tense, which is common in Old French and is formed by the verb *aller* 'to go' and the infinitive of a main verb (10). Therefore, the specific labels LJ, LX, LPP for the auxiliary *aller* from the corpus of Old French are mapped to the corresponding tags of the main verb, such as VJ, VX, VPP, VG (see Table 1).

(10) *Et que iroi ge faire?*
and what will go I do
'What am I going to do?'

Finally, the French tagset contains a label FP for focus particle, such as *seullement/seulement* and *ne...que* 'only' (11). The following comparison shows that the latter construction *ne..que* does not convey focus in Old Occitan (12):

(11) *le jeune homme* ***ne*** *le fait* ***que***
the young man not it does only
pour l' avarice
for the greed
' Young man does it **only** for greed' (MCVF corpus)

(12) *Don* ***non*** *cug* ***que*** *ja mais reveinha*
of it not think that ever returns
'I do **not** think **that** he ever recovers from it'

We trained TnT on 28 265 sentences from the Medieval French texts (MCVF). This trained model was used to POS tag Old Occitan without any modification to the lexicon. For the performance evaluation we extracted 50 sentences (1000 tokens) from the corpus and annotated them manually. Then, the tagger output was compared to the gold standard. The POS tagger reached an accuracy of 64.00% for all words and 72.64% for known words, cf. Table 2.

Accuracy	N	%
All Words	640/1000	64.00
Known Words	539/742	72.64
Unknown Words	101/258	39.15

Table 2: POS evaluation using the unmodified Medieval French lexicon

A manual analysis of randomly selected 20 sentences revealed that a number of errors are caused by lexical duplicates that have different meanings in each language. For example, in (13) *no* is a possessive determiner 'our' in Old French, while in Occitan *no* is a negation. The second type of errors is the result of TnT's algorithm for handling unknown words by a suffix trie. That is, unknown words are assigned to an ambiguity class depending on their suffix. For example, the unknown Occitan word *ancar* 'yet' is recognized as an infinitive (VX), based on the ending -ar which is common for French infinitives; whereas *entremes* 'involves' receives higher probability as an adjective (ADJ) because of its ending -es (see (13)) .

(13)

	Ancar	*d'*	*amor*	*no*	*s'*
TnT:	VX	P	NCS	DZ	ADV
gold:	ADV	P	NCS	NEG	PRO

entremes
ADJ
VJ
'he is not yet involved in the love affair'

Therefore, to improve the fit of the lexicon extracted from Medieval French with words specific to Occitan, we added 171 manually annotated sentences from 'Flamenca' to the training data. The validation on the test set yielded 78.10% accuracy for all words, 81.10% for known, and 57.48% for unknown words, see Table 3. The results prove that adding even a small set of high quality, manually annotated sentences in the target languages improves POS tagging quality considerably, bringing the tagger's performance close to results reached for modern languages (given

Accuracy	N	%
All Words	781/1000	78.10
Known Words	708/873	81.10
Unknown Words	73/127	57.48

Table 3: Evaluation with the Occitan-enriched lexicon

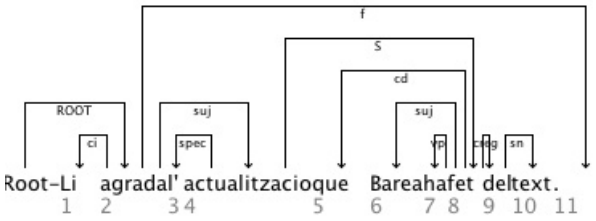

Figure 3: Example of Dependency Relation from the Catalan Treebank (Ancora)

ID	Word	Lemma	Pos	Head	Dep
1	Li	ell	PP3CSD00	2	ci
2	agrada	agradar	VMIP3S0	0	sent
3	l'	el	DA0CS0	4	spec
4	actualització	actualització	NCFS000	2	suj
5	que	que	PR0CC000	8	cd
6	Barea	Barea	NP00000	8	suj
7	ha	haver	VAIP3S0	8	v
8	fet	fer	VMP00SM	4	S
9	del	del	SPCMS	8	creg
10	text	text	NCMS000	9	sn
11	.	.	Fp	2	f

Table 4: Annotation in AnCora

a morphologically richer language and a small training set), thus validating our approach.

Furthermore, in the course of our project we have compiled an Occitan dictionary, consisting of 2 800 entries. The glossary to Flamenca (Meyer, 1901) was used as a reference guide. Thus, our tagged corpus was augmented with lemmas from the dictionary. We have further manually checked and corrected 8 284 tags in the corpus. Finally, the POS tagged version was converted to XML, again using EXMARaLDA.

6 Syntactic Cross-Linguistic Parsing

While it has been widely accepted that syntactic annotation in terms of constituent trees provides a rich internal tree structure, recent years have shown an increased interest in dependency graphs (Civit et al., 2006). Dependency graphs provide an immediate access to lexical information for words, word pairs, and their grammatical relations. For example, each word in the Catalan sentence (14) has exactly one head, as demonstrated in Figure 3. The arcs show dependencies from heads to dependents.

(14) *Li* *agrada* *l'actalització* *que*
to him likes the modernisation that
Barea *ha* *fet* *del* *text.*
Barea has made of the text.
'He likes how Barea modernized the text'.

In addition, it is argued that dependency grammars "deal especially well with languages involving relatively free word order" (Bamman and Crane, 2011). Since Old Occitan has relatively free word order, we aim for a syntactic annotation in form of dependency graphs. At present, dependency treebanks are available only for modern Romance languages, namely French (Abeillé et al., 2003), Catalan, and Spanish (AnCora), (Civit et al., 2006). For training a dependency parser, we use Catalan rather than modern French since the syntactic characteristics of Catalan are more similar to Occitan than French. For example, Old Occitan is a pro-drop language while French is not. We used the Catalan treebank from AnCora[5].

The Catalan treebank consists of 16 591 sentences extracted from newspapers and annotated syntactically and semantically. The dependency treebank has been converted automatically from a constituency format with the help of a table of head finding rules (Civit et al., 2006). The sentence in (14), for example, is annotated in the treebank format as shown in Table 4.

As shown in Table 4, the Catalan tagset describes information about the major POS classes, represented as letters, and morphological features, such as gender, number, case, person, time and mode, represented as digits. The tagset has a total of 280 different labels (Taulé et al., 2008). The rich morphological information allowed us to map the Catalan tags to our Occitan tagset with high accuracy. For example, in the sentence from

[5] http://clic.ub.edu/corpus/en/ancora-descarregues

Table 4, verbal features such as VMIP (main verb indicative present) and VMP (main verb past participle) were mapped to Occitan tags VJ and VPP, respectively. The example of the mapped Catalan tags is shown in (15).

(15) *Li* *agrada* *l'actalització* *que*
PRO VJ NCS WPRO
Barea *ha* *fet* *del* *text* .
NPC AJ VPP P NCS PONFP
'He likes how Barea modernized the text'.

The Catalan dependency representation contains a large set of grammatical relations. We found 48 different labels in the AnCora Corpus. We decided to map these dependencies to the Penn Treebank core dependencies, such as subject (SBJ), direct object (OBJ), predicate (PRED), nominal modifier (NMOD), verbal modifier (VMOD). In addition, we added a language specific relation - CL (clitic) (16). The complete list of Occitan dependency labels and their corresponding labels in Catalan is shown in Table 5.

(16) *Pero* *vostre* *sen* ***m'*** ***en*** *digas*
but your opinion me about it tell
'But tell me you opinion about it.'

It is necessary to note that verbal head selection in the AnCora Corpus differs from the constituency head assignment (Civit et al., 2006). In the dependency annotation, the head is assigned to the rigthmost element in the verbal phrase. For example, past participles, and gerunds are heads, whereas auxiliaries are dependents. In addition, the treebank is automatically augmented by empty elements to represent null subjects.

In order to annotate our Old Occitan texts, we trained a transition-based dependency parser, MaltParser (Nivre et al., 2007) on the Catalan treebank with the reduced tagset. We then used the trained model to parse our corpus.

We manually annotated 30 sentences to evaluate the accuracy of the parser. For the evaluation we used MaltEval[6], an evaluation tool for dependency trees. The results yielded 63.1% of label accuracy and 55.8% of labeled attachment. The highest score of precision and recall was for nominal modifiers, prepositional modifiers, clitic, negation and punctuation, cf. Table 6.

As can be seen from Table 6, subject, object and predicate relations are the least accurate. This is due to a relatively free word order in Old Occi-

[6] http://w3.msi.vxu.se/users/jni/malteval/

Occitan	Relation type	Catalan
ROOT	Main clause	Sentence
S(bar)	Infinitival sentence	infinitiu
	Subordinate complement	ao
SBJ	Subject	suj
OBJ	Nominal Direct Object	cd
CL	Pronominal direct object	cd
	Pronominal morpheme	morfema.pron
	Impersonal	impers
	Passive 'se'	pass
PRED	Predicative	cpred
	Atribute	atr
NMOD	Determiner	d
	Numbers	z
	Prepositional phrase	sp
	Adjectival phrase	s.a
	Determiner	spec
	Second adj inside sn	grup.a
VMOD	Prepositional complement	creg
	Adverbial phrase	sadv
	Verb adjunct	cc
	Indirect object	ci
PMOD	Noun phrase	sn
	Adverbial modifier	r
	Second noun phrase inside sp	grup.nom
	Direct object	ci
V	Auxilliary	v
NEG	Verb modifier 'no'	mod
CONJ	Conjunction que	conj
COORD	Coordination	c
REL	Relative pronoun	relatiu
INC	Inserted phrase	inc
P	Punctuation	f

Table 5: Dependency Labels

Deprel	precision	recall
P	91.2	86.7
CL	89.3	78.1
NEG	77.8	77.8
NMOD	86.1	87.3
PMOD	76.9	78.9
PRED	42.9	16.7
ROOT	44.8	78.9
OBJ	42.3	34.4
S	45.8	16.2
SBJ	37.9	55.0
V	53.3	57.1
VMOD	46.2	56.3

Table 6: Parsing results

```
$<$sentence id="19" user="" date=""$>$
     $<$word id="1" form="Le" postag="D" head="2" deprel="NMOD"/$>$
     $<$word id="2" form="coms" postag="NCS" head="3" deprel="SBJ"/$>$
     $<$word id="3" form="fes" postag="VJ" head="0" deprel="ROOT"/$>$
     $<$word id="4" form="sa" postag="DZ" head="5" deprel="NMOD"/$>$
     $<$word id="5" form="mollier" postag="NCS" head="3" deprel="OBJ"/$>$
     $<$word id="6" form="venir" postag="VX" head="5" deprel="S"/$>$
     $<$word id="7" form="." postag="PONFP" head="3" deprel="P"/$>$
$<$/sentence$>$
```

Figure 4: An example of the syntactic annotation

tan and its pro-drop feature. The example in (17) illustrates that in some cases, the noun, here *Flamenca*, can be ambiguous between subject or object of the sentence.

(17) *Que Flamenca penra*
that Flamenca will take
'that Flamenca will take' / 'that he will take Flamenca'

In contrast, the low precision of ROOT is due to MaltParser's design which may lead to incomplete syntactic annotations. To repair this type of errors, we performed post-editing, which readjusts ROOT labels, increasing its accuracy to 71%, instead of 44.8%. In addition, it yielded better results for label accuracy - 76.4% and label attachment score - 63.4%.

Finally, the parsed output was converted to XML format using Malt Converter[7]. An example of the resulting XML file is presented in Figure 4.

7 Conclusion and Future Work

While the annotation of historical corpora requires precision, the first steps in building a syntactically annotated corpus can be resource-light. If we use a cross-language transfer, we can profit from existing resources for historical or modern languages sharing similar morphological and syntactic features. We have shown that Old French presents a good source for POS tagging of Old Occitan while modern Catalan is a good source for the syntactic annotation of Occitan.

The POS tagger model trained on the Old French Corpus MCVF (Martineau et al., 2007) yielded 78% accuracy when we added a small Occitan lexicon into training, whereas dependency parsing results yielded 63.4% of labeled attachment.

For the future, we plan to annotate the whole corpus of 8 000 lines, manually correct them, and make it available on the web. At present, 3 095 lines of tagged and parsed XML files are available upon request.

[7] http://w3.msi.vxu.se/~nivre/research/MaltXML.html

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In *Treebanks*. Kluwer.

Joseph Anglade. 1921. *Grammaire de l'ancien provençal ou ancienne langue d'oc*. Librairie C. Klincksieck.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 79–98. Springer.

Pierre Bec. 1973. *La langue occitane*. Number 1059 in Que sais-je? Paris: Presses universitaires de France.

Simon Belasco. 1990. France's rich relation: The Oc connection. *The French Review*, 63(6):996–1013.

Thorsten Brants. 2000. 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.

Monserat Civit, Antònia Martí, and Nuria Bufí. 2006. Cat3LB and Cast3LB: From constituents to dependencies. In *Advances in Natural Language Processing*, pages 141–153. Springer.

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

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

Dinh Dien, Thuy Ngan, Xuan Quang, and Chi Nam. 2004. The parallel corpus approach to building the syntactic tree transfer set in the English-to-Vietnamese machine translation. In *2004 International Conference on Electronics, Information, and Communications (ICEIC2004)*, pages 382–386, Ha Noi, Vietnam.

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

Suzanne Fleischmann. 1995. 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.

Jirka Hana, Anna Feldman, Chris Brew, and Luiz Amaral. 2006. Tagging Portuguese with a Spanish tagger using cognates. In *Proceedings of the EACL Workshop on Cross-Language Knowledge Induction*, pages 33–40, Trento, Italy.

Greg Hanneman, Vamshi Ambati, Jonathan H. Clark, Alok Parlikar, and Alon Lavie. 2009. An improved statistical transfer system for French–English machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 140–144.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Anthony Kroch and Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania.

France Martineau, Constanta Diaconescu, and Paul Hirschbühler. 2007. 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.

France Martineau. 2010. Corpus MCVF, modéliser le changement: les voies du français. `http://www.arts.uottawa.ca/voies/voies_fr.html`.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, UK.

Paul Meyer. 1901. *Le Roman de Flamenca*. Librarie Emile Bouillon.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637, Jeju Island, Korea.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

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

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 3405–3408, Marrakech, Morocco.

Taro Watanabe and Eiichiro Sumita. 2003. Example-based decoding for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 9–12, Boston, MA.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, PA.