CorpusArièja: Building an Annotated Corpus with Variation in Occitan

Clamença Poujade, Myriam Bras, Assaf Urieli
CLLE, Université de Toulouse Jean Jaurès, CNRS, UT2J, France ; Joliciel, Foix, France
{clamenca.poujade, myriam.bras}@univ-tlse2.fr
assaf.urieli@gmail.com

Abstract

The Occitan language is a less resourced language and is classified as 'in danger' by the UNESCO. Thereby, it is important to build resources and tools that can help to safeguard and develop the digitisation of the language. CorpusArièja is a collection of 72 texts (just over 41,000 tokens) in the Occitan language of the French department of Ariège. The majority of the texts needed to be digitised and pass within an Optical Character Recognition. This corpus contains dialectal and spelling variation, but is limited to prose, without diachronic variation or genre variation. It is an annotated corpus with two levels of lemmatisation, POS tags and verbal inflection. One of the main aims of the corpus is to enable the conception of tools that can automatically annotate all Occitan texts, regardless of the dialect or spelling used. The Ariège territory is interesting because it includes the two variations that we focus on, dialectal and spelling. It has plenty of authors that write in their native language, their variety of Occitan.

Keywords: less-resourced language, occitan, POSTagging, diversity, corpus, deep learning

1. Introduction

Many languages, mostly minority and endangered ones, have no official standard for writing. This exacerbates their status as under-resourced languages because the surface variations are an important challenge in NLP.

The Occitan language deals with plenty of these variations: spelling, dialectal, formal, etc. Our aim is to provide resources and tools to help processing these variations in Occitan NLP.

In this article, we are going to describe the particularities of the Occitan language and some of its variations. Then, we will present our work to build and annotate a corpus of Occitan texts.

We build an annotated (lemma, supra-lemma, POS and verbal flexion) collection of texts that contains different types of variations present in the language. We selected texts from the French departement, Ariège. This departement and the texts provided are quite representative of the variations we focus on in this research.

2. Occitan is an Under Resourced Language

2.1. What is Occitan?

The Occitan is a Romance language spoken in the south of France, the Aran valley in Spain and some valleys in the Italian Alps. Traditionally, it is divided into six dialects (Bec, 1978) (Figure 1). The language has no official standard for spelling or speaking, as it has no official recognition in France nor Italy. Therefore, Occitan texts contain many variations in spelling and dialect.

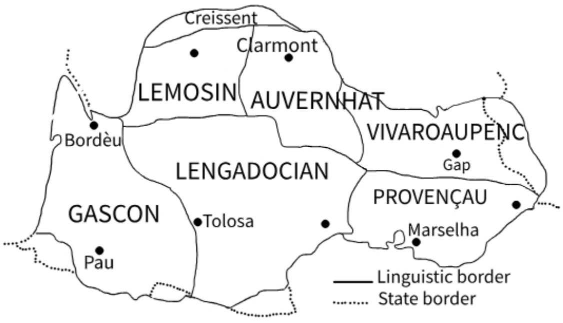

Figure 1: Map of Occitan Dialects.

Occitan has nearly a million speakers, the majority of which are over 60 years old and live in rural areas (OPLO, 2020). It is a language classified as "in danger" of disappearance by the UNESCO (Moseley, 2010). Thus, it is important to work on its safeguarding.

2.2. Resources and Tools for Occitan NLP

As discussed, Occitan is a minority language and many of these languages have fewer resources that can be used in natural language processing.

However, during the past ten years, some studies have been done to provide resources for the natural processing of the Occitan language. For processing of written text, three major funded projects had helped increase the digitization of Occitan.

SIGUL2024 Workshop, pages 66–71

BaTelÒc (Bras and Vergez-Couret, 2013)[1] was a project to build a digital collection of nearly 3.4 million words of Occitan texts. From this collection, and other texts, Bernhard et al. (2019) built an annotated corpus (12,425 tokens), with lemmas and part-of-speech (POS) tags, and provided a first tool for the automatic annotation of an Occitan corpus with POS tags (Urieli, 2013). This tool was then used in the project TolosaTreebank (Miletic et al., 2020a) to annotate a collection of texts (25,000 tokens) with both POS tags and syntactic dependencies.

Moreover, the independent institution *Lo Congrès permanent de la lenga occitana*[2] is working on NLP tools for public applications, such as automatic translation[3] or speech synthesis[4].

2.3. A Low Resourced Language?

Thanks to the European Language Grid (ELG) (Rehm et al., 2020) we can compare the amount of resources and tools between European languages.

Nowadays, Occitan has more resources for NLP than many other endangered and minority languages, like Aragonese, Gallo or Friulian. On the other hand, we cannot say that it is a well resourced language, as there is a lot of work yet to be done. For example, the automatic annotation tools can be improved, it could be interesting to fine-tune or train an Large Language Model (LLM) for occitan tasks and have more tools for speech processing, among other aims. Nevertheless, we do not consider Occitan as a low resourced language. If we compare Occitan with other European languages in the ELG, we can observe similarities in term of number and quality of NLP resources and tools with Breton, Asturian, Aragonese and Basque for video processing tasks. Basque is considered as a less-resourced language (Urbizu et al., 2022), Breton as an under-resourced language (Guennec et al., 2022) and Asturian and Aragonese as low-resourced (Lignos et al., 2022). Many others European languages are low-resourced and have less resources than Occitan. We thus choose to classify Occitan as an under-resourced language more likely to be less-resourced than low-resourced.

3. The Need for a Corpus with Variation

Occitan is a language with many variations. We chose to focus on two of these variations in our work on Occitan texts: dialectal and spelling. These surface variations add an additional challenge to the NLP of under-resourced language, and it is important to study their effects on various NLP tasks.

3.1. The Different Variations

The first variation we chose to study is the dialectal variation. This variation can be observed on a lexical, morphological, phonetic level and sometimes on a syntactic level. As previously stated, Occitan has about six dialects (Bec, 1978). These six dialects are a linguistic continuum, meaning there are plenty of isoglosses that traverse the Occitan territory, constituting different varieties in the dialects.

For example, the sentence *Lo gos vegèt un caval.* ('The dog saw a horse.') is a variety of Lengadocian. In Provençau it could be *Lo chin veguèt le cavau.* and in Gascon *Eth can vedó eth chivau.*.

The second variation concerns spelling variation. Contemporary Occitan has commonly three different spelling conventions. The most widely used is "classical" spelling, inspired by medieval Occitan and Catalan spelling [5]. Another spelling widely used is "Mistral" spelling. It uses mostly French spelling to write Occitan. The third group is personal spelling conventions. Indeed, the majority of Occitan speakers are not in contact with people or institutions that can teach them how to write the language. Nevertheless, many want to write in their language, so they choose to write with the spelling learned in school, French, Spanish, Catalan or Italian spelling.

For example, the sentence *L'occitan es una lenga romanica.* ('Occitan is a romance language.')[6] is written with "classical" spelling. *L'occita es uno lengo roumanico.* is an example of "Mistral" spelling and *L'oxità és uno léngo roumaniko.* is an example of what could be a personal spelling.

These two forms of variation limit the use of texts if we do not have tools that are trained to take them into account. The collections TolosaTreebank (Miletic et al., 2020b) and Restaure (Bernhard et al., 2018) introduced some dialectal variation, and the first tool (Vergez-Couret and Urieli, 2015) has good results on this variation. However, there is no spelling variation in these collections. In order to automatically handle all types of Occitan texts we need to build a robust tool that can deal with spelling variation.

3.2. The Challenge of Variation

As mentioned before, we chose two types of variations to work on with the texts in our corpus. More-

[1] http://redac.univ-tlse2.fr/bateloc/

[2] 'the permanent congress of the Occitan language'

[3] https://revirada.eu/

[4] https://votz.eu/

[5] in this article, Occitan extracts will be written with "classical" spelling.

[6] It can be pronounced [lutsit'a ez yno l'engo ruman'iko] in Lengadocian.

over, these variations are present in the majority of Occitan texts. A lot of texts are written in spelling conventions other than the "classical" one. Therefore, it is a necessity to build collections and tools that are able to process these variations.

Furthermore, the personal spelling and the "Mistral" spelling are often very similar to the pronunciation of the writer. Thereby, it appears important to study texts written with these spellings to study some particular Occitan varieties.

To the best of our knowledge, no annotated corpus of contemporary Occitan texts contains spelling variation.

3.3. Aim of the Corpus

We decided to build a collection of texts with these two kinds of variation, dialectal and spelling. The objective is to annotate it with POS tags and verbal flexion. The corpus is divided into a part that is manually annotated and a part that will be automatically annotated. We used the manually annotated part to train tools that will automatically annotate texts with spelling and dialectal variation.

When the annotations are completed, the corpus will be accessible to the scientific community through an OpenScience platform.

4. Description of the CorpusArièja

The CorpusArièja is the collection of 72 texts of the French department of Ariège, for a total of 41,233 tokens. We selected 56 authors who are natives of Ariège and who write in their own variety of Occitan.

To limit the type of variations, we restrained the collection to contemporary texts (1850 to nowadays) and to prose (tales, legends, novels and journalistic texts). We feel that texts previous from 1850 would introduce too much diachrony whereas we wish to concentrate on synchronic variation. The choice of 1850 is purely subjective. Other genres of texts, such as poetry, are more likely to have some syntactic forms that differ from the natural speaking of the language.

Setting aside diachronic and genre variations, the corpus contains both types of variation that interest our research, dialectal and spelling.

The majority of these texts were not available in a digital form. We needed to scan them and perform Optical Character Recognition (OCR) to prepare them for downstream processing. The OCR tool we used[7] was quite good and fast. However, we corrected every text manually to eliminate errors.

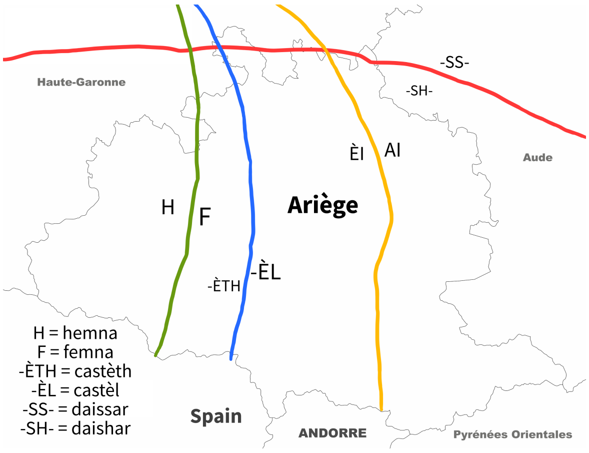

Figure 2: Map of some isoglosses in Ariège

4.1. The Choice of Ariège

Ariège is a border territory in the south of France. It has a frontier with Spain, but linguistically it borders Catalonia. This territory is also crossed by isoglosses that separate Lengadocian and Gascon dialects (Figure 2[8]).That makes Ariège an area of transition between two dialects and a land that has several linguistic variations. The proximity with Catalan creates, in the mountains, some language varieties that lay outside the continuum between Gascon and Lengadocian. Similarly, certain varieties in the high Pyrénées are very conservative in terms of their phonology and do not fit in the dialectal continuum between Lengadocian and Gascon.

Ariège is also an area that contains a lot of texts and especially texts written with different spellings. Indeed, there is an association, the "felibrige", that defends "Mistral" spelling and was very present in Ariège[9] with many publications with that spelling. Moreover, with the resurgence of occitanism in the early 1900s, many authors adopted the "classical" spelling which had just been created. In addition, many authors were not aware of or refused to adopt these spelling conventions, using a personal spelling convention instead. We believe that there isn't a significant difference between personal and "Mistral" spelling in the CorpusArièja collection, so we categorize them together under the "Mistral" label.

We feel it is important for tools and experiments to have a balanced distribution between the various types of variation (Table 1). As can be seen from the Linguatec project (Miletic et al., 2020a), we seem to have enough tokens of each dialect in our corpus in order to train a tool. However, it was complicated to maintain balanced numbers of tokens for the dialectal variation because the corpus

[7] https://www.ocr2edit.com/fr/convertir-en-txt with language parameters of Occitan, Catalan and French.

[8] Made from https://d-maps.com/carte.php?num_car=111145&lang=fr

[9] Particularly the institution 'Escòlo deras pirenéos'.

Dialect	# tokens
Lengadocian	20,194
Gascon	12,901
Other varieties	8,138

Spelling	# tokens
Mistral	19,887
Classical	21,346

Table 1: Distribution of variation in CorpusArièja

POS	meaning	count
ADJ	adjective	1,226
ADP	adposition	5,180
ADP+DET	adp.+determiner	762
ADV	adverb	2,000
AUX	auxiliary	865
CCONJ	coord. conjunction	1,397
DET	determiner	7,686
INTJ	interjection	143
NOUN	common noun	9,307
NUM	numeral	330
PART	particule	236
PRON	pronoun	4328
PROPN	proper noun	219
SCONJ	subord. conjunction	981
VERB	verb	7,683
X	foreign word	71

Table 2: Category distribution in CorpusArièja

contains a lot more Lengadocian texts than Gascon or other varieties.

4.2. Description of Annotations

We built a corpus with annotations of the POS tags, lemmas and verbal inflexion. For the annotation of the collection we adhere to the Universal Dependencies guidelines (Nivre et al., 2016).

We divided the corpus into two parts. One is annotated manually (21,691 tokens) to train and evaluate our tool and the other one (19,542 tokens) will be annotated automatically using the model of the automatic tool with the better results on spelling and dialectal variation.

The manual annotation of the corpus was performed by a single annotator. Indeed, it was a work that required a great expertise on the varieties of Ariège and of the spelling conventions used. Thereby, the annotator is a linguistic expert in these varieties.

4.2.1. Part-Of-Speech Annotation

For the annotation of POS we followed the guidelines used in Miletic et al. (2020a). These guidelines were made for the particularities of Occitan.

Table 2 is the description of the distribution of POS tags in the corpus.

4.2.2. Verbal Inflection Annotation

The annotation of verbal inflection is divided into six features, following the UD guidelines.

1. 'Gender', feminine or masculine, to describe the gender inflection for the past and present participles.
2. 'Number', singular or plural, is required for all verbal inflections except the infinitive form.
3. 'Person', 1, 2 or 3, is necessary to describe the person of conjugation for verbs that are not infinitive nor participles.
4. 'VerbForm', participle or infinitive, to tell the inflection form of the verb. If it is not present it means that is neither participle nor infinitive.
5. 'Mood', indicative, subjunctive, conditional or imperative, describes the mood inflection.
6. 'Tense', present, past, future or imperfect, is used to indicate the tense used in the conjugation of the verbal form.

```
     a | Number=Sing|Person=3|Mood=Ind|Tense=Pres
sautat | Gender=Masc|Number=Sing|VerbForm=Part|Tense=Past
```

Figure 3: Example of verbal inflection annotation

The Figure 3 is an example of an annotation of verbal inflection in the CorpusArièja.

4.2.3. The Lemma Annotation

The lemma is the form of citation of a word form. For example, *ostals* ('houses') is an inflected form of the lemma *ostal* ('house'). As already mentioned, the corpus contains variations of spelling and dialect, which makes the lemmatisation of the tokens in Occitan quite delicate. We have to make sure that we are not normalizing the language variety or spelling of the author. One of the interests of lemmas is to gather all of the inflections of a word together. However, we can go a little further saying that it could be interesting to unite all the variations of a word with a single lemma.

We therefore decided to create a second level of lemmatisation called *Supralemma*.

The first level of lemmatisation follows the spelling and language variety of the author, *oustals* is lemmatised *oustal* (spelling variation), the lemma of *ostaus* is *ostau* (dialectal variation) and *oustaou* is the lemma of *oustaous* (spelling and dialectal variation).

The second level, the *Supralemma* is an abstract lemma, it is not a normalisation or a standardisation, it is only a way to bring together all variations of a same word. *Oustals, oustaus* and *oustaous* have the same *Supralemma*, *ostal*. We chose to follow classical spelling and most of the Lengadocian dialect for the *Supralemma*. This choice is for the personal comfort of the expert annotator who is accustomed to this dialect and spelling in Occitan.

5. Conclusions and Perspectives

We presented the CorpusArièja, a corpus of Occitan texts. It has 42,413 tokens and it is divided into three dialects (Gascon, Lengadocian and other varieties of Ariège) and two spellings (mistralian and classical). We annotated the resource with POS tags, verbal inflection and two levels of lemmatisation: one level giving the presumed lemma that the author would use, and another more abstract level called *Supralemma* to lemmatise all the variations of a word together.

The annotated corpus can be modified to add others annotations, like the syntactic dependencies. Work is underway to train NLP tools for automatic POS annotation with good results on texts with and without spelling and dialectal variation. With the bests results of our POS tagger and Flex tagger, we want to automatically annotate the BaTelÒc collection. Our aim is to help the study of Occitan language and the development of public NLP applications.

We want to pursue this work introducing other variations, such as diachrony or a variation in the genre of the texts. There are numerous poems and songs available in Occitan that could be presented as variations.

We also want to try our tools on other Occitan dialects and test our methodology on other less or low resourced languages that have no writing standard. Indeed, we are willing to demonstrate that it is not necessary to have corpora with millions of words to build high-performance automatic annotation tools.

6. Bibliographical References

Pierre Bec. 1978. *La langue occitane*, 4e eédition corrigeée edition. Que sais-je ? 1059. Presses universitaires de France, Paris.

Delphine Bernhard, Myriam Bras, Pascale Erhart, Anne-Laure Ligozat, and Marianne Vergez-Couret. 2019. Language Technologies for Regional Languages of France: The RESTAURE Project. In *International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide*, Collection of Research Papers of the 1st International Conference on Language Technologies for All, page 272-275, Paris, France. European Language Resources Association (ELRA).

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steible, 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 *11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Myriam Bras and Marianne Vergez-Couret. 2013. BaTelÒc : a Text Base for the Occitan Language. In *First International Conference on Endangered Languages in Europe*, Minde, Portugal. University of Hawai'i Press .

David Guennec, Hassan Hajipoor, Gwénolé Lecorvé, Pascal Lintanf, Damien Lolive, Antoine Perquin, and Gaëlle Vidal. 2022. Breizhcorpus: a large breton language speech corpus and its use for text-to-speech synthesis. In *Odyssey Workshop 2022*, pages 263–270. ISCA.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. *arXiv preprint arXiv:2202.12288*.

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

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020b. A four-dialect treebank for Occitan: Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Christopher Moseley. 2010. *Atlas des langues en danger dans le monde*. Unesco.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16), pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

OPLO. 2020. *L'occitan aujourd'hui, enquête sociolinguistique*. Ofici Public de la Lenga Occitana.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajič, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez-Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. 2020. European language grid: An overview. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France. European Language Resources Association.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. Basqueglue: A natural language understanding benchmark for basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Theses, 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*, Caen, France.