Tagging Occitan using French and Castillan Tree Tagger

Marianne Vergez-Couret

Tagging Occitan using French and Castillan Tree Tagger

Vergez-Couret Marianne

University of Toulouse Le Mirail
5 allées Antonio Machado, Toulouse, France
{vergez}@univ-tlse2.fr

Abstract

Part-Of-Speech (POS) tagging, including tokenization and sentence splitting, is the first step in all Natural Language Processing chain. It usually requires substantial efforts to annotate corpora and produce lexicons. However, when these language resources are missing like in Occitan, rather than concentrate the effort in creating them, methods are settled to adapt existing rich-resourced languages tagger. For this to work, these methods exploit the etymologic proximity of the under-resourced language and a rich-resourced language. In this article, we focus on Occitan, which shares similarities with several romance languages including French and Castillan. The method consists in running existing morpho-syntactic tools, here Tree Tagger, on Occitan texts with first a translation of the frequent words in a rich-resourced language. We performed two distinct experimentations, one exploiting similarities between Occitan and French and the second exploiting similarities between Occitan and Castillan. This method only requires the listing of the 300 most frequent words (based on corpus) to construct two bilingual lexicons (Occitan/French and Occitan/Castillan). Our results are better than those obtained with the Apertium tagger using a larger lexicon.

Keywords: less ressourced language (LRL), POS tagging, Tree Tagger, Occitan, dialect, language resources, language technologies.

1. Introduction

Occitan is a romance language spoken in southern France and in several valleys of Spain and Italy. It is written since the middle age and a very important literature has been produced. The BaTelÒc project (Bras, 2006; Bras and Thomas, 2011; Bras and Vergez-Couret, 2013) aims at creating wide coverage text collections by gathering written texts of literature (prose, drama and poetry) and other genres such as technical texts and newspapers, for modern and contemporary periods. More than one million words have already been gathered. The text base is also designed to provide online tools for interrogating texts, for example a concordancer to observe key forms in context. In the future, the aim is to enrich the text base with linguistic annotations. In this paper, we focus on Part-Of-Speech (hereafter POS) annotations and tools for annotate them automatically. Within the framework of the text base, POS would allow new possibilities of request, for example the disambiguation of homographs such as *poder* (common noun *power*) and *poder* (verb *can*):

1) *Fau lo polit per [poder $_{verb}$] far lo gòrre (Los croquants de Roergue,F. Delèris).*
2) *Lo Prince de las tenèbras al siu servici : l'argent, lo [poder $_{common\ noun}$], las onors, la capitada… (L'estilò negre, B. Bergé.)*

POS tagging is the first step in all Natural Language Processing chain. It usually requires substantial efforts to annotate corpora and produce lexicons. But these resources are missing for Occitan. Creating them is more subsequent since languages with various dialects present spelling and dialectal variations and these languages are not necessarily standardized. As a consequence, direct translation of existing models for resourced-rich languages is difficult and rather more compromised since there are not enough annotated data and structured lexicon for considering supervised learning. For similar cases, some works present systems to bypass the need of annotated data and lexicons. We followed on Hana *et al.* (2011) and Bernhard and Ligozat (2013) who used methods that exploit the etymologic proximity between an under-resourced language and a rich-resourced language. We performed two distinct experimentations, one exploiting similarities between Occitan and French and the second exploiting similarities between Occitan and Castillan.

The paper is organized as follows: Section 2 gives additional information on Occitan, its spelling and dialectal variations and the impact of these variations on Apertium which includes a POS tagger for Occitan. Section 3 presents some works on the development of language technologies when language resources are missing. Finally, section 4 is dedicated to the presentation of our two experiments.

2. Occitan language

Occitan language belongs to romance languages. The number of speakers, in France, several valleys of Spain and Italy is hard to estimate: according to several studies it can be evaluated between 600,000 to 2,000,000 speakers (Sibille, 2007).
Occitan is not a unitary language and is not standardized as a whole. It has several varieties organized in dialects. The most accepted classification suggested by Bec (1995) includes Auvergnat, Gascon, Languedocien, Limousin, Provencal and Vivaro-Alpin.

2.1. Written Occitan

Occitan is written since the Middle-Age. The spelling used at that time is called the “troubadour spelling”. This spelling disappeared gradually with the decline of the literary production. Since the 19th century, one can distinguish two major types of spellings, the first ones was influenced by the French spelling, such as the Mistral's spelling, created in Provence and the Gaston Febus's spelling used in Bearn. The second type appeared during the 20th century. It is a unified spelling, said "classical spelling", inspired from the “troubadour

spelling" and diffused in all Occitan territories (Sibille, 2007).
So, the existence of numerous spellings is one cause of variation. Another cause is the dialectal state of the language. The classical spelling naturally integrates the geo-linguistic varieties (for instance *lo filh* vs. *eth hilh* ; *luna* vs. *lua* and *cabra* vs. *craba* (Bec, 1995)). Variations are also due to the fact that the effort of spelling normalisation is in progress (evolution for the spelling of conjugate verbs: *avian* vs. *avián*). And finally there are also phonological intra-dialectal variations (for example *contes* vs. *condes*).
Because of the spelling and dialectal variations, it is difficult to simply apply the existing system of POS tagging by creating annotated corpora and large coverage lexicons as it is currently done with rich-resourced standardized languages. We present in the next section Apertium (Forcada *et al*, 2011) which includes a POS tagger for Occitan based on a lexicon and raise some problems about this system.

2.2. POS tagging Occitan with Apertium

Apertium originally proposes open source systems for automatic translation, generally for related-language pairs. Armentano I Oller (2008) developed a translation system for the Occitan/Catalan pair, which includes a POS tagger for Occitan. It is based on the use of one lexicon containing 36 500 entries. Two difficulties can be raised:

a) If a word is not in the lexicon, no tag is proposed for this word. This results in important variations of performances from a text to another one, especially because all the possible spelling forms for all the dialects do not occur in the lexicon. Armentano I Oller announces an accuracy of 0.8 of correct tags for a text in Languedocien. We made the same experiment on a text in Gascon. We reached an accuracy of 0.6 of correct tags. Indeed more words were unknown in Gascon (19%) than in Languedocian (13 %).

b) The lexicon includes indiscriminately forms from various dialects. It would be required to evaluate if it is better to have one larger lexicon for all dialects or on the contrary one lexicon for each dialect.

Finally, to improve the performance of Apertium, the only way is to enrich the lexicon which is very time consuming. To cope with this kind of problem, some researchers develop system requiring a minimum of lexical resources.

3. POS tagging other less-resourced languages

For less-resourced languages, the main circumvention strategy is to use existing systems for a rich-resourced etymologically close language: Hana *et al* (2011) use the proximity between Old Czech and Modern Czech and Bernhard and Ligozat (2013) exploit the similarities between Alsatian and German.
Hana *et al* (2011) present two different strategies. The first one consists in transforming a text in Old Czech to make it looks like an approximate text in Modern Czech with a finite number of spelling changes rules. The data produced are then used to train a POS tagger for Old Czech. The text is finally restored in its original form and the result is an Old Czech text annotated in POS tagging. The second strategy consists in creating a lexical resource as much automatically as possible using resource-light morphological analyzers (which is based on word endings). This method permits to reduce the manual annotation to a list of 250 frequent words, manually analyzed with their possible tags.
Bernhard and Ligozat (2013) adopt the same idea, using the proximity between Alsatian and German. The relation between Alsatian and German is different than the one between Old and Modern Czech (two successive state of the language). Alsatian is considered as a German dialect. They propose to directly use the morpho-syntactic tools existing for German (Tree Tagger (Schmid, 1994) and the Stanford POS Tagger (Toutanova *et al*, 2003)) on Alsatian texts with pre-translations of grammatical words (articles, pronouns, prepositions and conjunctions) and auxiliaries. They obtain an accuracy of 0.8 with Tree Tagger and 0.86 with Stanford POS Tagger. We propose to adapt these two methods for Occitan and two etymologically related languages French and Castillan using the Tree Tagger software.

4. Experimental validations

4.1. Motivations

This paper describes experiments aiming at creating morpho-syntactic resources for Occitan, exploiting the similarities between Occitan and French and Castillan. The relation between the three languages is looser than the one between German and Alsatian. Occitan, French and Castillan belong to romance languages and are then etymologically related. We focus on Gascon dialect for which we have a corpus for evaluation (see section 2.2.).

4.1.1 Evaluation corpus

We extract and annotate with POS 1024 words (104 sentences) from a Gascon novel. The POS annotation has been done manually with a simplified tag set inspired by the one used in Frantext[1], an online text base with POS annotations and tools for French. The tag set is given in Table 1. The manual annotation was done by revising the Apertium POS tag (see section 2.2.). This annotated corpus is our Gold standard. An extract is given Table 2.

A	Adjective (except Ap)
Ap	Possessive adjective
Adv	Adverb
Cc	Coordination conjunction
Cs	Subordination conjunction
D	Article (except Dg)
Dg	Amalgamated article
Dca	Cardinal number as article
Pe	Enonciative particle
V	Finite verb (except Vi, Vpp, Vps)
Vi	Infinite verb
Vpp	Present participle
Vps	Past participle
Inj	Interjection

[1] www.frantext.fr

Np	Proper nom
S	Common noun
P	Pronoun
Pp	Preposition
Pr	Relative pronoun
<sent>	End of sentence
Cm	Comma

Table 1. Tag set

Que	Pe	d'	Pp
's	P	ompras	S
pòblan	V	darrèr	Pp
de	Pp	las	D
babaus	S	travetas	S
las	D	.	<sent>
plapas	S		

Table 2. Extract from the Gold standard

4.1.2. Tree Tagger Software

We choose Tree Tagger software (hereafter TT), training for both French and Castillan. The TT's performance relies on the use of a large coverage lexicon. But unlike Apertium, TT predicts tags, using the probability of POS tag sequences calculated on a manually tagged training corpus.

We assume that probabilities of POS tag sequences will be fairly similar between Occitan and French and Occitan and Castillan. Examples below show specificities in Gascon which may influence the word order (Massourre, 2012). We then indicate if yes or no these specificities are shared with the two related languages French and Castillan.

- Occitan is a pro-drop language (subject pronoun-dropping), as in Castillan (see example 3). The word order in Gascon and Castillan is then less limited than in French. For example subject may be more often placed after the verb in Occitan and Castillan
- Partitive article and plural indefinite article are also absent, as in Castillan (see example 3).

  3) [Que $_{Pe}$] [crompi $_{V}$] [pans $_{S}$] [e $_{C}$] [pomas $_{S}$].
  I buy bread and apples.
- Object pronouns can occur before or after the verb (see example 4 where it occurs after). In French, object pronouns always occur before the verb in affirmative sentences. In Castillan, the two places are possible but some object pronouns such as *ne* and *i* do not have direct translation as a pronoun.

  *4) [Non $_{Adv}$] [sembla $_{V}$] [pas $_{Adv}$] [avisar $_{Vi}$] **[-se $_{P}$]** [deu $_{Dg}$] [men $_{Ap}$] [estat $_{S}$] ?*
  He does not seem to find out about my condition.
- Gascon has enonciative particles for affirmative sentences (Que see example 3), exclamative sentences (Be see example 5) and interrogative sentences (E). There are no equivalent in both French and Castillan.
- The possessive adjectives in Gascon may be preceded by a definite article (*la mia (my)*) and it can be placed before or after the noun. This is not the case in both French and Castillan.
- Relative clauses start with the combination of a preposition and a relative pronoun (see example 5) although in French and Castillan prepositions are not allowed in front of relative pronouns.

  *5) Be diserem [**de $_{Pp}$**] [**que $_{P}$**] drom !*
  It seems that he sleeps !

With respect to all romance languages, it is well know that French is the most notable exception mostly because it is not a pro-drop language. It seems that the syntax of Gascon is more similar to Castillan probably because it has been historically strongly influenced by languages from North Spain (Massourre, 2012).

The question is now how to adapt the French and Castillan TT to tag Occitan. We use the method of translation described by Bernhard and Ligozat (2013) that only requires a small lexicon of translated words.

4.2. Methods

The two experiments using French and Castillan resources are based on the same methodology.

4.2.1. Lexicon creation

The lexicon is built on ad-hoc criteria. It is a trilingual lexicon, composed of the 300 most frequent words extracted from one novel in Gascon with their translation in French and Castillan. An example is given Table 3.

Gascon	French	Castillan
deu	Du	del
devath	Sous	bajo
dinc	jusqu'	hasta

Table 3. Extract from the lexicon

The 300 most frequent words belong mostly to grammatical categories. We stopped at 300 to reduce the presence of words from lexical categories which are specific to the novel. We assume that our lexicon is exportable for others novels.

Homonyms words which are translated in French or in Castillan with two different words have been deleted from the lexicon. For instance, the Occitan word *a* can be either a verb (have third person singular) or a preposition. In the first case it would be translated in French by *a*; and in the second case it would be translated by *à*. Any of these ambiguity was excluded from the lexicon.

6) Frasia de Gèrbet que m'acompanha dinc [a $_{Pp}$] la crampa mia.

7) Qué m'[a $_{V}$] deishat, Cohita, en airetatge ?

4.2.2. Translation on word-to-word basis

The next step consists in transposing (see Table 4) the Occitan words which are in the lexicon in French (Trans_Fr, for example entà → pour in 8a) and in Castillan (Trans_Sp, for example entà → para in 8b). Only the bold words have been translated. The other ones remain in their original form.

8) Original	**Alavetz**	,	**entà**	trompar	**lo**	som
8a) Trans_Fr	**Alors**	,	**pour**	trompar	**le**	som
8b) Trans_Sp	**Entonces**	,	**para**	trompar	**el**	som
	Then	*,*	***to***	*fool*	***the***	*sleep*

Table 4. Example of translation

4.2.3. POS tagging with TreeTagger

POS tagging was done twice, first on the original text and then on the translated text. Table 5 for French and 6 for Castillan give an example of tags for original texts and for translated texts. Translated words are in bold. The symbol ✓ means known words and × unknown words by TT. Finally correct tags are greyed out.

First, we deleted all the enunciative particles (strikethrough text in Table 5 and 6) because there is no equivalent in both French and Castillan and the tags for them would have been inevitably wrong. While these particles play a role at the enunciative level, the following proposition is still grammatical.

- **Running TT on original text**

Some words are graphically similar in each pair of languages. For example the feminine singular definite article is *la* both in Occitan, French and Castillan (see Table 5 and 6). TT will consider them as known words (✓ in Table 5 and 6) and will use the available information about this word. Nonetheless, it does not insure that the similar words in the two languages will have the same POS.

- **Running TT on translated text**

In the translated text, similar words and translated words are considered known by TT.

Original text		Tag	Trans_fr		Tag
Dab	×	**S**	**Avec**	✓	**Pp**
la	✓	D	la	✓	D
complicitat	×	Pp	complicitat	×	S
de	✓	D	de	✓	Pp
la	✓	S	la	✓	D
lua	×		lua	×	S
~~que~~			que		
vau	×	**S**	**vais**	✓	**V**
poder	×	**S**	**pouvoir**	✓	**Vi**
,		Cm	,		Cm
adara	×	**S**	**maintenant**	✓	**Adv**
,		Cm	,		Cm
tirar	×	S	tirar	×	S
camin	×	S	camin	×	A
.		<sent>	.		<sent>

Table 5. Extract from original texts and translated texts in French tagged with French TT

Original text		Tag	Trans_es		Tag
Dab	×	**Np**	**Con**	✓	**Pp**
la	✓	D	la	✓	D
complicitat	×	Pp	complicitat	×	S
de	✓	D	de	✓	Pp
la	✓	S	la	✓	D
lua	×		lua	×	S
~~que~~			que		
vau	×	**A**	**voy**	✓	**V**
poder	✓	Vi	poder	✓	Vi
,		Cm	,		Cm
adara	×	**A**	**ahora**	✓	**Adv**
,		Cm	,		Cm
tirar	✓	Vi	tirar	✓	Vi
camin	×	S	Camin	×	S
.		<sent>	.		<sent>

Table 6. Extract from original texts and translated texts in Castillan tagged with Castillan TT.

For French and Castillan, known words are correctly tagged as well as the words near them (see the correctly tagged common nouns *complicitat* and *lua*). After translation, translated words are correctly tagged. In this case, it does not result in additional correct tags but see Table 9 for more investigation on this point.

Table 7 gives the number of unknown words by TT before and after translation:

	Unknown words (before)	Unknown words (after)
Experiment_Fr	562 (54.9%)	347 (33.9%)
Experiment_Sp	524 (51.2%)	348 (33.9%)

Table 7. Percentage of unknown words

As expected, the number of unknown words reduces after translation.

On the basis of all the tags given for the known words, TT will predict the tags for the unknown ones. This should work if a) enough tags are first correctly predicted and b) the probabilities of POS tag sequences are fairly similar between Occitan, and French and Castillan.

4.3. Results and perspectives

4.3.1. Results

For the evaluation, we first compare the annotation done by Apertium and TT with the reference annotation. The detailed tags given by Apertium and TT were initially simplified according to our tag set (see Table 1).

	Original text	Translated text
Apertium	0.65	
French TT	0.49	0.75
Castillan TT	0.46	0.80

Table 8. TT precision

The precision of Apertium, 0.65, is rather low. As we explained above, Apertium only assigns tags to known words. As a consequence, the performances correlate significantly with the number of unknown words (19% for our evaluation corpus).

The TT precision for the original text is unsurprisingly low, less than 0.5. After translation, the precision reaches up to 0.75 for French and 0.8 for Castillan. As expected, the best results are obtained with Castillan translation, probably because of the higher distribution similarities between Occitan and Castillan than between Occitan and French.

We also calculated the precision only for non-translated words, in order to evaluate the method for tagging those words.

	Original text	Translated text
French TT	0.55	0.67
Castillan TT	0.52	0.76

Table 9. Precision for non-translated words

As Table 9 shows, even non-translated words are analyzed better in translated texts than in original texts, thanks to the probabilistic system of TT based on the tags previously assigned.

4.3.2. Difficulties

The main difficulty concerns words which can have two different POS, as for example the very frequent Occitan word *a*. As explained in Section 4.2.1., this word was excluded from the lexicon and corresponds to the third person singular form of the French verb *avoir*. Unsurprisingly, French TT always tags this word as a finite verb.

We also expectedly observe difficulties concerning "cross-language homographs", i.e. Occitan words having the same spelling but not the same POS than in French or Castillan. As for example, the word *ne* is a pronoun in Occitan and an adverb in French.

These two difficulties can be easily overcome.

4.3.3. Perspectives

Our approach is resource-free and gives a precision of 0.8 which is fairly good. Nevertheless, improvements are required to raise the precision up to 0.95, as usually expected for POS tagging. Other types of strategies to bypass the use of lexicon would help. As for example, Scherrer and Sagot (2013) acquire German/Palatin [2] cognates pairs with unsupervised automatic learning methods. Such cognates would be used in our case for pairing a) Occitan words with their translation(s) in other romance languages; b) dialectal variations and c) spelling variations. Moreover this strategy would also help for lemmatizing which is currently left undone in the project.

The main objective of these experiments is to create a gold standard for Occitan POS taggers training (Tree Tagger and Talismane (Urieli and Tanguy, 2013)). Experiments described in this paper propose strategies for reducing the cost of this very long and fastidious process. They will be extended to other dialects and also to other related-languages such as Catalan by using Freeling (Carreras *et al*, 2004), a Catalan POS tagger.

5. Conclusion

The main objective of this study is to help the very long and fastidious process of creating a gold standard for POS annotation in Occitan. This approach is almost resource-free and gives a precision of 0.8 using Castillan TT. It is a very good start to annotate a large amount of data and the manual checking of annotations will be speed up in order to build a Gold standard.

These experiments show that the methodology first used between a language and one of its dialects (Bernhard and Ligozat, 2013) is exportable with similar results for pairs of languages, less close even if etymologically related.

References

Armentano I Oller, C. (2008). *Traduction automatique occitan-catalan et occitan-espagnol: difficultés affrontées et résultats atteints.* IXème Congrès International de l'AIEO, Aix-La-Chapelle.

Bec, P. (1995). *La langue occitane*. Que sais-je n°1059. Paris.

Bernhard, D. and Ligozat A.-L. (2013). Es esch fàscht wie Ditsh, oder net? Étiquetage morpho-syntaxique de l'alsacien en passant par l'allemand. *In Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, pp. 209-220.

Bras, M. (2006). Le projet TELOC : construction d'une base textuelle occitane. *In Langues et Cité : bulletin de l'observation des pratiques linguistiques,* 8, p9.

Bras, M. and Thomas, J. (2011). Batelòc : cap a una basa informatisada de tèxtes occitans. *In* A. Rieger (ed.) *L'Occitanie invitée de l'Euregio. Liège 1981 - Aix-la-Chapelle 2008 Bilan et perspectives*, Actes du IXème Congrès International de l'AIEO, Aache, Shaker.

Bras, M. and Vergez-Couret, M. (2013). Batelòc: a text base for the Occitan language, International Conference on Endangered Language in Europe, Octobre 17-18th.

Carreras, X., Chao, I., Padró, L. and Padró, M. (2004). Freeling: An Open-Source Suite of Language Analyzers, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04).*

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. In *Machine Translation: Volume 25, Issue 2*, p. 127-144.

Hana, J., Feldman, A. and Aharodnik, K. (2011). A low-budget tagger for Old Czech. *In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanitie (LaTeCH'11)*, pp. 10-18.

Massourre, J.-L. (2012). *Le Gascon, les mots et le système*. Honoré Champion, Paris.

Scherrer, Y. and Sagot, B. (2013). Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche. *In Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, pp. 195-208.

Schmid, H. (1994). Probabilistic Part-Of-Speech Tagging Using Decision Trees. *In Proceedings of the International Conference on New Methods in Language Processing*, pp. 44-49.

Sibille, J. (2007). L'occitan, qu'es aquò ? *In Langues et Cité : bulletin de l'observation des pratiques linguistiques,* 10, p. 2.

Toutanova, K., Klein, V., Manning, C.D. and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1*, NAACL'03, pp. 173-180.

Urieli, Assaf and Tanguy, Ludovic (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2013)*. Les Sables d'Olonne, France.

[2] German dialect.