A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments

Aleksandra Miletic
CNRS UMR 5263 CLLE
University of Toulouse, France
aleksandra.miletic@univ-tlse2.fr

Myriam Bras
CNRS UMR 5263 CLLE
University of Toulouse, France
myriam.bras@univ-tlse2.fr

Marianne Vergez-Couret
EA 3816 FoReLLIS
University of Poitiers, France
marianne.vergez.couret@univ-poitiers.fr

Louise Esher
CNRS UMR 5263 CLLE
University of Toulouse, France
louise.esher@univ-tlse2.fr

Jean Sibille
CNRS UMR 5263 CLLE
University of Toulouse, France
jean.sibille@univ-tlse2.fr

Clamença Poujade
CNRS UMR 5263 CLLE
University of Toulouse, France
clamenca.poujade@univ-tlse2.fr

Abstract

Occitan is a Romance language spoken mainly in the south of France. It has no official status in the country, it is not standardized and displays important diatopic variation resulting in a rich system of dialects. Recently, we created a first treebank for this language (Miletic et al., 2020). However, this corpus is based exclusively on texts in the Lengadocian dialect. Our paper describes the work aimed at extending the existing corpus with content in three new dialects, namely Gascon, Provençau and Lemosin. We describe both the annotation of initial content in these new varieties of Occitan and experiments allowing us to identify the most efficient method for further enrichment of the corpus. We observe that parsing models trained on Occitan dialects achieve better results than a delexicalized model trained on other Romance languages despite the latter training corpus being much larger (20K vs 900K tokens). The results of the native Occitan models show an important impact of cross-dialectal lexical variation, whereas syntactic variation seems to affect the systems less. We hope that these results, as well as the associated corpus, incorporating several Occitan varieties, will facilitate the training of robust NLP tools, capable of processing all kinds of Occitan texts.

1 Introduction

Occitan is a Romance language spoken in southern France (except in the Basque and Catalan areas), in several valleys of the Italian Piedmont and in the Val d'Aran in Spain. It does not have the status of an official language, and as many such languages, it is not standardized. It displays a rich system of diatopic varieties, organized into dialects. The variation can be appreciated at all levels of linguistic structure: it can be lexical or phonetic, but also morphological and syntactical (see Section 2). Also, there are two different spelling norms in use today, one called the classical, based on the Occitan troubadours' medieval spelling, and the other closer to the French language conventions (Sibille, 2002).

Since all these factors contribute to data sparsity, they make Occitan particularly challenging for natural language processing (hereafter NLP). In fact, Occitan is still relatively low-resourced, although recent efforts have started to remedy this situation. The firsts of them was the creation of the BaTelÒc text base (Bras and Vergez-Couret, 2016) and the RESTAURE project, which resulted in the creation

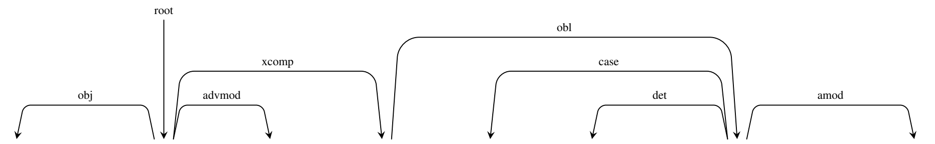
Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 140–149
Barcelona, Spain (Online), December 13, 2020

of an electronic lexicon (Vergez-Couret, 2016; Bras et al., 2020) and a POS tagged corpus (Bernhard et al., 2018). Even more recently, the first treebank for Occitan was created (Miletic et al., 2020) following Universal Dependencies guidelines[1]. Whereas the RESTAURE corpus contains several dialects, the Loflòc lexicon and the treebank are based on only one dialect – the Lengadocian. However, our goal is to be able to train robust machine learning tools capable of successfully processing texts in all varieties of Occitan. There are two main possible solutions to this: we need either training corpora representative of all Occitan varieties or extensive compensation methods for NLP which would allow us to adapt tools across varieties, such as delexicalized cross-lingual parsing. This technique consists in training a parsing model on a delexicalized corpus of a source language (i.e., using only POS tags and morphosyntactic features while ignoring tokens and lemmas) and then using the model to process data in the target language. It has been successfully used on a number of language pairs in the past (McDonald et al., 2013; Lynn et al., 2014; Duong et al., 2015; Tiedemann, 2015), including on Occitan (Miletic et al., 2019b). Both the corpus building solution and the parsing transfer solution are explored in the remainder of the paper.

In Section 2, we give a brief description of the linguistic properties of Occitan and of its main dialects. Section 3 describes the addition of three new dialects to the existing treebank and the resulting corpus. Section 4 is dedicated to experiments looking to identify the most effective method for training a robust parsing model for all Occitan dialects in the corpus. Finally, in Section 5, we give our conclusions and directions for future work.

2 Occitan: Main Linguistic Properties and Diatopic Variation

Occitan belongs to the Gallo-Romance group of Romance languages, together with standard French and “langues d’oïl”, Francoprovençal and Catalan. It is closer to Catalan than to French and forms with Catalan a subgroup called occitano-roman (Bec, 1970). It is a null subject language with tense, person and number inflection marks on finite verbs for each person. Many dialects mark number and gender inflection on all components of the noun phrase. Unlike contemporary French, Occitan maintains the use of the preterite (*passat simple*), which contrasts with the perfect tense (*passat compausat*), and the use of the imperfect subjunctive, even in oral colloquial speech. An example in Lengadocian illustrating some of these properties is given in Example 1.



(1)	Vos	vòli	pas	espaurugar	amb	lo	rescalfament	planetari
	you.ACC.PL	want.1SG	NEG	frighten	with	the.SG.M	warming	planetary.SG.M

‘I don’t want to scare you with global warming.’

As mentioned above, Occitan has several diatopic varieties, organized into dialects. The most widely accepted classification proposed by Bec (1995) includes Auvernhat, Gascon, Lengadocian, Lemosin, Provençau and Vivaroaupenc (see Figure 1), but these dialects are not homogeneous: they form a continuum with areas of greater or lesser variation. In this article we focus on four of them: Lengadocian, Gascon, Provençau and Lemosin[2], since they are the ones for which the greatest number and variety of texts are currently available. Differences in dialects can be appreciated at different levels (lexical, phonological, morphological and syntactical), as shown below.

The variation can be lexical: e.g., the word *potato* translates as *mandòrra* in some Gascon varieties, but as *trufa/trufet* or *patana/patanon* in Lengadocian. A large part of this type of variation stems from different phonological processes, many of which appear in Gascon: the aspirated *h* in word-initial position in

[1] https://universaldependencies.org/

[2] Names of dialects are given in Occitan (each one in its dialect) as there is no standardized orthographic form for those names in English.

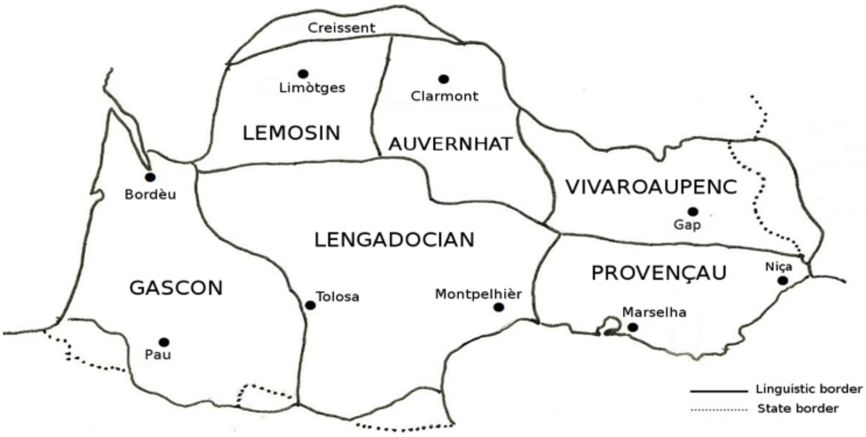


Figure 1: Occitan dialects map

words such as *hilh* 'son', *hèsta* 'celebration' (cf. *filh* in Lengadocian and Lemosin, *fiu* in Provençau and Vivaroaupenc, *fèsta* in Lengadocian and Provençau, *festa* in Lemosin, Auvernhàt and Vivaroaupenc), the drop of the intervocalic *n* in words such as *lua* 'moon' (cf. *luna* in Lengadocian) and the *r* metathesis in words such as *craba* 'goat', *dromir* 'to sleep' (cf. respectively *cabra, dormir/durmir* in a large part of the Lengadocian area). It is also caused by the existence of several spelling norms. Since the 19th century, two major spelling conventions can be distinguished: the first was influenced by French; the second, called the "classical spelling" and inspired by the medieval troubadour spelling, appeared in the 20th century. The latter is a unified spelling convention distributed across all of the Occitan territories (Sibille, 2002).

On the morpho-syntactic level, verb inflection varies from one dialect to another as illustrated in Table 1, which gives the present indicative of the verb *èsser/èstre* 'to be' in the most common paradigm for each of the four dialects (there is also intradialectal variation).

Number	Person	Gascon	Lemosin	Lengadocian	Provençau
sg	1st	soi	sei	soi	siáu
	2nd	ès	ses	ès/siás	siás
	3rd	ei/es	es	es	es
pl	1st	èm	sem	sèm	siam
	2nd	ètz	setz	sètz	siatz
	3rd	son	son	son	son

Table 1: Verb *èsser/èstre* 'to be' in present indicative across dialects

When it comes to syntax, there is more homogeneity across dialects, but Gascon exhibits several important specificities. First, it has enunciative particles which mark the sentence modality: the most frequent are *que* for affirmative sentences, *be* and *ja* for exclamative sentences and *e* for interrogative sentences and subordinate clauses. They appear between the subject and the verb and their presence is even obligatory in some Gascon areas. They have no equivalent in the Lengadocian, Provençau and Lemosin dialects (cf. Example 2.a). The interrogative and relative pronoun *qui* 'who'(cf. Example 2.b), which is scarcely used in Lengadocian, Provençau and Lemosin and only as an interrogative pronoun, has many functions in Gascon, such as the subject or direct object functions (where the other dialects use *que*), and it can be precedeed by the preposition *de* regardless of the verbal rection (cf. Example 2.c), which does not occur in other dialects. *Qui* and *de qui* can also be a subordinating conjunction introducing a completive clause. Furthermore, object clitics and reflexive pronouns are more often found in post-verbal position than it is the case in Lengadocian, in which they are typically pre-verbal (cf. Example 2.d). Finally, unlike in other dialects, there are no indefinite or partitive articles in Gascon (cf. Example 2.e).

(2) a.

Lo	vent	**que**	s’	èra	lhevat	e	**que**	hasó	drin	fresc.
the.SG.M	wind	PART	REFL	was	risen	and	PART	did	slightly	cold

‘The wind had started blowing and it was a bit cold.’
Leng.:*‘lo vent s’èra levat e fasiá un pauc freg.’*

b.

l’	atge	**qui**	a
the	age	that	has

lit. *‘the age he has’, ‘his age’*
Leng.:*‘l’atge qu’as / que as’*

c.

los	guardians	d’	un	concèpte	de	civilizacion	**de**	**qui**	calèva	preservar
the.PL.M	guardians	of	a.M.SG	concept	of	civilization	PREP	which	ought	preserve

‘the guardians of a concept of civilization that needed to be preserved’
Leng.:*‘los gardians d’un concèpte de civilizacion que caliá preservar’*

d.

Ne	cau	pas	està	**’s**	darrèr	mieidia
NEG	ought	NEG	stay	REFL	after	noon

‘You shouldn’t stay after noon’
Leng.:*‘vos cal pas demorar après miègjorn’*

e.

entà	véner	**Ø**	objècts	de	pietat.
in.order.to	sell		objects	of	piety

‘in order to sell objects of piety’
Leng.:*‘per vendre d’objèctes de pietat’*

It is important to note the potential impact of these characteristics on NLP tools based on machine learning. The lexical and morpho-syntactical variation is potentially problematic, but it is mostly a question of coverage: it can be alleviated either by a carefully engineered training corpus, representative of as many dialects as possible, or by an extensive, dialect-diversified lexicon. If a combination of the two were available, we could reasonably suppose that the effect of the variation would be minimized and that a tool trained and used in such conditions would be able to POS-tag, lemmatize and parse texts independently of dialect. However, the syntactic variation has a potentially more profound effect: a Gascon corpus can be expected to have more ambiguity related to relative pronouns, but also a different distribution of POS tags (i.e., more particles, fewer determiners) and of syntactic structures (i.e., more right-branching verb dependents) than corpora in other dialects. This could compromise the transfer of parsers trained on other dialects to Gascon and *vice versa*.

3 Extending the Existing Corpus with Content in Other Dialects

3.1 Initial Treebank in Lengadocian

The first treebank for Occitan (TTB: Tolosa Treebank) was created recently (Miletic et al., 2020). It contains 19K tokens of Lengadocian texts spanning 5 different genres (literature, newspaper, encyclopedia, scientific text and blog) (cf. row *Lengadocian* in Table 2). It is annotated for POS-tags, lemmas and syntactic dependencies. The annotation was done according to the Universal Dependencies (UD) guidelines, with some adaptations (see (Miletic et al., 2020) for a detailed description). A part of the content was taken from the RESTAURE project and the initial GRACE tagset was converted automatically to the UD tags (Miletic et al., 2019a). The remainder was tagged and lemmatized manually.

As for the syntactic annotation, the whole Lengadocian corpus was processed following the same procedure. We offer here a quick overview of the experiment; for a detailed account, see (Miletic et al., 2019b). Given the absence of training data for Occitan at the start of our project, we explored delexicalized cross-lingual parsing based on existing UD corpora for Romance languages. We trained a total of 21 models and evaluated them on a manually annotated Occitan sample. The top 5 models achieved LAS between 70.0 and 71.6 points[3]. These models were trained on individual or combined UD treebanks in Italian (ISDT, ISDT+ParTUT), French (GSD, GSD+ParTUT+Sequoia), and Portuguese (Bosque). One model per language was selected (Italian ISDT, French ParTUT+GSD+Sequoia, Portuguese Bosque) for manual pre-annotation of new samples in Occitan. During manual validation of their output, the human annotator observed that their performances were rather homogeneous. In the following stages of our work, we therefore decided to train a delexicalized model on a merged trilingual corpus (bringing the train size to 900K tokens) in an effort to further improve the performances. Using this model for pre-annotation brought the mean manual validation time from 340 tokens/h (for a fully manual annotation) to 650 tokens/h.

The quality of the manual annotation was regularly evaluated through inter-annotator agreement both in terms of Cohen's *kappa* and as a simple agreement ratio (raw percentage of consistent annotations between annotators)[4] (Miletic et al., 2020). The values stabilized around 0.87 for Cohen's *kappa* and 88% for the agreement ratio during the last stages of annotation.

	literature	newspaper	encyclopedia	science	blog	TOTAL
Lengadocian	13056	974	3546	776	621	18973
Gascon	1672	2379	-	-	-	4051
Lemosin	1323	-	-	-	-	1323
Provençau	1275	-	-	-	-	1275
TOTAL	17326	3353	3546	776	621	25622

Table 2: Distribution of the TTB corpus content by dialect and by genre

3.2 Adding Gascon, Lemosin and Provençau

In order to enrich the existing Lengadocian corpus with initial samples of other dialects, we draw on the RESTAURE corpus described above. More precisely, we transferred all available texts in Gascon, Lemosin and Provençau to the TTB corpus. The samples contain around 4K, 1,3K and 1,2K tokens respectively. There is less genre diversity than in the Lengadocian corpus: the Gascon sample contains literature and newspaper content, whereas the Lemosin and Provençau sections of the corpus are limited to literary texts. The distribution of content by genre is given in Table 2.

The content in the new dialects was lemmatized and POS-tagged as part of the RESTAURE project, and the initial GRACE tagset was converted to the Universal Dependencies tags. The syntactic annotation method we used is the same one used for the initial Lengadocian subcorpus: we pre-annotated the samples using the delexicalized parsing model trained on French, Italian and Portuguese UD corpora described above and then corrected manually. The manual annotation stage was relatively fast and simple. No significant performance decrease was noted by the annotators[5] compared to the Lengadocian subcorpus. This brought the total size of the corpus to 25K tokens. Table 3 gives some basic statistics for the whole corpus and by dialect. Tables 4 and 5 give counts for POS tags and syntactic labels, respectively, for the whole corpus.

It should be noted that the added samples remain fairly small, especially given their intended use as training and evaluation data for NLP tools based on machine learning, whose performances notoriously depend on the size of the data. This is due to two factors. First, as with many non-standardized varieties that are most often not written, it was not easy to acquire content in these dialects, especially if the licensing issues are taken into account. We therefore worked with the content that was already available, leaving further extensions to later efforts. Second, we wanted to identify promising methods

[3] LAS (labelled attachment score): the percentage of tokens for which the parser correctly identifies both the head and the label.

[4] We are aware that both of these measures have their deficiencies: the former is intended for classification tasks and dependency annotation is more complex, whereas the latter does not correct for chance agreement. However, both have been used in treebank building projects (cf. (Uria et al., 2009; Bhat and Sharma, 2012; Urieli, 2013) for Cohen's *kappa*, (Skjærholt, 2013; Voutilainen and Purtonen, 2011) for the agreement ratio) and they allow to estimate the agreement level in the corpus.

[5] The annotation campaign was managed by the first author of this paper, whereas the remaining authors acted as annotators.

for the annotation process as early as possible in order to simplify the work of our annotators. From our experience, this has an important impact on ergonomic issues during the annotation stage.

	All	Lengadocian	Gascon	Limousin	Provençal
Tokens	25622	18973	4051	1323	1275
Types	5786	4191	1333	570	547
Lemmas	4196	3045	1088	474	472
No. of sentences	1522	1113	255	77	77
Mean sent. length	16.83	17.04	15.89	17.18	16.56

Table 3: Annotated corpus information

Tag	Count
ADJ	1056
ADP	3174
ADV	1360
AUX	671
CCONJ	769
DET	3560
INTJ	90
NOUN	4468

Tag	Count
NUM	298
PART	148
PRON	1915
PROPN	707
PUNCT	3706
SCONJ	449
VERB	3233
X	18

Table 4: POS tag counts in the corpus

Label	Meaning	Count
acl	adjectival clause	520
advcl	adverbial clause	379
advmod	adverbial modifier	1224
amod	adjectival modifier	798
appos	apposition	99
aux	auxiliary	350
case	case mark	2661
cc	coordinating conjunction	754
ccomp	clausal complement	174
compound	compound word element	5
conj	coordination conjunct	964
cop	copula	335
csubj	clausal subject	12
dep	dependency	10
det	determiner	3556
discourse	discourse element	62
dislocated	dislocated element	85
expl	expletive element	506
fixed	element of a fully grammaticalized MWE	271

Label	Meaning	Count
flat	element of an exocentric construction	191
iobj	indirect object	275
mark	subordination mark	803
nmod	nominal modifier	1159
nsubj	nominal subject	1070
nummod	numeral modifier	172
obj	direct object	1382
obl	oblique dependent	1509
orphan	element orphaned by ellipsis	47
parataxis	paratactic element	305
punct	punctuation	3706
reparandum	overriden speech disfluency	5
root	sentence root	1496
vocative	vocative	69
xcomp	open clausal complement	535

Table 5: Dependency labels in the corpus

4 Exploring Methods for Expanding the Multi-Dialect Part of the Corpus

4.1 Evaluation Setup

In order to further enrich our treebank with content in different dialects, we explore several possibilities to improve the quality of the automatic pre-annotation. As stated above, the delexicalized model trained on Italian, Portuguese and French corpora from the UD collection was useful in the first round of annotation. However, given the Occitan content at our disposal, we examined if parsing models trained on the same language (although on much smaller amounts of text) yielded better results. We consider three main scenarios:

1. delexicalized cross-lingual parsing with the model trained on UD corpora in Italian, Portuguese and French;
2. direct parsing transfer with a lexicalized model trained on Occitan;
3. delexicalized cross-dialectal parsing with a delexicalized model trained on Occitan.

The first scenario is our point of comparison, given the fact that it has already been used in pre-processing Occitan texts. The second scenario (direct transfer of lexicalized models trained on Occitan) corresponds to the most straightforward strategy: since all the varieties belong to the same language, a model trained on one of them should be able to process others. In the third scenario, we are looking to investigate if the delexicalization benefits the model by allowing it to abstract the lexical variation or hurts it by the fact that it reduces the amount of information available for learning.

We further refine the second and third scenario by using Lengadocian as the basis for the training corpus, then adding training material in each of the other dialects. This is done in order to evaluate if these additions lend robustness to the model so as to make it sufficient for processing all Occitan varieties or if variety-based parsing should be considered in the future.

4.2 Results and Discussion

Each of the models (9 in total) is evaluated on a test sample in each of the dialects. Train and test sample sizes are given in Table 6. The results are given in Table 7 as LAS and UAS scores[6]. Since the CoNLL shared tasks in 2006 (Buchholz and Marsi, 2006) and 2007 (Nivre et al., 2007), these metrics are widely used in dependency parsing evaluations and can be considered as a *de facto* standard in the domain. In our context (automatic pre-annotation intended for manual validation), both metrics can help estimate the extent of human intervention needed: the LAS gives us the percentage of tokens that need no correction, whereas UAS indicates how much of the tree structure will need no modification. Since correcting the tree structure is more time-consuming than simply changing a syntactic label, if a model has lower LAS, but significantly higher UAS than another, it can be more adapted to our purpose.

For each dialect, the best-performing model in terms of LAS is given in bold, and the worst in italics.

Sample	Lengadocian	Gascon	Lemosin	Provençau	UD
train	17 081	3 635	905	867	912 121
test	1 894	416	418	408	na

Table 6: Train and test sample sizes for each of the subcorpora (in tokens)

The best results for Lengadocian and Provençau were achieved by the delexicalized model trained on a combination of Lengadocian and Gascon, whereas the best scores for Gascon and Lemosin were achieved by the lexicalized models trained on the combination of Lengadocian and the dialect in question. However, the delexicalized model trained on Lengadocian and Gascon was on par with the best performing model for Lemosin in UAS, and it scored second best in LAS and UAS on Gascon. It can therefore be considered as the most useful across the board.

[6] LAS (labelled attachment score): see section 3.1. UAS (unlabelled attachment score): the percentage of tokens for which the parser correctly identifies the head, regardless of the label.

Model	Lengadocian		Gascon		Lemosin		Provençau	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
Lex_Leng	79.1	88.6	77.2	87.7	80.6	85.6	73.3	83.4
Lex_Leng+Gasc	79.4	88.4	**79.1**	**89.9**	80.6	**86.1**	76.2	86.0
Lex_Leng+Lem	78.5	88.2	75.7	86.8	**81.3**	**86.1**	74.0	84.3
Lex_Leng+Prov	79.0	88.6	77.6	88.0	80.7	85.9	74.7	85.0
Delex_Leng	79.5	88.9	77.9	87.8	80.6	85.4	76.7	85.5
Delex_Leng+Gasc	**80.2**	**89.2**	77.9	87.7	80.1	**86.1**	**77.0**	**86.3**
Delex_Leng+Lem	79.6	88.7	76.7	86.8	80.6	85.4	76.7	85.3
Delex_Leng+Prov	79.0	88.2	77.9	87.5	79.4	84.7	76.5	84.8
Delex_UD	*66.5*	*76.4*	*65.4*	*76.6*	*71.0*	*77.9*	*67.4*	*78.0*

Table 7: Parsing evaluation results

It is also interesting to note that despite the observed non-lexical variation in Gascon, this dialect does not seem to be the hardest to parse. Somewhat surprisingly, it is on Provençau that the models almost systematically obtain the lowest results. It remains to be determined if this is due to the properties of the dialect itself or potentially to the properties of the test sample.

More globally, the delexicalized models almost systematically outperformed their lexicalized counterparts. This seems to indicate that abstracting away from the lexical level allows for better generalization when dealing with different varieties of Occitan. However, this effect could also be due to other factors, such as the limited sample size and some genre variation in Lengadocian and Gascon data. We will therefore repeat these evaluations once the samples are extended. It is also interesting to note that the addition of Gascon seems to make the model the most robust, but this can also be due to the fact that the Gascon train sample is larger than the other two (3K tokens vs around 900 tokens).

Another interesting observation that can be made is that the delexicalized cross-lingual model trained on UD corpora in Italian, Portuguese and French scored systematically the worst, with 10-15 points of difference compared to the best-performing model for the given dialect. It is worth pointing out that the cross-lingual model's training corpus is an order of magnitude greater than for the models trained on Occitan: it contains 912K tokens, whereas the various training samples in Occitan contain between 17,5K and 20,5K tokens. This indicates once again that a small amount of annotated data in the target language can be as useful (or more useful, as we can see here) than a truly large corpus in a related language. This fact underlines the importance of developing data sets for under-resourced languages: even though transfer techniques for NLP tasks are useful, resources in each given language can be particularly valuable in achieving solid processing results.

5 Conclusion

In this paper we presented an extension of an existing Occitan treebank based on the Lengadocian dialect with content in three additional dialects, namely Gascon, Lemosin and Provençau. The resulting corpus contains 25K tokens, it is annotated following Universal Dependencies guidelines and will be made available as part of the November 2020 release.

Our experiments in parsing show that parsing models trained on dialects of Occitan achieve better results than a delexicalized model trained on UD corpora of Romance languages even though the latter is much larger (20k vs 900K tokens). This underlines once again the need to foster development of corpora for low-resourced languages: our results indicate that even a small amount of data in the target language can be expected to yield important improvements. We also observe that delexicalized models trained on Occitan dialects perform better than their lexicalized counterparts. While this is in line with the observed degree of lexical variation across Occitan dialects, it remains to be confirmed on larger amounts of data. On the other hand, contrary to our expectations, the syntactic variation observed in Gascon did not seem to overly affect the models' performance. The most useful model across the board was the delexicalized model trained on a combination of Lengadocian and Gascon content. Thanks to these experiments, we will be able to continue our work on corpus enrichment using a more efficient parsing model. Hopefully, this will further simplify the work of human annotators and allow for faster and easier additions to the treebank, ultimately leading to the creation of robust parsing models capable of processing all Occitan texts.

Acknowledgements

The present work is supported by the EFA 227/16 LINGUATEC Project, financed by the POCTEFA Interreg European funds.

References

Pierre Bec. 1970. *Manuel pratique de philologie romane*, volume Vol. 1. Picard.

Pierre Bec. 1995. *La langue occitane*. PUF, 6th edition.

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. 2018. Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard. In *International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May.

Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. A dependency treebank of Urdu and its evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop (LAW 2012)*, pages 157–165, Jeju Island, South Korea. Association for Computational Linguistics (ACL).

Myriam Bras and Marianne Vergez-Couret. 2016. BaTelÒc : a Text Base for the Occitan Language. In Vera Ferreira and Peter Bouda, editor, *Language Documentation and Conservation in Europe*, pages 133–149. Honolulu: University of Hawaï Press .

Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benazet Dazéas. 2020. Loflòc : Lexic obèrt flechit occitan. In Jean-François Courouau, editor, *Fidélités et dissidences (Actes du XIIe congrès de l'Association Internationale d'Études Occitanes)*, Albi. Centre d'Etude de la Littérature Occitane.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL2006)*, pages 149–164, New York City, USA. Association for Computational Linguistics (ACL).

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 845–850.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Aleksandra Miletic, Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, and Marianne Vergez-Couret. 2019a. Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan. TALN19, July. Poster.

Aleksandra Miletic, Myriam Bras, Louise Esher, Jean Sibille, and Marianne Vergez-Couret. 2019b. Building a treebank for Occitan: what use for Romance UD corpora? In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 2–11, Paris, France, August. Association for Computational Linguistics.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. Building a Universal Dependencies Treebank for Occitan. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2932–2939, Marseille, France, May. European Language Resources Association.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 915–932. Association for Computational Linguistics (ACL).

Jean Sibille. 2002. Ecrire l'occitan : essai de présentation et de synthèse. In Dominique Caubet, Salem Chaker, and Jean Sibille, editors, *Les langues de France et leur codification. Ecrits divers – Ecrits ouverts*, Paris, France, May. Inalco / Association Universitaire des Langues de France, L'Harmattan.

Arne Skjærholt. 2013. Influence of preprocessing on dependency syntax annotation: speed and agreement. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 28–32.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349.

Larraitz Uria, Ainara Estarrona, Izaskun Aldezabal, Maria Jesús Aranzabe, Arantza Díaz De Ilarraza, and Mikel Iruskieta. 2009. Evaluation of the syntactic annotation in EPEC, the reference corpus for the processing of Basque. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 72–85. Springer.

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

Marianne Vergez-Couret. 2016. Description du lexique Loflòc. Research report, CLLE-ERSS, Apr.

Atro Voutilainen and Tanja Purtonen. 2011. A double-blind experiment on interannotator agreement: The case of dependency syntax and finnish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 319–322.