Pos-tagging different varieties of Occitan with single-dialect resources

Marianne Vergez-Couret
CLLE-ERSS
Université de Toulouse
vergez@univ-tlse2.fr

Assaf Urieli
CLLE-ERSS
Université de Toulouse
assaf.urieli@univ-tlse2.fr
Joliciel Informatique
Foix, France
assaf@joli-ciel.com

Abstract

In this study, we tackle the question of pos-tagging written Occitan, a lesser-resourced language with multiple dialects each containing several varieties. For pos-tagging, we use a supervised machine learning approach, requiring annotated training and evaluation corpora and optionally a lexicon, all of which were prepared as part of the study. Although we evaluate two dialects of Occitan, Lengadocian and Gascon, the training material and lexicon concern only Lengadocian. We concluded that reasonable results ($> 89\%$ accuracy) are possible with a very limited training corpus (2500 tokens), as long as it is compensated by intensive use of the lexicon. Results are much lower across dialects, and pointers are provided for improvement. Finally, we compare the relative contribution of more training material vs. a larger lexicon, and conclude that within our configuration, spending effort on lexicon construction yields higher returns.

1 Introduction

Pos-tagging is one of the first steps in many Natural Language Processing chains, and generally requires annotated corpora and lexicons to function properly. Substantial efforts are needed to create such resources, few of which exist in the required format for less-resourced languages like Occitan. Creating them is more challenging since less-resourced languages present spelling and dialectal variations and are not necessarily standardized. In this paper, we apply a tool that was initially developed for rich-resourced languages (French and English), the pos-tagger Talismane, to different varieties and dialects of literary Occitan. We evaluate whether adapting this tool with only little annotated data is worthwhile.

Various efforts have been made recently to adapt pos-taggers to lesser-resourced languages. Täckström et al. (2013) use a semi-supervised approach based on aligned bitext between a resource-rich and resource-poor language, and achieve substantial gains. In our case, without an aligned bitext resource, we were unable to attempt this approach. Garrette et al. (2013) perform an experiment giving annotators limited time (4 hours) to annotate either training corpora or lexicons (which they call token and type annotation) for 2 low-resourced languages. They conclude that lexicons provide higher initial gains. However, whereas their lexicons are constructed by automatically selecting the most frequent words from large unannotated corpora, our study can make use of existing wide-coverage lexical resources. Scherrer and Sagot (2013) use an approach where lexical cognates are identified between a resource-rich and resource-poor language, and their pos-tags are then used to help tagging the resource-poor language. Their approach is interesting for languages, unlike Occitan, with no lexical resources available. However, even cross-language approaches require a small manually-annotated corpus for accurate evaluation. It seems simpler to begin by using this corpus for both training and evaluation before attempting more complex approaches. A finer evaluation would then be required to determine whether data quality (a small purpose-built corpus) or quantity (a large cross-language corpus) are more important for the present task.

A pos-tagger for Occitan was also developped as an intermediate step for machine translation in Apertium (Armentano i Oller and Forcada, 2006; Sánchez-Martınez et al., 2007), where the most likely translation is used to select the correct pos-tags. However, since they only evaluate the resulting translation quality, and since Apertium is not available as a standalone pos-tagger, we were unable to perform comparisons.

Our article is organized as follows: in Section 2, we give an overview of the Occitan language and its dialects. In Section 3, we present the software used, Talismane, as well as the feature and rule sets applied. In Section 4, we discuss the various resources that were constructed for this study, including corpora and lexica. In Section 5 we give the experimental setup, and discuss the results in Section 6.

2 Occitan language

Occitan is a romance language spoken in southern France and in several valleys of Spain and Italy.

The number of speakers is hard to estimate: according to several studies it might reasonably be situated around 500,000 speakers. It is even harder to evaluate the number of people with an interest in Occitan. According to a socio-linguistic survey carried out in the Midi-Pyrénées Region in 2010, 4% of the population are native or fluent speakers, 14% are speakers with an average competence and 32% understand the language, with different degrees of competence, giving an estimated total of 1.5 million people for this region alone. The interest in Occitan is supported by a sizable network of non-profit associations. Among others, the primary and secondary immersive bilingual school system Calandreta, IEO (*Institut d'Estudis Occitans*) and CFPO (*Centre de Formacion Professionala Occitan*) provide opportunities for learning Occitan at any age. Occitan is also present in the French national education system in bilingual classes at the primary school level; as optional courses at the secondary school level; and as a major or optional classes in several universities.

2.1 Occitan dialects

Occitan is not standardized as a whole. It has several varieties organized into dialects. The most widely accepted classification proposed by Bec (1995) includes Auvernhat, Gascon, Lengadocian, Lemosin, Provençau and Vivaroaupenc.

In this article we focus on two Occitan dialects: Lengadocian, spoken in a zone delimited by the Rhône, the Garonne and the Mediterranean Sea; and Gascon, spoken in a zone delimited by the Pyrenees, the Garonne, and the Atlantic Ocean. Some examples of lexical variation from Lengadocian to Gascon include the transformation of a Latin f into an h (filh/hilh), dropping the intervocalic n (luna/lua) and metathesis of the r (cabra/craba) (Bec, 1995).

We assume that probabilities of pos-tag sequences will be fairly similar between Lengadocian and Gascon in most cases. However, several examples below show non-lexical differences between the two dialects that result in different pos-tag distributions.

1. Gascon has enunciative particles: "que" for affirmative sentences, "be" for exclamatory sentences, and "e" for interrogative sentences and subordinate clauses. There is no equivalent in Lengadocian.
 - Example: "I'm buying bread and apples". Gascon: *"**Que** crompi pans e pomas."* Lengadocian: *"Compri de pans e de pomas."*

2. There is no indefinite or partitive article in Gascon.
 - Example: "He's catching birds." Gascon: *"Que gaha ausèths."* Lengadocian: *"Trapa **d'**aucèls."*
 - Example: "I want some water." Gascon: *"Que vòli aiga."* Lengadocian: *"Vòli **d'**aiga."*

3. Object and reflective clitics occur more often after the verb in Gascon than in Lengadocian.
 - Example: "To come in and get served?" Gascon: *"Entrar e hèr-**se** servir ?"* Lengadocian: *"Dintrar e **se** far servir ?"*

4. Double-negatives in Gascon: the preceding "ne/no" is mandatory in Gascon, but not in Lengadocian.
 - Example: "He can't hear anything." Gascon: *"**N'**enten pas arren."* Lengadocian: *"Enten pas ren."*

2.2 Written Occitan

Written Occitan first appeared in medieval times, with all dialects represented in literature. This results in a lot of inter- and intra-dialectal variation within the texts. This geolinguistic variation corresponds to (i) variations in spelling reflecting variations in pronunciation (for instance *contes/condes*) and (ii) lexical variations (for instance *pomas de terra/mandòrra*). Numerous spelling conventions account for additional variation within Occitan text. The spelling used in medieval times is nowadays called the "troubadour spelling". This spelling gradually disappeared with the decline of literary production. Since the 19$^{\text{th}}$ century, two major spelling conventions can be distinguished: the first was influenced by French spelling, and includes Mistral's spelling in Provence and the Gaston Febus' spelling in Bearn; the second, called "classical spelling" and inspired by the troubador spelling, appeared in the 20$^{\text{th}}$ century. It is a unified spelling convention distributed across all of the Occitan territories (Sibille, 2007). Diachronic variation corresponds to changes in spelling conventions over time (for instance the evolution in the spelling of conjugated verbs: *avian* vs. *avìan*). Embracing all dialectal and spelling variations is one of the main objectives of the BaTelÒc project.

2.3 BaTelÒc Project

The BaTelÒc project (Bras and Thomas, 2011; Bras and Vergez-Couret, 2013) aims at creating a wide-coverage collection of written texts in Occitan, including literature (prose, drama and poetry) as well as other genres such as technical texts and newspapers. The texts aim to cover the modern and contemporary periods, as well as all dialectal and spelling varieties. 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, such as pos-tags. These would allow new querying possibilities, e.g. the disambiguation of homographs such as *poder* as a common noun ("power") and *poder* as a verb ("be able to"). In order to provide such annotations, Part-Of-Speech annotation tools are required. We therefore decided to use a probabilistic pos-tagger based on supervised machine learning methods: Talismane.

3 The Talismane pos-tagger

The present study trained the open source Talismane pos-tagger (Urieli, 2013) on an Occitan training corpus. Talismane has already been applied to English and French pos-tagging, attaining an accuracy $\approx 97\%$ (Urieli, 2014). It allows for the incorporation of a lexicon both as training features and as analysis rules. In terms of features, this comes down to saying, "if the word X is listed in the lexicon as a common noun, then it is more likely to be a common noun". This information is incorporated into the statistical model during training, along with other features listed below. Analysis rules override the statistical model's decisions during analysis, either imposing or prohibiting the choice of a certain category. For example, a rule might say, "the word X cannot be assigned the closed category *preposition* unless it is listed as a preposition in the lexicon".

To select the machine learning configuration of the Occitan pos-tagger, we performed a grid search of different classifier types and parameters, and settled on a linear SVM classifier with $\epsilon = 0.1$ and $C = 0.5$.

3.1 Features

We used the identical feature set for Occitan as the one used by Talismane for French and English. These include, for the token currently being analysed: W the word form; P each of the token's possible pos-tags according to the lexicon; L each of the token's possible lemmas according to the lexicon; U whether the current token is unknown in the lexicon; *1st* whether the token is the first in the sentence; *Last* whether the token is the last in the sentence; *Sfx* the last n letters in the token; as well as various regular expression features testing whether the token starts with a capital letter, contains a dash, a space or a period, or contains only capital letters.

We also used the following additional features for the tokens before and after the current token (where the subscript indicates the position of the token with respect to the current token):

$W_{-1}, W_1, P_{-1}, P_1, L_{-1}, L_1, U_1$, where P_{-1} looks at the pos-tag assigned to the previous token, and is thus the standard bigram feature. We also included various two-token and three-token combinations of all of the above basic features, e.g. $P_{-2}P_{-1}$ giving the standard trigram feature.

3.2 Rules

The following rules were defined around closed class pos-tags (i.e. non-productive functional categories) and open class pos-tags (i.e. productive lexical categories).

- Closed classes: for each closed class pos-tag (e.g. prepositions, conjunctions, pronouns, etc.), only allow the pos-tagger to assign this pos-tag if it exists in the lexicon. This prevents us, for example, from inventing new prepositions.
- Open classes: do not assign an open class pos-tag (e.g. common noun, adjective, etc.) to a token if it is only listed with closed classes in the lexicon. This prevents us, for example, from assigning a tag such as "common noun" to the token "*lo*" ("the").
- Rules which automatically assign the pos-tags `Card` and `Pct` respectively to numbers and punctuation. These were applied systematically in all experiments.

4 Resources

For Talismane to function properly, various resources are required: a training corpus from which the statistical model is learned, one or more evaluation corpora to evaluate performance, and optionally a lexicon for wide-coverage features and rules. These resources all rely on a tagset specifically designed for Occitan, shown in Table 1.

4.1 Lexicon and tagset

In the present study, we decided to construct a lexicon for one dialect only, the Lengadocian dialect, corresponding to our training corpus.

The lexicon was built from available digital resources: the Laus dictionary of Lengadocian (Laus, 2005), as well as certain closed-class entries and proper nouns from the Apertium lexicon. The Laus dictionary in particular covers different varieties of Lengadocian. For example, the entry for "night" includes three variants: *nuèch / nuèit / nuòch*. Inflected forms for verbs were gathered from *Lo congrès permanent de la lenga occitana*, which provides a complete verb-conjugation module[1]. A script was written to automatically generate inflected forms for adjectives, nouns and past participles from the base form entries. The number of entries for each pos-tag and total count are given in Table 1.

4.2 Training corpus

For training Talismane, a homogeneous corpus in the Lengadocian dialect was extracted from a single novel: *E la barta floriguèt* by Enric Molin, an Occitan author from the Rouergue region. Since the present study concentrates on differences between dialects and varieties, no attempt was made to construct a balanced training corpus. The corpus contains around 2500 tokens manually annotated with pos-tags, lemmas, and additional morpho-syntactic information (grammatical gender, number, person, tense and mood). The first 1000 tokens were annotated separately by three annotators, who then consolidated their annotations into a single gold standard, with an annotation guide. The remaining 1500 tokens were annotated by a single annotator, who consulted the others in cases of doubt.

In the present study, the annotated lemmas and additional morpho-syntactic information were not used.

4.3 Evaluation corpora

For evaluation, three different corpora were compiled: the first one, the *Rouergue* corpus, was extracted from: *Los crocants de Roergue* by Ferran Delèris, another author from the Rouergue region; the second one, the *Lot* corpus, was extracted from *Dels camins bartassièrs* by Marceu Esquieu, written in another

[1] `http://www.locongres.org/oc/aplicacions/verboc/conjugar`

Tag	Description	Lexicon size
A	Adjective (general)	29,638
A$	Adjective (possessive)	85
Adv	Adverb (general)	751
Adv$	Adverb (negative, quantifier, exclamatory and interrogative)	46
Cc	Coordinating conjunction	8
Cs	Subordinating conjunction	150
Det	Article	127
Card	Cardinal number	42
Cli	Clitic	72
CliRef	Reflexive clitic	17
Inj	Interjection	7
Nc	Common noun	25,817
Np	Proper noun	4,603
Pct	Punctuation	15
Pe	Enunciative particle (Gascon only)	0
Pp	Present participle	4,530
Pr	Preposition	521
Prel	Relative pronoun	37
Pro	Pronoun	81
Ps	Past participle	17,963
PrepDet	Amalgamated preposition and article	499
Vc	Conjugated verb	135,731
Vi	Infinitive verb	4,643
Z	Consonant for phonetic liaison	3
Total		**225,386**

Table 1: Tagset

variety of Lengadocian; the third one, the *Gascon* corpus, was extracted from *Hont blanca* de Jan Loís Lavit, representing a variety of Gascon. The three corpora aim at representing different varieties of Occitan: firstly, two different dialects: Lengadocian and Gascon; secondly, two varieties of Lengadocian: Rouergue and Lot.

Table 2 shows a statistical comparison of the different corpora. As we can see, the percent of tokens unseen in the training corpus (excluding punctuation) ranges from 46% for the same dialectal variant (Rouergue) to 56% for a different dialect (Gascon). The difference is even more striking in terms of the Lengadocian lexicon: 17% unknown forms in the Rouergue corpus vs. 40% unknown forms in the Gascon corpus. Closed class coverage is particularly good for the two Lengadocian variants, with only 1.5% and 1% unknown forms, as opposed to 20% in the Gascon corpus.

5 Experiments

The resources we built were designed with several questions in mind:

- Which is the best strategy for each evaluation corpus?
- Is it always useful to apply closed-class rules?
- To what extent can a model built from a training corpus for a single dialectal variety be applied to other varieties and dialects?
- To what extent can a lexicon for one dialect be applied to another dialect?
- What methods can be used to improve analysis for a dialect different from the training/lexicon dialect?

Corpus	**Training**	**Rouergue**	**Lot**	**Gascon**
Size	2501	701	467	469
Size (without punct.)	2078	591	388	399
% unknown in training corpus		46.36	48.97	56.39
% unknown in lexicon	0.10	16.58	19.85	40.10
Open class tokens	1111	324	201	203
% unknown in training corpus		76.23	82.59	87.68
% unknown in lexicon	0.18	29.01	37.31	59.11
Closed class tokens	967	267	187	196
% unknown in training corpus		10.11	12.83	23.98
% unknown in lexicon	0.00	1.50	1.07	20.41

Table 2: Training and evaluation corpora

A second range of experiments was designed to answer the following question: Given limited resources, is it better to annotate a larger training corpus, or compile a larger lexicon?

To this end, we divided the training corpus into two halves, `train1` and `train2`. We also created several sub-lexica: closed classes only (`closed`), closed classes + half of the open class entries (`half1`) closed classes + the other half of the open class entries (`half2`), the full lexicon (`full`) and an empty lexicon (`empty`). Finally, we tested with and without closed class rules. This gave us a total of 3 training corpus options × 5 lexicon options × 2 rule options = 30 evaluations per evaluation corpus.

We measured in each evaluation the total accuracy, the precision, recall and f-score for each pos-tag, and for all open pos-tags and all closed pos-tags combined. These were also measured separately for the set of tokens known and unknown in the lexicon.

6 Results

6.1 Overall results

Figure 1 shows results for the different lexicons and with/without closed-class rules (+rules on the figure). Not surprisingly, the best configuration for all evaluation corpora was the full training corpus, the full lexicon, and closed-class rules applied. This gives an accuracy of 87.02% for the Rouergue corpus, 89.08% for the Lot corpus, and 66.17% for the Gascon corpus. We can see that even a small training corpus provides reasonable results: almost 90% with only 2500 annotated tokens.

Within a given dialect, variation in style and genre seem more important than variation due to dialectal varieties: indeed, a training corpus in the Rouergue variety gave better results for an author in the Lot variety than for another author in the Rouergue variety. Another reason for handling dialects as a whole is that it would be very difficult and time consuming to construct a separate lexicon for each variety within a given dialect.

The much lower results for Gascon are expected, given the much lower training corpus coverage and lexicon coverage shown in Table 2, and the differences in pos-tag distribution presented in Section 2.1.

6.2 Closed class rules

The use of closed-class rules presented in Section 3.2 improved accuracy for all three corpora. The accuracy rose from 85.88% to 87.02% for the Rouergue corpus, from 88.01% to 89.08% in the Lot corpus, and 66.10% to 67.16% in the Gascon corpus. The last result is somewhat surprising, given the fact that 20% of the closed class tokens in the Gascon corpus are unknown in the lexicon.

6.3 Lexicons

The five lexicon setups described above allowed us to compare the contribution of different parts of the lexicon. Using a lexicon with only closed classes gives a fairly radical increase in all cases: together with rules, we gain 7.13% for Rouergue, 11.99% for Lot, and 4.9% for Gascon. When we add the full

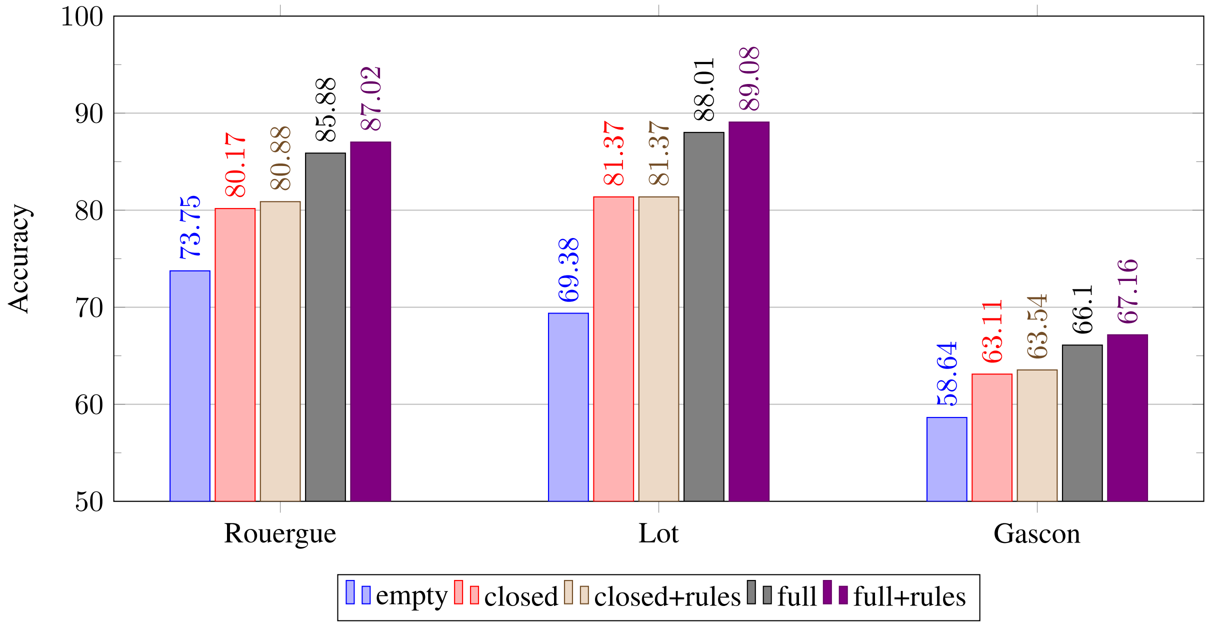

Figure 1: Pos-tagging lexicon/rules comparison: accuracy by corpus

lexicon with open and closed classes, we see an additional increase of 6.14% for Rouergue, 7.71% for Lot, and 3.62% for Gascon with respect to a closed-class lexicon only.

The open class gains are not directly correlated to the percentage of unknown words: the Lot corpus has far more unknown words than the Rouergue corpus, and yet gains more in terms of accuracy when the lexicon is added. Furthermore, the gains affect unknown words as well, probably through improvement in tagging of neighboring words and n-gram features: we see an average gain of 8.54% in accuracy for unknown words in Rouergue between the half1+rules/half2+rules and full+rules configurations, and 17.96% for unknown words in Lot.

6.4 Improving accuracy for other dialects

Given the relatively low score for Gascon, the question is, what can be done to improve this accuracy? In view of the training corpus in Lengadocian and the differences described in Section 2.1, it is clear that certain phenomena will be very difficult to detect, especially when Gascon lexical items are combined with uniquely Gascon pos-tag sequences. Additionally, one Gascon part-of-speech, the enunciative particle (annotated `Pe`), is entirely missing from Lengadocian. However, this pos-tag happens to be the most common one for the word *“que”*, and the only possibility for the word *“be”*.

We thus tested the addition of a new rule for Gascon only, stating that *“be”* is always annotated `Pe`, and *“que”* is annotated `Pe` whenever it’s found at the start of a sentence, after a coordinating conjunction, or after a comma. For a total of 30 ennunciative particles, this rule gives us 17 true positives, 1 false positive, and 13 false negatives, for an f-score of 70.83%. It increases the total accuracy from 67.16% to 69.72%.

Beyond this rule (and possibly other similar rules), improving the accuracy necessarily requires more resources. Given the gains provided by small but complete closed-class lexica, a priority should thus be given to constructing a full-coverage closed-class lexicon for Gascon, and replacing the Lengadocian closed-class lexicon with this one during analysis. It is an open question whether it is better to use a higher-recall lexicon covering all dialects, or a higher-precision lexicon covering only Gascon. A similar question concerns training corpora, which are typically much more costly to construct than lexica, given that dictionaries in digital form are generally already available. Is it better to use a small training corpus per dialect, or to mix training corpora for all dialects into a larger training corpus? This of course depends on the degree of similarity between the dialects, and cannot be answered without empirical testing.

6.5 Build a training corpus or a lexicon?

To answer the question regarding the relative importance of annotating more training data or compiling larger lexica, we ran an experiment where the training corpus and open-class lexicon were each divided into two halves. We then compared the results provided by a single half of the training corpus and a single half of the lexicon (4 possible combinations) with results provided when including either the entire training corpus or the entire lexicon, but not both. Since the lexicon covers Lengadocian, we concentrate on the two Lengadocian corpora only, considering them as a single corpus.

The mean gain for doubling the training corpus from 1,250 tokens to 2,500 tokens is 1.46%, whereas the mean gain for doubling the open-class lexicon from 110K entries to 220K entries is 4.16%. It is thus much more productive to double the lexicon size, in our configuration. Note of course that there is no guarantee that this tendency would continue if we doubled the size of the training corpus and lexicon again. Also, while it is always possible (albeit costly) to annotate more text, there is a limit to the available lexical resources that can easily be compiled.

7 Conclusion and perspectives

In the present study, we show that supervised approaches, usually considered too costly for lesser-resourced languages, can achieve good results ($> 89\%$) with very little annotated material, as long as wide-coverage lexicon is available. We determined that given a limited amount of time, it is better to construct a larger lexicon than to annotate more training material. It would be interesting to repeat this experiment when we have gathered more training material and a wider-coverage lexicon, in order to view the tendencies in a graphical form.

One of the main objectives of the present study was to test a proof-of-concept for Occitan pos-tagging and identify guidelines for future efforts in this area. One of the first benefits of our work is that, in addition to the training and evaluation corpora and lexicon, we now have a functioning pos-tagger which can help efficiently construct more training and evaluation material, and an annotation guide to help correct this material.

Many recent studies have used semi-supervised cross-language pos-taggers, resulting in a larger quantity but lower quality of training data. It would be interesting to compare such an approach to our present supervised approach, as well as seeing whether the two can be combined (e.g. by giving more weight to the higher quality material during training).

The use of Talismane as a pos-tagger gives us a certain degree of robustness for handling language variants. Talismane is a hybrid toolkit: on the one hand, it provides robust supervised machine learning techniques, allowing us to ensure that as more data gets annotated, the results improve. On the other hand, it allows us to override the statistical models with symbolic rules, thus compensating for the low representativity of less common phenomena in the limited training material, as well as allowing us to take into account phenomena specific to the dialect or variety being analysed. The use of rules needs to be explored more deeply and extended to other phenomena than those explored in the present study.

In terms of the Gascon dialect, although the results are much better than random chance, they still leave much to be desired. Nevertheless, all of the phenomena observed for Lengadocian applied to Gascon as well, albeit to a lesser extent: the closed-class lexicon and related rules provided substantial gains (despite 20% unknown closed-class tokens in the lexicon), and additional gains were provided by the open-class lexicon. We tested with success a single rule for Gascon around the enunciative particle. Efforts would now be required to identify additional rules. However, the most promising perspective is the construction of a lexicon for Gascon, in particular giving full coverage for all closed classes. It is yet to be determined whether this lexicon should replace the Lengadocian lexicon during analysis, or complete it. A similar question applies to training corpora: if we annotate a Gascon training corpus, should it be combined with the Lengadocian corpus or should Gascon be trained separately.

Finally, there is another practical perspective from the present study: to use lists of unknown pos-tagged words as the initial input for the construction of wider-coverage lexica.

References

Carme Armentano i Oller and Mikel L Forcada. 2006. Open-source machine translation between small languages: Catalan and aranese occitan. *Strategies for developing machine translation for minority languages*, page 51.

P. Bec. 1995. *La langue occitane*. Number 1059. Que sais-je ? Paris.

M. Bras and J. Thomas. 2011. Batelòc : cap a una basa informatisada de tèxtes occitans. In *L'Occitanie invitée de l'Euregio. Liège 1981 - Aix-la-Chapelle 2008 Bilan et perspectives*, Aix-la-Chapelle. Aache, Shaker.

M. Bras and M. Vergez-Couret. 2013. Batelòc : a text base for the occitan language. In *Proceedings of the International Conference on Endangered Languages in Europe*, Minde, Portugal.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *ACL 2013*, pages 583–592, Sofia, Bulgaria.

C. Laus. 2005. *Dictionnaire Français-Occitan*. IEO del Tarn.

Felipe Sánchez-Martınez, Carme Armentano-Oller, Juan Antonio Pérez-Ortiz, and Mikel L Forcada. 2007. Training part-of-speech taggers to build machine translation systems for less-resourced language pairs. In *Procesamiento del Lenguaje Natural (XXIII Congreso de la Sociedad Espanola de Procesamiento del Lenguaje Natural)*, volume 39, pages 257–264, September.

Yves Scherrer and Benoît Sagot. 2013. Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources. In *RANLP Workshop on Adaptation of language resources and tools for closely related languages and language variants*.

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

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Assaf Urieli. 2014. Améliorer l'étiquetage de "que" par les descripteurs ciblés et les règles. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014)*, Marseille, France.