HAL open science

Building a treebank for Occitan: what use for Romance UD corpora?

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

To cite this version:

Aleksandra Miletic, Myriam Bras, Louise Esher, Jean Sibille, Marianne Vergez-Couret. Building a treebank for Occitan: what use for Romance UD corpora?. Syntax Fest, Aug 2019, Paris, France. hal-02380554

HAL Id: hal-02380554

https://hal.science/hal-02380554v1

Submitted on 26 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Building a treebank for Occitan: what use for Romance UD corpora?

Aleksandra Miletic*, Myriam Bras*, Louise Esher*,
Jean Sibille*, Marianne Vergez-Couret**
* CLLE-ERSS (CNRS UMR 5263), University of Toulouse Jean Jaurès, France
firstname.lastname@univ-tlse2.fr
** FoReLLIS (EA 3816), University of Poitiers, France

Abstract

This paper describes the application of delexicalized cross-lingual parsing on Occitan with a view to building the first dependency treebank of this language. Occitan is a Romance language spoken in the south of France and in parts of Italy and Spain. It is a relatively low-resourced language and does not have a syntactically annotated corpus as of yet. In order to facilitate the manual annotation process, we train parsing models on the existing Romance corpora from the Universal Dependencies project and apply them to Occitan. Special attention is given to the effect of this cross-lingual annotation on the work of human annotators in terms of annotation speed and ease.

1 Introduction

Occitan is a Romance language spoken across the south of France and in several areas of Italy and Spain. Although it has no official status in France, it has been recognized – among other regional languages – as part of the cultural heritage of France by the constitutional amendment Article 75-1 published in 2008. Ever since, there have been more efforts to strengthen the preservation and the dissemination of the language through the creation of electronic resources. The most notable such project was RESTAURE (Bernhard et al., 2018), which resulted in the creation of an electronic lexicon (Vergez-Couret, 2016) and a POS tagged corpus (Bernhard et al., 2018) for Occitan. However, Occitan does not yet have a syntactically annotated corpus. This paper presents the first efforts towards the creation of such a resource.

It is well-known that manual annotation is time-consuming and costly. In order to facilitate and accelerate the work of human annotators, we implement direct delexicalized cross-lingual parsing in order to provide an initial syntactic annotation. This technique consists in training a parsing model on a delexicalized corpus of a source language and then using the model to process data in the target language. The training is typically only based on POS tags and morphosyntactic features, whereas lexical information (i.e. the information related to the token and the lemma) is ignored. Thus, the model is able to parse the target language even though no target language content was present in the training corpus.

In the past, delexicalized cross-lingual parsing was used with mixed results due to the divergent annotation schemes in different corpora (McDonald et al., 2011). The Universal Dependencies project (Nivre et al., 2016) offers a solution to this issue: version 2.3 comprises over 100 corpora in over 70 different languages[1], all annotated according to the same annotation scheme. The use of such harmonized annotations has lead to cross-lingual parsing results consistent with typological and genealogical relatedness of languages (McDonald et al., 2013). These corpora have since been successfully applied to delexicalized parsing of numerous language pairs (Lynn et al., 2014; Tiedemann, 2015; Duong et al., 2015).

Lexicalized cross-lingual parsing was also considered as a possible solution, but was rejected for two main reasons. Firstly, to the best of our knowledge, there are no parallel corpora of Occitan that could have been of immediate use for techniques such as annotation projection. Secondly, Occitan data could have been adapted to lexicalized parsing through different techniques such as machine translation or devoweling (Tiedemann, 2014; Rosa and Mareček, 2018), but the effort needed for such an approach is not negligible. As already stated above, the work presented here was conducted as part of a corpus-building project, with the primary goal of accelerating the manual annotation process. The methods used to facilitate

[1] https://universaldependencies.org/

the annotation where therefore not to be more costly than manual annotation itself. Given this constraint, delexicalized cross-lingual parsing was chosen as the most straightforward approach.

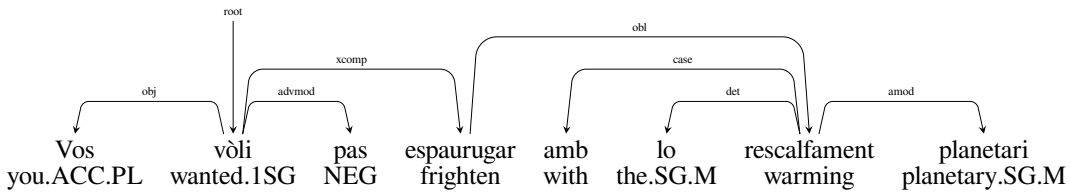
Direct delexicalized cross-lingual parsing has been used to initiate the creation of an Old Occitan treebank. Scrivner and Kübler (2012) used Catalan and Old French corpora for cross-lingual transfer of both POS tagging and parsing. Unfortunately, we were unable to locate the resulting corpus. We therefore decided to implement delexicalized cross-lingual parsing based on the Romance corpora made available by the UD project. In this paper we present the quantitative evaluation of this process, but also the effects of this technique on the work of human annotators in terms of manual annotation speed and ease.

The remainder of this paper is organized as follows. First, we give a brief linguistic description of Occitan (Section 2); in Section 3 we describe the resources and tools used in our experiments; we then present the quantitative evaluation of the parsing transfer (Section 4) and analyze the impact of this method on the manual annotation (Section 5). Lastly, we draw our conclusions and discuss future work in Section 6.

2 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. It shares numerous linguistic properties with several other Romance languages: it displays number and gender inflection marks on all members of the NP, and it has tense, person and number inflection marks on finite verbs (cf. example 1). It is a pro-drop language with relatively free word order and as such it is closer to Catalan, Spanish and Italian than to French and other regional languages from the north of France.

(1)



'I didn't want to scare you with global warming.'

Another crucial property of Occitan from the NLP point of view is that it has not been standardized. It has numerous varieties organized in 6 dialectal groups (Auvernhàs, Gascon, Lengadocian, Lemosin, Provençau and Vivaro-Aupenc). Also, there is no global spelling standard, but rather two different norms, one based on the Occitan troubadours' medieval spelling, and the other closer to the French language conventions (Sibille, 2000). This double diversity which manifests itself both on the lexical level and in the spelling makes Occitan particularly challenging for NLP.

To avoid the data sparsity issues that can arise in such a situation while working on small amounts of data, we decided to initiate the treebank building process with only two dialects: Lengadocian and Gascon. Other dialects will be added once we produce a training corpus sufficient to generate stable parsing models for the first two dialects.

3 Resources and tools

To implement cross-lingual delexicalized parsing, we used the Romance language corpora from the UD project as training material, we created a manually annotated sample of Occitan to be used as an evaluation corpus, and we used the Talismane NLP suite to execute all parsing experiments. Each of these elements is presented in detail below.

3.1 UD Romance corpora

Universal Dependencies v2.3 comprises 22 different corpora in 8 Romance languages (Catalan, French, Galician, Italian, Old French, Portuguese, Romanian, and Spanish). These corpora vary in size (from 23K tokens in the PUD corpora in French, Italian, Portuguese and Spanish to 573K tokens in the FTB corpus

of French), as well as in terms of content: they include newspaper texts, literature, tweets, poetry, spoken language, scientific and legal texts.

Some of these corpora were excluded from our experiments. Some were eliminated based on the text genre. The Occitan corpus we are working on consists mainly in literary and newspaper texts. We therefore did not include corpora containing spoken language and tweets. Secondly, in order to ensure the quality of the parsing models trained on the corpora, we only selected those built through manual annotation or converted from such resources. Lastly, for practical reasons, we only kept the corpora that already had designated train and test sections. This resulted in a set of 14 corpora, but all 8 languages are represented (for the full list, see Section 4.1).

These corpora integrate different sets of morphosyntactic traits, and some of them implement a number of two-level syntactic labels. In order to maintain consistency between the training corpora, but also with the Occitan evaluation sample, no morphosyntactic traits were used in training, and syntactic annotation was reduced to the basic one-level labels.

3.2 Manually annotated evaluation sample in Occitan

In order to evaluate the suitability of the delexicalized models for the processing of our target language, we created an evaluation sample in Occitan. This sample contains around 1000 tokens from 4 newspaper texts, 3 of which are in Lengadocian and 1 in Gascon (cf. Table 1). The sample is tagged with UD POS tags, obtained by a conversion from an existing Occitan corpus which was manually tagged using EAGLES and GRACE tagging standards (Bernhard et al., 2018). As of yet, the sample contains no fine-grained morphosyntactic traits[2].

Sample	Dialect	No tokens	No POS	No labels
jornalet-atacs	Lengadocian	272	13	25
jornalet-festa	Lengadocian	353	13	24
jornalet-lei	Lengadocian	310	12	20
jornalet-estanguet	Gascon	217	12	24
TOTAL		1152	14	27

Table 1: Occitan evaluation sample

At the moment, the syntactic annotation is limited to first-level dependency labels (no complex syntactic labels). This is due to the fact that the annotation of this evaluation sample was in fact the first round of syntactic annotation in the project. It was therefore used to test and refine the general UD guidelines, but also to gather information as to which two-level labels may be necessary. The result of this analysis will be included in the next round of annotation.

The syntactic annotation of the sample was done manually using the brat annotation tool (Stenetorp et al., 2012). Each text was processed by one annotator who had extensive experience with dependency syntax, UD guidelines and the annotation interface (although not on Occitan), and one novice. The inter-annotator agreement on the sample in terms of Cohen's *kappa* (excluding punctuation marks) is 88.1. This can be considered as a solid result given that this was the very first cycle of annotation. All disagreements were resolved in an adjudication process, resulting in a gold-standard annotated sample.

3.3 Talismane NLP suite

For all parsing experiments described in this paper, we used Talismane (Urieli, 2013). It is a complete NLP pipeline capable of sentence segmentation, tokenization, POS tagging and dependency parsing. It currently integrates 3 algorithms: perceptron, MaxEnt, and SVM. The Talismane tagger has already been successfully used on Occitan for POS tagging in a previous project (Vergez-Couret and Urieli, 2015), on the outcomes of which the current project is founded. Talismane gives full access to the learning features, which can be defined by the user. Thus, it suffices to adapt the feature file in order to define the desired

[2]The original corpus annotation does encode some lexical traits, which will be recuperated and included in the UD conversion in immediate future. However, the original corpus does not contain any inflectional traits.

learning conditions: in our case, no lemma-based or token-based features were included in the feature set, which dispensed the user from the need to modify the learning corpora. This was particularly useful given the number of corpora used. However, numerous recent works have shown that tools based on neural networks outperform classical machine learning algorithms in tasks including dependency parsing, while often offering comparable practical advantages (Zeman et al., 2017; Zeman et al., 2018). One of the future steps in the continuation of this work will be to test neural network parsers on our data.

4 Transferring delexicalized parsing models to Occitan

We used Talismane's SVM algorithm to train models on the selected corpora. Learning was based on the POS tag features of the processed token and its linear and syntactic context, and different combinations thereof (34 features in total). Since the features were light, the training generated relatively compact models even for the largest corpora (the biggest at 130MB). The generated models were evaluated first on their respective test samples and then on the manually annotated Occitan sample. The results are discussed below.

4.1 Baseline evaluation

The goal of this first evaluation was to establish the baseline results for each model. This baseline was to be used to assess the stability of the models when transferred to Occitan. The results are given in Table 2. The corpus names contain the language code and the name of the corpus in lowercase. The top 5 models in terms of the LAS are highlighted in bold.

Corpus	Train size	Test size	LAS[3]	UAS[4]
ca_ancora	418K	58K	77.82	82.20
es_ancora	446K	52.8K	76.75	81.29
es_gsd	12.2K	13.5K	74.88	78.81
fr_partut	**25K**	**2.7K**	**82.41**	**84.60**
fr_gsd	364K	10.3K	78.51	81.81
fr_sequoia	52K	10.3K	78.29	80.71
fr_ftb	470K	79.6K	68.93	73.08
gl_treegal	16.7K	10.9K	73.91	78.79
it_isdt	**294K**	**11.1K**	**81.03**	**84.19**
it_partut	**52.4K**	**3.9K**	**82.66**	**85.22**
ofr_srcmf	136K	17.3K	69.41	79.09
pt_bosque	222K	10.9K	77.41	81.27
pt_gsd	**273K**	**33.6K**	**80.2**	**83.2**
ro_rrt	185K	16.3K	71.87	78.92
ro_nonstandard	155K	20.9K	65.59	75.45
es_ancora+gsd	458.2K	66.3K	73.14	78.24
fr_partut+gsd+sequoia	441K	23.3K	73.69	77.57
fr_partut+gsd+sequoia+ftb	911K	102.9K	74.87	78.55
it_isdt+partut	**346.4K**	**15K**	**81.78**	**84.66**
pt_bosque+gsd	495K	44.5K	76.09	81.47
ro_nonstand+rrt	340K	37.2K	67.21	76.06

Table 2: Baseline evaluation of models trained on UD Romance corpora

The LAS varies from 65.59 (ro_nonstandard) to 82.41 (fr_partut), and the UAS from 73.08 (fr_ftb) to 85.22 (it_partut), with the top 5 models acheiving an LAS > 80 and a UAS > 83. We also tested the option of merging several corpora in the same language (cf. the lower half of the table) under the supposition that, given the shared annotation scheme, this would equate to having a larger training corpus and boost the results. However, none of the combined corpora produced a model that surpassed the best performing individual model, although it_isdt+partut did score among the top 5. This seems to indicate that there are divergences

in the application of the UD annotation scheme between different corpora of the same language, resulting in inconsistent annotations in the merged corpora. Indeed, at least one such discrepancy was spotted in the French corpora during this work: the temporal construction *il y a* 'ago' is annotated in three different ways in the GSD, ParTUT and Sequoia corpora. Nevertheless, it should be noted that such effects can also be due to the fact that the content of the combined corpora was simply concatenated and not reshuffled, which may have had a negative effect on the learning algorithm.

Nevertheless, since the baseline performances were not necessarily directly indicative of the results that each model would achieve on Occitan, all models generated in this step were tested on Occitan too.

4.2 Evaluation on Occitan

Table 3 details the results of the parsing evaluation on the manually annotated Occitan sample presented in section 3.2. The models are listed from best to worst in terms of LAS. Since the test sample contains around 1000 tokens, a different annotation of a single token constitutes roughly a 0.1% change in the parsing scores. Therefore, the scores are rounded to one decimal point.

Train corpus	LAS	UAS	Train corpus	LAS	UAS
it_isdt	71.6	76.0	ca_ancora	68.6	75.2
it_isdt+partut	71.3	75.9	fr_sequoia	68.6	73.3
fr_partut+gsd+sequoia	70.8	75.7	es_gsd	67.8	73.4
fr_gsd	70.4	75.9	fr_ftb	67.4	72.5
pt_bosque	70.0	75.3	ro_rrt	67.1	72.2
it_partut	69.7	74.1	ro_nonstand+rrt	66.6	72.0
fr_partut+gsd+sequoia+ftb	69.6	74.4	pt_bosque+gsd	66.4	74.3
fr_partut	69.4	74.6	pt_gsd	63.1	73.3
es_ancora+gsd	69.1	74.9	ro_nonstand	60.2	72.7
es_ancora	69.0	75.3	ofr_scmrf	59.2	66.0
gl_treegal	68.7	73.4			

Table 3: Evaluation on the manually annotated Occitan sample. (Bold: models selected for further experiments.)

In this evaluation scenario, the LAS varies from 59.2 (ofr_scmrf) to 71.6 (it_isdt), whereas the UAS ranges from 66.0 (ofr_scmrf) to 76.0 (it_isdt). Rather surprisingly, among the top 5 models we find three based on French and Portugese corpora, although these languages are not traditionally considered as close to Occitan. What is more, the languages that have already been used for delexicalized parsing transfer on Occitan, namely Catalan and Old French (Scrivner and Kübler, 2012), come in as 14th and last, respectively. Also, the pt_bosque model scores here as 5th, whereas it was only 10th in the baseline evaluation. It is also interesting to note that the best results here come from large corpora, the smallest in the top 5 being pt_bosque with 222K tokens. Finally, the only model that did not suffer important performance loss is fr_partut+gsd+sequoia: it lost 2.9 LAS points and 1.9 UAS points, whereas the other four lost 7-10 LAS points and 6-8 UAS points. This may indicate that the diversity of linguistic content that was a disadvantage in the baseline evaluation actually provided robustness to the model which allowed it to maintain its performance when transferred to Occitan. This however has to be further investigated.

For the following step, we selected the best performing model for each of the languages in the top 5 (it_isdt, fr_partut+gsd+sequoia, pt_bosque) and used them to pre-annotate new Occitan samples. It is important to note that the difference in scores between it_isdt and it_idst+partut is explained by different annotation of 3 tokens when it comes to LAS, and 1 token when it comes to UAS, whereas the difference between it_isdt+partut and e.g. pt_bosque is much more important. However, we preferred having models based on different languages and comparing their performances rather than adhering strictly to the quantitative results.

5 Annotating Occitan: parsing process and manual correction analysis

The models selected in the previous section were applied to new samples of Occitan text. Coming from an existing corpus, these samples already had a manual POS annotation needed to put the delexicalized models to work. The resulting annotation was then submitted for validation to an experienced annotator. The corrected analysis was used as a gold standard against which the initial automatic annotation was evaluated. The manual annotation process also allowed us to observe the specificities of the annotation produced by the models and their impact on the manual annotation process.

5.1 Parsing new Occitan samples with selected UD models

Each of the 3 selected models was used to parse a new, syntactically unannotated sample of some 300 tokens of Occitan text. In order to minimize the bias related to the intrinsic difficulty of the text, we selected samples from the same source[5]. The annotation produced by the models was filtered: since it can be very time-consuming to correct erroneous dependencies, we only retained the dependencies for which the parser's decision probability score was >0.7. This was possible thanks to a Talismane option allowing to output the probability score for each parsing decision. Several other thresholds were tested (0.5, 0.6, 0.8, 0.9), and 0.7 was chosen for a balanced ratio between the confidence level and the sample coverage. Although research on parser confidence estimation has shown that more complex means may be needed to obtain reliable confidence estimates (Mejer and Crammer, 2012, e.g.), Talismane probability scores have already been used in this fashion and have been judged as adequate by human annotators (Miletic, 2018).

Table 4 shows the size of each sample, the model used to parse it and the coverage of the sample by the model when the 0.7 probability filter is applied. This partial annotation was then imported into the brat annotation tool and validated by an experienced annotator. Using this manually validated annotation as the gold standard, we calculated the percentage of correct annotations in the initial partial annotation submitted to the annotator (cf. Table 4, columns LAS and UAS). Punctuation annotation was excluded, since punctuation marks are always attached to the root of the sentence. We also give the duration of annotation for each sample.

Sample	Model	Size (tokens)	Coverage at prob. level >0.7	LAS	UAS	Duration of man. annot.
viaule_1	it_isdt	352	84.7 %	81.2	88.7	30'
viaule_2	fr_partut+gsd+sequoia	325	86.5 %	74.8	85.2	32'
viaule_3	pt_bosque	337	88.3 %	84.5	89.4	21'

Table 4: Results of the manual annotation of new Occitan samples

The elevated LAS and UAS scores show that the annotator's job consisted mostly in completing the partial annotation, whereas the actual corrections were less frequent, which is in line with our annotator's observations. The ergonomic value of such input is corroborated by the annotation duration times, which point towards a mean annotation speed of around 650 tokens/h. Since this annotator's speed during the annotation of the initial evaluation sample in Occitan was around 340 tokens/h, the utility of pre-processing with transferred models is certain. In order to verify if there were any noticeable differences between the outputs of different models, we proceeded to a more detailed analysis of the validation process.

5.2 Manual annotation analysis

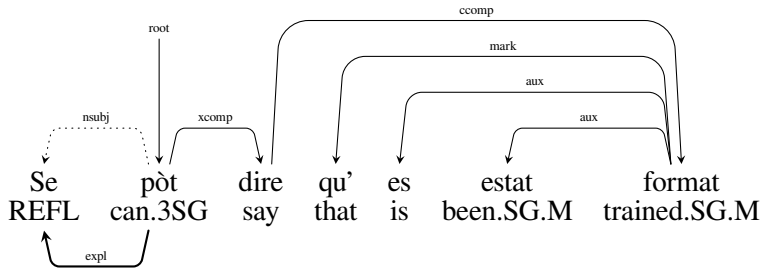
Given the differences between the languages on which the three models were trained, we could expect some differences in their output. However, the three models performed in a largely consistent way: the annotator observed that in the three samples the internal structure of the NP was mostly well processed, whereas verbal dependents seemed to be more challenging.

An issue related to lexical information occurred with reflexive pronouns: according to the UD guidelines, these should be treated as expletives with the *expl* syntactic label. However, given the minimal POS

[5] Sèrgi Viaule: *Escorregudas en Albigés*. Lo Clusèl, 2012.

annotation in the Occitan corpus and the fact that the models had no access to lexical information, it was impossible to distinguish these pronouns from any others. They were therefore often annotated as nominal subjects, direct objects and indirect objects, which are common functions for other types of pronouns (cf. example 2)[6].

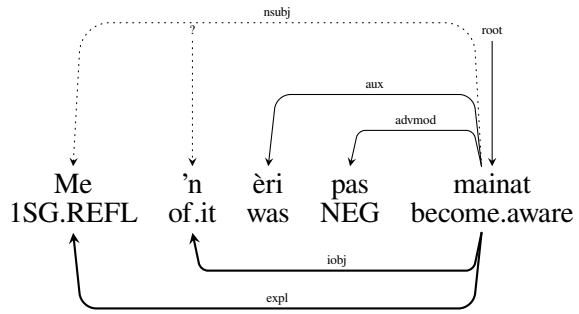
(2)



'You could say that he has been trained.'

In general, the annotation of pronouns proved difficult for the three models. Pronouns in sentence-initial position were often annotated as nominal subjects (*nsubj*), and in the case of pronominal clusters, pronouns other than the first often had no annotation, indicating that the one produced by the parser was not sufficiently reliable to pass our filtering criteria (cf. example 3). This is not surprising for the model trained on French, which has an obligatory subject, but it is for the ones learned on Italian and Portuguese, which allow the dropping of the subject.

(3)



'I hadn't noticed it.'

Although this type of error was recurrent, it was relatively easy to detect and correct.

Another less frequent but interesting issue retained the attention of the annotator: the auxiliaries. The Occitan verb *èsser* 'to be' can behave both as a copula and as an auxiliary in complex verbal forms, and whereas both of these usages receive the tag *AUX* on the POS level, their treatment on the syntactic level differs. The auxiliaries receive the label *aux* and are governed by the main verb of the complex form. The copulas are typically treated as *cop* and governed by their complement, except for the cases where they introduce a clause (cf. *The problem is that this has never been tried*), in which case they are treated as the head of the structure and carry the label most appropriate to the context in which they appear[7].

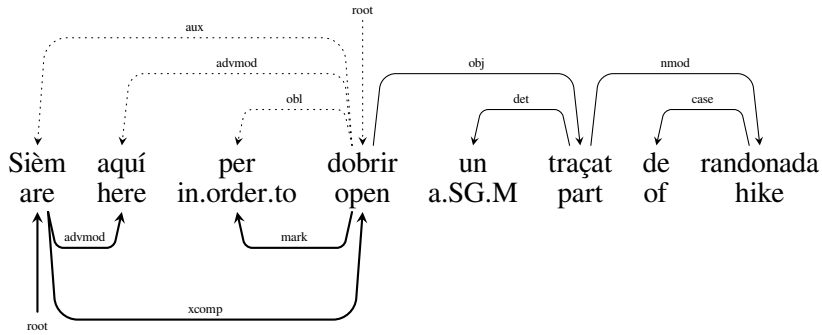
The annotator noticed that the forms of *èsser* tended to be treated as auxiliaries even when they were in reality a copula, especially if there was a main verb in their proximity (cf. example 4).

Correcting these structures was particularly time consuming because the annotator not only had to correct the annotation of the verb *èsser*, but also to remove and then redo several other dependencies in its neighbourhood. This also applies to all cases of root miss-identification.

[6]In the following examples, the syntactic annotation produced by the model is given above the sentence, with the incorrect part marked by dotted arcs. The correct analysis of the incorrect arcs is given below the sentence, in boldface arcs. The dependencies missing from the original annotation are indicated as having no governor, with *?* as label.

[7]Cf. the UD guidelines: `https://universaldependencies.org/u/dep/all.html#al-u-dep/cop`.

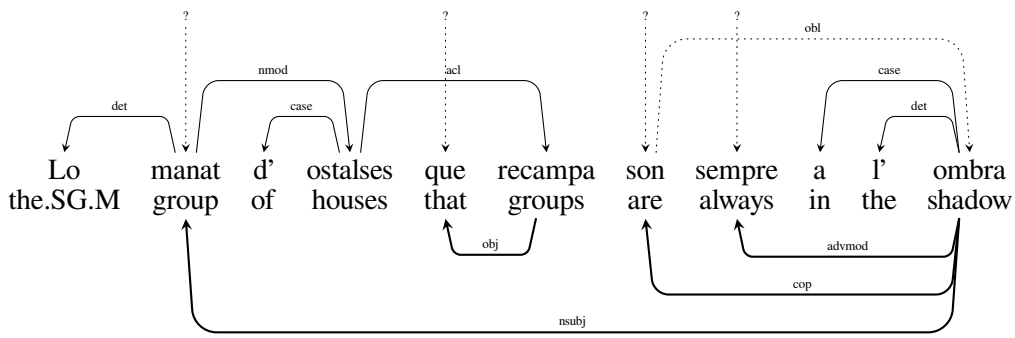
(4)



'We are here to open a part of a hike.'

More globally, all three models had difficulties with long-distance dependencies[8]. The models produced relatively few of them in each of the samples, and their accuracy rate was relatively low in two of the texts (cf. Table 5). The presence of a long-distance dependency often entailed the absence of other relations in the sentence, such as in example 5. However, it should be noted that this type of dependency is a long-standing issue in parsing and may not be due to model transfer.

(5)



'The hamlet of houses it regroups are always in the shade'

Sample	Model	Total long-distance deps.	Correct long-distance deps.
viaule_1	it_isd	18	12
viaule_2	fr_partut+gsd+sequoia	12	7
viaule_3	pt_bosque	13	10

Table 5: Long-distance dependency annotation per text sample

As mentioned above, some of these issues are undoubtedly related to the lack of lexical information in the models. Pronoun processing may be improved simply by including the pronoun type in the morphosyntactic traits of the corpus. This step is already planned for the next cycle of syntactic annotation. The issue with the distinction between the copulas and the auxiliaries is more complex, but even here, a presence of a morphosyntactic trait indicating the nature of the main verbs in the corpus (specifically, infinitive *vs* past participle) may contribute to the solution. This information will also be added to the corpus. Finally, the consistency of the output across the three models indicates that it could be useful to merge their training corpora and learn one global model, which is another direction we will be taking in the immediate future.

[8]For the scope of this paper, we define long-distance dependencies as having 6 or more intervening tokens between the governor and the dependent.

6 Conclusions and future work

In this paper we presented the application of cross-lingual dependency parsing on Occitan with the goal of accelerating the manual annotation of this language. 14 UD corpora of 8 Romance languages were used to train 21 different delexicalized parsing models. These models were evaluated on a manually annotated Occitan sample. The top 5 models achieved LAS scores ranging from 70.0 to 71.6, and UAS scores from 75.3 to 76.0. They were trained on Italian, Portuguese and French. From the top 5 models, 3 were selected (one per language) and used to annotate new Occitan samples. These were then submitted to an experienced annotator for manual validation. The annotation speed in these conditions went from 340 tokens/h to 650 tokens/h and the annotator also reported greater facility in facing the task. Their observations show that the three models had largely consistent outputs, but they also note several recurring issues, such as erroneous processing of copula structures and pronouns, and problems in the identification of long-distance dependencies.

Some of these problems can be tackled by simple strategies. In order to improve pronoun processing, the morphosyntactic trait encoding the pronoun type will be included in our corpora in the following annotation cycle. We will also test combining the now available annotated material in Occitan with the delexicalized models with the aim of enhancing the processing of copula constructions. Given the consistent output of the three models, we will also combine their training corpora and learn one last global model in the hope of achieving further output improvements.

Regardless of these issues, the positive impact of the application of cross-lingual delexicalized parsing on the manual annotation of Occitan is clear. The annotation speed achieved by the annotator shows that they were able to almost double the amount of annotated text in around half the time needed to process the initial evaluation sample. Using the delexicalized models to pre-process the data also had an important ergonomic and psychological effect: the annotator noted that it was less daunting to correct the output of the models than to face completely blank sentences.

Finally, it is important to point out that this was a reasonably quick process. Since the goal was to accelerate manual annotation, this work had to be less costly than manual annotation itself. This condition was met: thanks to the general quality of the UD corpora and their documentation, the work described in this paper was an efficient exercise with satisfying results.

References

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, et al. 2018. Corpora with part-of-speech annotations for three regional languages of france: Alsatian, Occitan and Picard. In *11th edition of the Language Resources and Evaluation Conference*.

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, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics.

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.

Avihai Mejer and Koby Crammer. 2012. Are you sure? confidence in prediction of dependency tree edges. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 573–576.

Aleksandra Miletic. 2018. *Un treebank pour le serbe: constitution et exploitations*. Ph.D. thesis, Université de Toulouse - Jean Jaurès.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Rudolf Rosa and David Mareček. 2018. Cuni x-ling: Parsing under-resourced languages in CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 187–196.

Olga Scrivner and Sandra Kübler. 2012. Building an Old Occitan corpus via cross-language transfer. In *KONVENS*, pages 392–400.

Jean Sibille. 2000. Ecrire l'occitan: essai de présentation et de synthèse. In *Les langues de France et leur codification. Ecrits divers–Ecrits ouverts,*. L'Harmattan.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 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*, pages 102–107. Association for Computational Linguistics.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864.

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.

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 and Assaf Urieli. 2015. Analyse morphosyntaxique de l'occitan languedocien: l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*.

Marianne Vergez-Couret. 2016. Description du lexique Loflòc. Technical report.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. CoNLL 2017 shared task: multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.