Building a Universal Dependencies Treebank for Occitan

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, Jean Sibille

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2932–2939
Marseille, 11–16 May 2020

Building a Universal Dependencies Treebank for Occitan

Aleksandra Miletic*, Myriam Bras*, Marianne Vergez-Couret, Louise Esher***
Clamença Poujade*, Jean Sibille*
*CNRS UMR 5263 CLLE-ERSS (University of Toulouse), **EA 3816 FoReLLIS (University of Poitiers)
{aleksandra.miletic, myriam.bras, louise.esher, clamenca.poujade, jean.sibille}@univ-tlse2.fr
marianne.vergez.couret@univ-poitiers.fr

Abstract

This paper outlines an ongoing effort to create the first treebank for Occitan, a low-resourced regional language spoken mainly in the south of France. We briefly present the global context of the project and report on its current status. We adopt the Universal Dependencies framework for corpus annotation. Our methodology is based on two main principles. Firstly, we rely on pre-processing using existing tools (taggers and parsers) to facilitate the work of human annotators, mainly through a delexicalized cross-lingual parsing approach. Secondly, we use agile annotation to ensure annotation quality. We present the results available at this point: annotation guidelines and an initial corpus annotated with PoS tags, lemmas and syntactic dependencies.

Keywords: Occitan, treebank, Universal Dependencies, agile annotation, delexicalized cross-lingual parsing, low-resource languages

1. Introduction and Background

Low-resourced regional, non-official or minority languages often find themselves in a similar situation: building NLP resources requires substantial human and financial resources, but given their status and their often limited number of speakers, investing in NLP research on these languages is typically not seen as profitable. Nonetheless, these languages are part of the world's cultural heritage and their preservation and study can shed significant light on various scientific questions, be it in theoretical or contrastive linguistics, in linguistic typology, or in NLP itself. Luckily, this fact has been recognized both by the NLP community and by cultural institutions, leading to specialized workshops and conferences, but also to greater financial support from official sources. Occitan is one of the languages that has benefited from this paradigm shift.

1.1. Occitan

Occitan is a Romance language spoken in a large area in the south of France, in several valleys in Italy and in the Aran valley in Spain. As illustrated in example (1), it shares numerous trademark traits of the Romance language family, such as: overt inflection for number and gender on all members of the NP; overt inflection for tense, aspect, mood, person and number on finite verbs; relatively free word order; and non-obligatory subject pronouns (Olivieri and Sauzet, 2016). As such, it is closer to Catalan, Spanish and Italian than to French or to regional languages of northern France. Like many other low-resourced languages, Occitan is not standardized. It has six varieties organized in dialectal groups (Auvernhàs, Gascon, Lengadocian, Lemosin, Provençau and Vivaro-Aupenc). There is no universal spelling standard, but rather two different spelling norms, one called the *classical*, based on the Occitan troubadours' medieval spelling, and the other closer to the French language conventions (Sibille, 2000). This diversity, which manifests itself on the lexical and morphological levels and also in the spelling, makes Occitan particularly challenging for NLP.

Nevertheless, some recent efforts have provided first elements towards endowing Occitan with essential NLP resources. Two of them are described below.

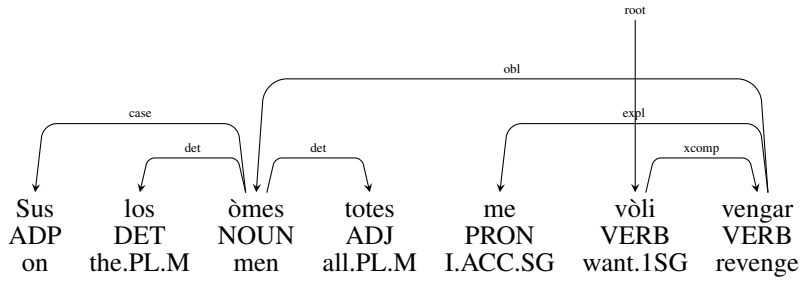
1.2. First Resource-Building Endeavours: the RESTAURE project

The main goal of the RESTAURE project (2016-2018) was to develop electronic resources and processing tools for three regional languages of France: Alsatian, Occitan and Picard. Although recognized as part of the cultural heritage of France by the constitutional amendment Article 75-1, these languages have no official status in France, and as such they have suffered from a lack of institutional support. The idea behind the RESTAURE project was to foster collaborative work on these languages, which at the time faced similar challenges concerning NLP tools. RESTAURE also represents the earliest endeavours to lend impetus to the preservation and dissemination of Occitan through the creation of digital resources, resulting in the creation of an electronic lexicon (Vergez-Couret, 2016; Bras et al., 2017) (850K entries), a textual database of 3,4M words (Bras and Vergez-Couret, 2016) and a PoS tagged corpus of 12K tokens (Bernhard et al., 2018). However, these resources remain relatively small compared to those available for well-resourced languages and Occitan does not yet have a syntactically annotated corpus. This point is being addressed in the current LINGUATEC Project.

1.3. Current Endeavours: the LINGUATEC Project

LINGUATEC is a European cross-border cooperation project, part of the France-Spain-Andorra POCTEFA Interreg Program for 2014-2020, which aims to promote knowledge transfer in language technologies. The project partners are based in France and Spain and work on Aragonese, Basque and Occitan. The goal is to develop new linguistic resources and tools for these languages in order to advance their digital development and dissemination, and to provide their speakers with innovative applications (automatic

(1)



'Suls òmes totes me vòli vengar.' ('I want revenge on all men.')

translation, spell- and grammar-checking, speech recognition and speech synthesis). Since Occitan lacks a syntactically annotated corpus and a parser, these resources were judged to be of highest priority for this language. The remainder of this paper describes our work creating a seed treebank for Occitan on which parsing experiments and further annotation work can be based.

In Section 2., we describe the annotation framework chosen for this project. Section 3. gives details on the annotation methodology we are using in order to ensure efficient, high-quality manual annotation. Section 4. describes the current status of the corpus. Finally, we draw conclusions and indicate directions for our future work in Section 5.

2. Applying Universal Dependencies Framework to Occitan

We adopt the Universal Dependencies (UD) Framework for our corpus. Universal Dependencies (Nivre et al., 2016) is a treebank building project whose goal is to create syntactically annotated corpora in a range of languages using shared annotation principles. Such an approach has the advantage of producing comparable linguistic annotation across corpora, which in turn facilitates research in cross-lingual parsing and machine translation, but also in linguistic typology and contrastive linguistics. Since its first release in January 2015, the UD corpus collection has grown continuously: its latest version at the time of writing (v2.5) contains 157 treebanks in 90 different languages. Adopting this framework for our project thus has a double advantage: it ensures resource visibility for our future corpus, and it also allows us to use the already existing UD annotation guidelines instead of defining our own from scratch. However, the universal character of the annotation choices made by UD results in some specific demands at different levels of processing. The most important aspects of applying these requirements to Occitan are given below.

2.1. Tokenisation

UD guidelines require that the texts be tokenized into *syntactic* units. Therefore, all *orthographic* units embodying more than one syntactic unit need to be split into separate tokens. Occitan has contracted article forms, in which a preposition and a definite article are fused, such as *sul* 'on.the.M.SG' < *sus* 'on' + *lo* 'the.M.SG'. In the column-based format used by UD (called CoNLL-U), a double representation of such tokens is recommended: one line with the original form followed by two lines with the split forms (cf. Table 1, illustrating the same sentence as example 1). Identification and tokenization of these forms were done using Python scripts based on closed lists of possible contracted article forms.

2.2. Part-of-Speech Tagging

Universal Dependencies implements a two-level morpho-syntactic annotation: the Part-of-Speech tagging is done using 17 basic tags such as verb (`VERB`), common noun (`NOUN`), proper noun (`PROPN`), auxiliary (`AUX`), etc. More detailed, language-specific morpho-syntactic information can be encoded through a rich system of morpho-syntactic features. The project proposes a set of 23 lexical and morphological features (e.g. `Gender, Animacy, Mood, Tense`, etc.) from which the relevant features can be selected for each language.

Currently, we use only the global PoS tags and no morphosyntactic features. This decision simplifies the manual annotation process, but has the disadvantage of incurring some information loss compared to the more detailed PoS tagset used in the RESTAURE project, which was based on the GRACE standard (Rajman et al., 1997). In the GRACE PoS tagset, grammatical subcategories are systematically taken into account (e.g., there are 8 different pronoun tags based on the pronoun subcategories). However, given the limited duration of our project (2018-2020), we decided to leave the annotation of grammatical subcategories and other morphosyntactic features for later stages of the project. For more details, see Section 4.

Table 4 lists all the PoS tags used in the corpus. Note that the UD tagset contains one additional tag, `SYM` (*symbol*) that has no occurrences in our corpus at this point.

For the application of the UD PoS-tagging guidelines to Occitan, some deviations were necessary. The most important one concerns the possessive forms. UD Guidelines require that all such forms be tagged either as pronouns or as determiners (cf. `https://universaldependencies.org/u/pos/all.html#al-u-pos/DET`). However, certain possessive forms in Occitan more closely resemble the category of adjectives. Compare *mon filh* 'my son', which uses the possessive determiner *mon* 'my' and does not appear with an article in Lengadocian, with the alternative construction *lo mieu filh* (lit. 'the my son'), which has the same meaning, but uses a different possessive form that can be preceded by a determiner and can appear both to the left and to the right of the noun (cf. *lo filh mieu*, lit. 'the son my'). Unlike in Italian, in which the

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1-2	Suls	_	_	_	_	_	_	_	_
1	sus	sus	ADP	Sp	_	3	case	_	Gloss=sur
2	los	lo	DET	Da	_	3	det	_	Gloss=le
3	òmes	òme	NOUN	Nc	_	7	obl	_	Gloss=homme
4	totes	tot	DET	Ai	_	3	amod	_	Gloss=tout
5	me	me	PRON	Pp	_	7	obj	_	Gloss=me
6	vòli	voler	VERB	Vm	_	0	root	_	Gloss=vouloir
7	vengar	vengar	VERB	Vm	_	6	xcomp	_	Gloss=venger
8	.	.	PUNCT	F	_	0	punct	_	Gloss=.

Table 1: Tokenizing contracted articles. UPOS=UD PoS tag, XPOS=language-specific PoS tag, FEATS=morphosyntactic features, HEAD=syntactic governor, DEPREL=dependency label, DEPS=enhanced dependencies if any, MISC=other information.

prototypical possessive constructions are preceded by an article, these are two different constructions using two distinct possessive paradigms. In order to distinguish between them, we annotate the latter as adjectives (since their syntactic behaviour is identical to that of adjectives), and their possessive character will be expressed as a morpho-syntactic feature in future versions of the corpus.

2.3. Syntactic Annotation

At the syntactic level, Universal Dependencies proposes a set of 37 basic syntactic dependency labels, denoting syntactic relations such as nominal subject (`nsubj`), direct object (`obj`), nominal modifier (`nmod`), etc. There is also a much larger set of two-level labels[1] encoding finer syntactic distinctions. These labels can be language-specific or more general.

At the current stage of our treebank building process, we only use the basic dependency labels. Table 5 gives all the labels used in our corpus. The UD syntactic tagset also includes `clf` (*classifier*), `reparandum` (overridden disfluency in spoken data), `list` (list element) and `goeswith` (ill-tokenized element), but these are absent from our data.

In the syntactic annotation process, we follow the global UD Guidelines, but have also defined some language-specific annotation rules. Some of these are based on the examples found in the UD corpora for other Romance languages (mostly French and Catalan). Others we proposed ourselves since they pertain to constructions specific to Occitan, such as the periphrastic constructions with *tornar 'return' + INF* or *V + tornar 'return'* to mark the repetition of an event.

UD Guidelines analyse an infinitive complement that inherits its subject from the main proposition as an open clausal complement (`xcomp`). This analysis captures well the nature of the *tornar + INF* construction and we apply it to all such occurrences (cf. examples 2 and 3). On the other hand, this analysis is not suitable for the *V + tornar* construction, in which the verb *tornar* appears in the same position as adverbials *tornamai* and *tòrna* ('again'); in this construction, we consider *tornar* as an adverbial (cf. example 4).

[1] `https://universaldependencies.org/ext-dep-index.html`

3. Optimizing the Annotation Process

Since our goal is to create a gold standard corpus for Occitan, our annotations are done manually. However, it is well-known that such an approach is both time-consuming and error-prone. In order to mitigate this, we combine two annotation strategies. Firstly, we use automatic data pre-annotation to lighten the task for human annotators and thus accelerate the process, relying on the well-established positive effects of this approach on various types of linguistic annotation (Xue N., 2005; Fort, 2012; Tellier I., 2014). Secondly, we adopt the agile annotation method (Voormann and Gut, 2008) to ensure the quality of the manual annotation.

3.1. Automatic Pre-processing

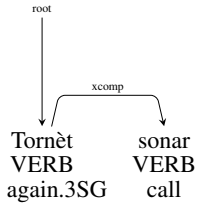
Given the absence of training data for Occitan at the start of our project, we trained a parsing model on existing UD corpora for Romance languages using delexicalized cross-lingual parsing. This technique consists in training a parsing model on a delexicalized corpus of a source language (i.e., using only PoS tags and morphosyntactic features and ignoring tokens and lemmas) and then using the model to process data in the target language. Specifically, we used 14 corpora in the 8 Romance languages from the UD collection to train 21 different delexicalized parsing models and tested them on a manually annotated Occitan sample of 1100 tokens. The top-performing model based on the LAS score[2] was trained on a combination of French, Portuguese and Italian data. Thanks to the pre-annotation of Occitan texts produced by this model, the manual annotation speed went from 340 tokens/h (for a fully manual annotation) to 650 tokens/h. The annotator working on this experiment also reported greater ease from an ergonomic point of view. For a detailed account of this part of our work, see (Miletic et al., 2019b). The delexicalized model was subsequently used to pre-annotate new texts in our treebank.

3.2. Agile Annotation

The overall organisation of our annotation process is given in Figure 1. Following Fort et al. (2012), we divide the work into four stages: campaign preparation (blue), pre-campaign (yellow), manual annotation campaign (green)

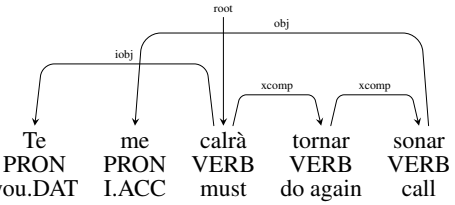
[2] *Labelled Attachment Score:* percentage of tokens for which a parsing model determines the correct governor and the correct dependency label.

(2)



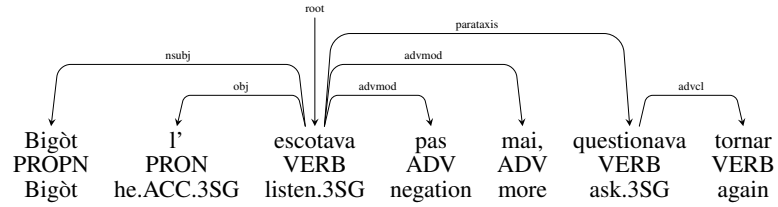
'Tornèt sonar.' ('He called again.')

(3)



'Te me calrà tornar sonar.' ('You'll have to call me again.')

(4)



'Bigòt l'escotava pas mai, questionava tornar.' ('Bigòt was no longer listening to him, he asked again')

and corpus finalisation (red). For the campaign stage of the process, we adopt the agile annotation approach defined by Voormann and Gut (2008): annotation is iterative, with each iteration followed by an evaluation step. A similar approach has been successfully used on Serbian, adapted here for Occitan following (Miletic, 2018; Miletic et al., 2019c).

1. **Campaign preparation** includes defining the tagset, selecting texts for the corpus, gathering other relevant resources (such as a morphosyntactic lexicon), choosing pre-annotation tools (taggers, lemmatisers and parsers) and the manual annotation interface, and preparing the initial version of the annotation guidelines.

2. **Pre-campaign** involves recruiting annotators and training them on the guidelines and the use of the annotation interface.

3. **Annotation campaign** comprises iterative cycles of manual annotation and evaluation, and, since we use automatic pre-annotation, also includes tool training and automatic pre-processing of the data. Initial tool training is performed with minimal resources resulting from the first two stages of the process and different compensation strategies, such as cross-lingual parsing (cf. Section 3.1.). A first sample of the corpus is then automatically pre-processed and manually corrected. During the evaluation step, inter-annotator agreement is calculated, annotation problems are discussed and the annotation guidelines are updated. Each subsequent iteration includes training a parser on newly annotated data and using the new model to annotate fresh texts, thus increasing the quality of the pre-annotation and facilitating the manual annotation.

4. **Finalisation** involves annotation coherency checks and corpus distribution. As the annotation guidelines are updated after each annotation cycle, it is essential to harmonize annotations in order to ensure coherent linguistic analysis throughout the corpus. Once this step is done, the corpus can be published.

Fort et al. (2012) also advocate for a clear attribution of roles in the annotation project. In our case, the first author of the paper acts as the campaign manager, whose main duties are campaign organization and time management, and also as an NLP expert, in charge of the automatic pre-annotation strategy, data processing and corpus curation. The five Occitan-speaking authors have the role of annotators tasked with enriching the corpus with different levels of linguistic analysis, and also provide linguistic expertise crucial to the writing of the guidelines.

4. Corpus Description and Project Status

We have successfully completed the campaign preparation and the pre-campaign and are currently in the third iteration of the campaign stage. The state of the annotation process is given in Table 2.

The corpus contains around 20K tokens representing texts written by 20 different authors and spanning 5 genres (literature, newspaper, encyclopedia, blog and scientific text). For the time being, the content is based on only one dialect

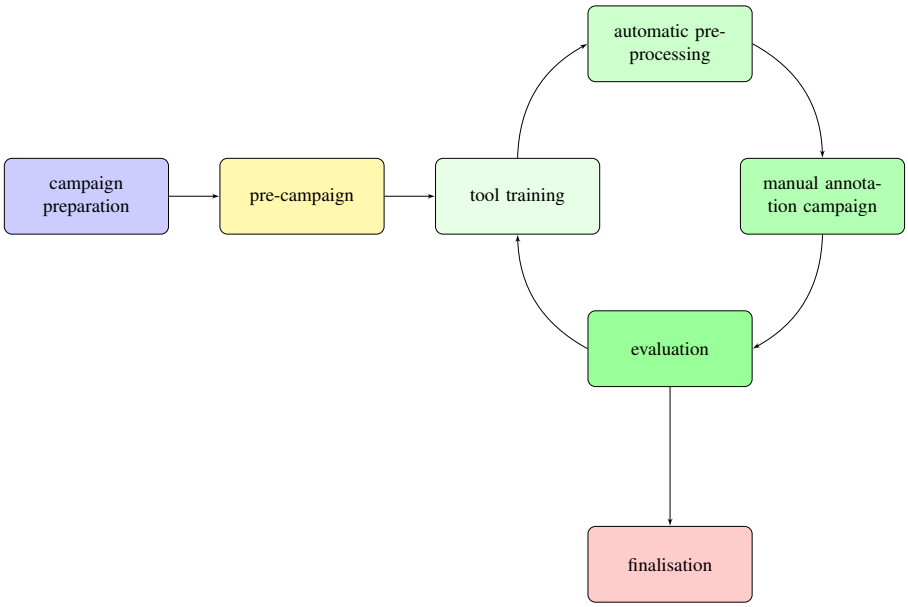


Figure 1: Annotation process

Origin	Round	Genre	Tok.	Annotation status		Inter-annot. agreement	
				Tok+PoS+Lem	Syntax	Cohen's *kappa*	%
RESTAURE	Round1	newspaper	974	yes (by conversion)	completed	0.88	90.9 %
	Round2	literature	3217	yes (by conversion)	completed	0.81	82.3 %
New	Round3	literature	7486	yes (manually)	completed	(single annotation)	
		encyclopedia	1527	yes (manually)	completed	(single annotation)	
		blog	607	yes (manually)	completed	(single annotation)	
	Round4	literature	3802	yes (manually)	no	-	-
		encyclopedia	1966	yes (manually)	no	-	-
		Completed	13811				
		Remaining	6535				
		TOTAL	20346				

Table 2: Corpus content

and one spelling norm: we selected texts in Lengadocian written in the classical spelling. This choice was made in order to avoid data sparsity issues while working on small amounts of data (cf. Section 1.1.). We chose Lengadocian for its central position in the Occitan dialect continuum from the linguistic point of view, but also because it is the dialect for which we have already built lexical resources (cf. Section 1.2.). Once we have produced a training corpus yielding stable parsing models in these conditions, other linguistic varieties of Occitan will be added to the treebank.

We benefit from the work done in the RESTAURE Project by integrating all Lengadocian texts from the RESTAURE corpus, for a total of around 4K tokens[3]. These were already tokenized, lemmatised and tagged. However, the annotation was done following the GRACE and EAGLES annotation guidelines (Rajman et al., 1997). The initial annotation was therefore converted into the Universal Dependencies tagset (Miletic et al., 2019a).

The remainder of the content comes from previously unprocessed texts in Lengadocian (around 16K tokens) which needed to be annotated from scratch.

As for other aspects of the campaign preparation, Universal Dependencies framework was adopted for the annotation process (cf. Section 2.). The automatic pre-annotation described in Section 3.1. is done using the Talismane NLP suite (Urieli, 2013), which has already been successfully used on Occitan in the RESTAURE project (Vergez-Couret and Urieli, 2015). The Brat rapid annotation tool (Stenetorp et al., 2012) was selected as the manual annotation interface.

All manual annotations (PoS tagging, lemmatization, dependency annotation) were carried out by our team of five annotators. They were trained by the campaign manager, who has extensive experience in dependency syntax, UD guidelines and the Brat annotation interface (although not in Occitan).

As shown in Table 2, the entire corpus has now been tokenized, PoS tagged and lemmatized. We have also completed the syntactic annotation of around 13K tokens (annotation rounds 1-3). The first two rounds of annotation were carried out on relatively small samples so as to allow

[3]The remaining content of the RESTAURE corpus is in dialects other than Lengadocian.

Annotated tokens:	13806
Types:	3499
Lemmas:	2435
No. of sentences:	867
Mean sent. length	15.9

Table 3: Annotated corpus information

for a quick update of the guidelines based on the questions arising from the data.

These batches were annotated by two annotators each and their productions were adjudicated in order to create the final version of the annotation. We report the inter-annotator agreement on these samples both in terms of Cohen's *kappa*[4] and as a simple agreement ratio (percentage of consistent annotations between annotators). Neither of these measures is perfect: the former is intended for classification tasks and dependency annotation is more complex, whereas the latter does not correct for chance agreement. However, at the time being there seems to be no consensus on an alternative measure, and both Cohen's *kappa* and agreement ratio 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). We provide them as a simple means of assessing the annotation coherence in the corpus.

The drop in agreement between rounds 1 and 2 can be explained by the change in genre: round 2 was based on literary texts as opposed to newspaper articles in round 1, and the annotators reported encountering longer and syntactically more complex sentences. Texts from round 3 were treated by one annotator only for the time being. A part of this batch of texts will be doubly annotated in order to check if the agreement is stabilizing.

Some basic counts for the annotated part of the corpus are given in Table 3, and the distribution of PoS tags and dependency labels is given in Table 4 and Table 5, respectively. This version of the corpus is available for download under the `CC BY-NC-SA 4.0` license[5] at the following address: `https://zenodo.org/record/3708268#.XmuLQ3VKg5k`.

The end of all annotations (cf. Round 4) is scheduled for 15 April 2020. The full corpus (20K tokens) will be submitted for publication as part of the Universal Dependencies v2.6 in May 2020.

5. Conclusions and Future Work

We presented an ongoing endeavour to produce the first treebank for Occitan, a low-resourced language that has suffered from a lack of institutional support until recently. We have gathered a 20K word corpus consisting of texts in one dialect of Occitan (Lengadocian), following one spelling norm (classical norm), spanning 5 genres (newspaper, literature, encyclopedia, blog, scientific text), and written by several authors. The PoS tagset and the dependency labels are based on the Universal Dependencies framework with some adaptations to accommodate constructions specific to Occitan. We defined our annotation methodology following previous works in Occitan PoS annotation and

Tag	Count	Tag	Count
ADJ	521	NUM	135
ADP	1642	PART	6
ADV	786	PRON	1100
AUX	349	PROPN	340
CCONJ	388	PUNCT	2063
DET	1952	SCONJ	256
INTJ	65	VERB	1855
NOUN	2345	X	8

Table 4: PoS tag counts in the corpus

Label	Meaning	Count
acl	adjectival clause	274
advcl	adverbial clause	176
advmod	adverbial modifier	683
amod	adjectival modifier	393
appos	apposition	60
aux	auxiliary	180
case	case mark	1354
cc	coordinating conjunction	377
ccomp	clausal complement	106
compound	compound word element	2
conj	coordination conjunct	430
cop	copula	182
csubj	clausal subject	3
dep	dependency	17
det	determiner	1944
discourse	discourse element	48
dislocated	dislocated element	51
expl	expletive element	257
fixed	element of a fully grammaticalized MWE	132
flat	element of an exocentric construction	84
iobj	indirect object	126
mark	subordination mark	456
nmod	nominal modifier	565
nsubj	nominal subject	591
nummod	numeral modifier	78
obj	direct object	725
obl	oblique dependent	841
orphan	element orphaned by ellipsis	38
parataxis	paratactic element	188
punct	punctuation	2063
root	sentence root	839
vocative	vocative	37
xcomp	open clausal complement	323

Table 5: Dependency labels in the corpus

[4]Cohen's *kappa* was calculated using an integrated option of our chosen parser (Urieli, 2013).

[5]`https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en`

Serbian and French dependency annotation. Our annotation process integrates an agile annotation approach with the automatic pre-annotation of the data. It allowed us to complete the annotation of 13K tokens with PoS tags, lemmas and syntactic dependencies in an efficient and ergonomic manner. The annotation process will be completed and the corpus submitted for publication in the Universal Dependencies v2.6 release in May 2020. These results also speak to the fact that low-resourced languages can benefit from resources and experience of better-resourced languages through the use and adaptation of existing annotation standards, tools and resources.

6. Acknowledgements

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

7. Bibliographical References

Bernhard, D., Ligozat, A.-L., Martin, F., Bras, M., Magistry, P., Vergez-Couret, M., Steiblé, L., Erhart, P., Hathout, N., Huck, D., Rey, C., Reynés, P., Rosset, S., Sibille, J., and Lavergne, T. (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.

Bhat, R. A. and Sharma, D. M. (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).

Bras, M. and Vergez-Couret, M. (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 .

Bras, M., Vergez-Couret, M., Hathout, N., Sibille, J., Séguier, A., and Dazéas, B. (2017). Lofl̀oc : Lexic obert flechit occitan. In *XIIème Congrès de l'Association Internationale d'Etudes Occitanes*, Albi, France, July.

Fort, K., Nazarenko, A., and Rosset, S. (2012). Modeling the complexity of manual annotation tasks: a grid of analysis. In *International Conference on Computational Linguistics (COLING 2012)*, pages 1–16, Mumbai, India, 08/12 au 15/12.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus.* Thèse de doctorat en informatique, Université de Paris XIII.

Miletic, A., Bernhard, D., Bras, M., Ligozat, A.-L., and Vergez-Couret, M. (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.

Miletic, A., Bras, M., Esher, L., Sibille, J., and Vergez-Couret, M. (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, 26 August. Association for Computational Linguistics.

Miletic, A., Fabre, C., and Stosic, D. (2019c). De la constitution d'un corpus arboré à l'analyse syntaxique du serbe. *Traitement Automatique des Langues*, January.

Miletic, A. (2018). *Un treebank pour le serbe : constitution et exploitations.* Thèse de doctorat en linguistique, Université de Toulouse Jean Jaurès.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., and others. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Olivieri, M. and Sauzet, P. (2016). Southern gallo-romance (occitan). In Adam Ledgeway et al., editors, *The Oxford Guide to the Romance Languages*, pages 319–349. Oxford University Press, Oxford.

Rajman, M., Lecomte, J., and Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.

Sibille, J. (2000). Ecrire l'occitan : essai de présentation et de synthèse. In Dominique Caubet, et al., editors, *Les langues de France et leur codification. Ecrits divers – Ecrits ouverts,*, Paris, France, May. Inalco / Association Universitaire des Langues de France, L'Harmattan.

Skjærholt, A. (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.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tellier I., Eshkol-Taravella I., D. Y. W. I. (2014). Peut-on bien chunker avec de mauvaises étiquettes pos ? In *Actes de TALN*, pages 125–136.

Uria, L., Estarrona, A., Aldezabal, I., Aranzabe, M. J., De Ilarraza, A. D., and Iruskieta, M. (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.

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

Vergez-Couret, M. and Urieli, A. (2015). Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*, Caen, France.

Vergez-Couret, M. (2016). Description du lexique Lofl̀oc. Research report, CLLE-ERSS, April.

Voormann, H. and Gut, U. (2008). Agile corpus creation.

Corpus Linguistics and Linguistic Theory, 4:235–251, 12.

Voutilainen, A. and Purtonen, T. (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.

Xue N., Xia F., C. F.-D. P. M. (2005). The Penn Chinese TreeBank : Phrase structure anno-tation of a large corpus. *Natural language engineering*, 11 (02):207–238.