Loflòc: A Morphological Lexicon for Occitan using Universal Dependencies

Marianne Vergez-Couret[1] Myriam Bras[2] Aleksandra Miletić[3] Clamença Poujade[2]
[1]UR15076 FoReLLIS, Université de Poitiers, France
[2]UMR5263 CLLE, CNRS & Université de Toulouse Jean Jaurès, France
[3]Department of Digital Humanities, University of Helsinki, Finland
marianne.vergez.couret@univ-poitiers.fr
{myriam.bras, clamenca.poujade}@univ-tlse2.fr
aleksandra.miletic@helsinki.fi

Abstract

This paper presents Loflòc (Lexic obèrt flechit Occitan – Open Inflected Lexicon of Occitan), a morphological lexicon for Occitan. Even though the lexicon no longer occupies the same place in the NLP pipeline since the advent of large language models, it remains a crucial resource for low-resourced languages. Occitan is a Romance language spoken in the south of France and in parts of Italy and Spain. It is not recognized as an official language in France and no standard variety is shared across the area. To the best of our knowledge, Loflòc is the first publicly available lexicon for Occitan. It contains 650 thousand entries for 57 thousand lemmas. Each entry is accompanied by the corresponding Universal Dependencies Part-of-Speech tag. We show that the lexicon has solid coverage on the existing freely available corpora of Occitan in four major dialects. Coverage gaps on multi-dialect corpora are overwhelmingly driven by dialectal variation, which affects both open and closed classes. Based on this analysis we propose directions for future improvements.

Keywords: low resource languages, Occitan, morphological lexicon

1. Introduction

Loflòc (Lexic obèrt flechit Occitan) is a morphological lexicon for Occitan. Its initial version was produced during the past few years as part of the ANR project RESTAURE[1] (Bernhard et al., 2021) and it was subsequently expanded during the European POCTEFA project LINGUATEC[2], in collaboration with Lo Congrès Permanent de la Lenga Occitana[3] (Bras et al., 2020). It contains 650,000 entries for 57,000 lemmas. The entries contain an inflected form, the corresponding lemma and the part-of-speech tag based on the Universal Dependencies framework[4]. The resource is available under the Creative Commons BY-NC-SA 4.0 license through Zenodo[5].

The creation of this lexicon is part of a wider drive to provide linguistic resources for Occitan, which was a low-resource language at the outset of this endeavour. The overarching goals are, on the one hand, the preservation and dissemination of linguistic heritage, and on the other hand, the creation of resources suitable for the development of NLP tools, in particular for basic processing such as lemmatization and morphosyntactic and syntactic analysis (Vergez-Couret and Urieli, 2015; Bernhard et al., 2018; Miletić et al., 2020; Miletić, 2023).

With the advent of large language models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020), the morphological lexicon no longer occupies the same position in the NLP pipeline: contemporary algorithms relying on LLMs are capable of achieving state-of-the-art results without it. However, as pointed out by Wang et al. (2022), a crushing majority of the world's 7000 languages do not have vast amounts of written text which could be used as training data for LLMs, which negatively impacts the results achieved on these languages, including through cross-lingual transfer (Hu et al., 2020). As argued by Joshi et al. (2020), this paradigm shift towards unsupervised learning based on large amounts of text "make[s] the 'poor [languages] poorer'".

On the other hand, a far larger number of languages possess lexical resources (see, e.g., Kamholz et al., 2014), and Wang et al. (2022) propose a methodology to leverage such resources to improve language representation and results of LLMs. While our own work is not directly aimed at improving LLMs, it has proven useful both in the creation of initial labeled datasets for Occitan (Bernhard et al., 2018; Miletic et al., 2020) and more recently in annotating a large silver-standard corpus and thus jumpstarting work on lemmatization (Miletić and Siewert, 2023). Note also that cre-

[1]French Research Agency ANR-14-CE24-0003
[2]European Regional Development Fund EPT 227/16
[3]https://locongres.org/
[4]https://universaldependencies.org/u/overview/morphology.html
[5]https://doi.org/10.5281/zenodo.10838802

LREC-COLING 2024, pages 10716–10724

ating and using lexicons often requires less human expertise in NLP and less extensive computational resources than working with LLMs. Both of these factors can play an important role for low-resource communities. We share our work in the hope of helping other languages and communities in a similar situation.

2. Beginnings of NLP for Occitan

2.1. 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 many languages that do not have official status, it is not standardized as a whole. It has six recognised varieties or dialects (hereinafter named by their names in the dialect: Auvernhàs, Gascon, Lengadocian, Lemosin, Provençau and Vivaro-Aupenc), which also display internal variation (Bec, 1995). These varieties form a continuum in which the Lengadocian dialect occupies a central position. (Figure 1).

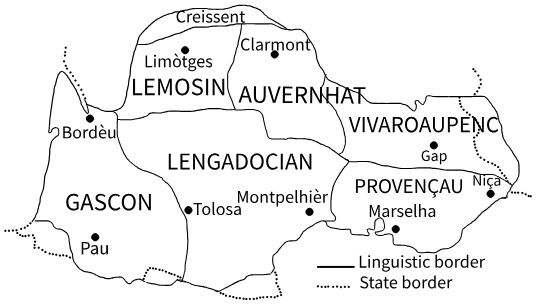

Figure 1: Occitan dialect continuum.

Furthermore, different spelling norms have been in use since the Middle Ages. Currently, two different spellings are prevalent. One of them, called the *classical spelling*, is based on the Occitan troubadours' medieval spelling, and the other, called the *Mistralian spelling*, is closer to the French orthographic conventions (Sibille, 2002).

Dialectal and orthographic variation are challenging from the point of view of NLP since this diversity manifests itself on the lexical and morphological levels and thus aggravates the data sparsity issue. Nevertheless, some recent efforts have provided an initial set of resources with the end goal of training essential NLP tools for Occitan.

2.2. Strategies for Building resources for Occitan

Resource building for Occitan started much later than it did for high-resource languages such as English or French. While this delay was regrettable, it had a silver lining: the work on Occitan benefited from the considerable advances made in NLP in the meantime, in particular from the advent of machine learning systems which replaced symbolic systems. This made it possible to reuse language-independent algorithms and to draw on available lexicographic resources in order to produce initial annotated data and thus bootstrap the creation of annotated corpora.

In these initial stages, the resources built were restricted to the Lengadocian dialect and the classical spelling, in order to limit variation (Vergez-Couret and Urieli, 2015; Bernhard et al., 2021). The choice of Lengadocian was guided by pragmatic reasons, since the most lexicographic resources and printed texts were available for this dialect, but it was also linguistically motivated. As mentioned in Section 2.1, Lengadocian occupies a central position in the continuum both geographically and linguistically, so it was expected that resources created for this dialect could be successfully transferred to others, as demonstrated in Miletic et al. (2020). To guarantee their reusability and durability, these resources were built in accordance with international NLP standards. We used the same strategy for the lexical resources: the first version of Loflòc, that will be presented in the next section, is a Lengadocian lexicon, written in classical spelling. Further steps of lexical resources development will aim at a multidialectal and multispelling lexicon.

3. Loflòc

3.1. Building Loflòc: Methodology

Loflòc was based on two major lexicographic resources: the Occitan parts of the bilingual Occitan-French and French-Occitan Lengadocian dictionaries by Laux (2001, 2005). The extracted information was then enriched with grammatical indications where these were incomplete, or modified where they were deemed unsuitable. For instance, some categories from traditional grammar were adapted to be in line with current linguistic theories. Most notably, indefinite and possessive adjectives (such as *cada* 'each' or *mon* 'my') were recoded as determiners.

Since the lexicon was extracted from a traditional dictionary, the extracted entries corresponded only to base forms of words. In the next step, the lexicon was completed by adding inflectional paradigms: plural for nouns, feminine and plural forms for adjectives, inflected forms for verbs. Verbal inflected forms were provided by the Verb'Òc application, a verb conjugator developed by Lo Congrès Permanent de la Lenga Occitana [6].

[6] https://dicodoc.eu/oc/conjugasons

Initially, the morphosyntactic information was represented using a set of labels adapted from the GRACE standard (Rajman et al., 1997). This choice was motivated by the fact that other related languages (French and Catalan) had used this particular tagset (Bras et al., 2020). More recently, the original tags were converted into the Universal Dependencies tagset[7]. This was done with the goal of including Occitan in the UD community, as we have already done for annotated corpora (Miletic et al., 2019).

Currently, lexicon entries are triples containing an inflected form, the corresponding UD PoS tag, and the lemma. Fine-grained morphosyntactic features will be added in future versions. The content of the lexicon is presented in detail below.

3.2. Content

Loflòc contains 680k entries corresponding to 650k unique wordforms for 57k lemmas (see Table 1) of the Lengadocian dialect. The distribution of entries per category is given in Table 2. The categories follow the definitions in the Universal Dependencies guidelines[8], with two exceptions: `ADP+DET` and `X`. In Occitan, some prepositions fuse with masculine and plural forms of the definite article, so, for instance, instead of *a lo* 'to/at the-sg', the single form *al* is used; instead of *per los* 'for/to the-pl', the single form *pels* is used. This phenomenon also exists in other Romance languages. In order to allow for the correct identification of these forms in unprocessed text, we include them in the lexicon and give them the combined tag `ADP+DET`. We use `X` to mark epenthetic consonants which do not fit any other category, such as 'n' in *a'n aquò 'to that'*.

Loflòc	
Entries	680,205
Unique wordforms	650,577
Lemmas	57,200

Table 1: Lexicon size

Only around 4% of inflected forms in Loflòc are ambiguous in the sense of having more than one entry, and <0.5% have more than two entries (Table 3). One of the examples of highly ambiguous inflected forms is *seguda*, which corresponds to six lexicon entries. It can be a common noun ('continuation') or the feminine form of the adjective *segut* ('seated'). It can also be the past participle form of two pairs of verbs : *segudar*/*sègre* ('to follow'), and *sèire*/*sèser* ('to sit'). Note that the pairs of infinitives *segudar*/*sègre* and *sèire*/*sèser* correspond to verbs which have different lemmas but partially overlapping paradigms. This is due to intradialectal variation (within Lengadocian).

POS	meaning	count
ADJ	adjective	42,657
ADP	adposition	111
ADP+DET	adp.+determiner	22
ADV	adverb	1,170
AUX	auxiliary verb	184
CCONJ	coord. conjunction	9
DET	determiner	125
INTJ	interjection	189
NOUN	common noun	66,095
NUM	numeral	55
PRON	pronoun	294
PROPN	proper noun	1,755
SCONJ	subord. conjunction	24
VERB	verb	567,512
X	epenthetic consonants	3

Table 2: Category distribution in Loflòc

# entries	# wordforms	(%)
1	623,951	91.73
2	24,017	3.53
3	2,304	0.34
4	261	0.04
5	20	0.00
6	8	0.00
7	14	0.00
8	1	0.00
10	1	0.00

Table 3: Ambiguity in Loflòc

Form	Lemma	PoS
seguda	segut	ADJ
seguda	seguda	NOUN
seguda	segudar	VERB
seguda	sèire	VERB
seguda	sèser	VERB
seguda	sègre	VERB

Table 4: Loflòc entries for *seguda*

4. General Coverage Analysis

To illustrate the utility of Loflòc, we examine its coverage of the available annotated corpora in Occitan (Table 5). For this, we use two manually annotated resources: the RESTAURE corpus (Bernhard et al., 2018) and the Tolosa Treebank produced during the LINGUATEC project (Miletic et al., 2020), and one automatically annotated corpus: OcWikiAnnot (Miletić, 2023). Since RESTAURE and Tolosa

[7]The conversion table is available in Appendix A.

[8]https://universaldependencies.org/u/pos/all.html

Corpus	Dialect	# Tokens	Coverage (%)	# Types	Coverage (%)
Restaure	Gascon	3,311	67.02	1,367	51.06
	Lengadocian	3,608	91.57	1,424	83.92
	Provençau	1,085	85.90	544	77.94
	Lemosin	1,975	76.20	941	60.47
	All	9,979	79.77	3,636	62.95
Tolosa Treebank	Gascon	3,465	69.38	1,351	51.89
	Lengadocian	16,192	91.06	4,314	79.37
	Provençau	1,113	87.51	539	79.04
	Lemosin	1,147	74.54	559	60.47
	All	21,917	86.59	5,941	69.77
OcWikiAnnot	All	1,812,127	82.25	143,160	30.56

Table 5: Coverage of existing annotated corpora

Treebank are stratified by dialect, we use them to examine the coverage of different varieties, and rely on OcWikiAnnot to evaluate the coverage on a larger resource, in which frequency effects are more readily observed.

In this step, we approximate the scenario in which the lexicon is used to provide basic information for raw text. We therefore only check if each corpus token has an entry in Loflòc. We calculate coverage both on token level (percentage of tokens from the corpus found in the lexicon) and on type level (percentage of unique wordforms from the corpus found in the lexicon). While type-level information can be an indicator of a lexicon's general coverage, token-level information can show if the lexicon contains frequent forms. The count excludes punctuation marks and lowercases forms which are not annotated as proper nouns.

As can be seen in Table 5, Loflòc covers 79.77% of tokens and 62.95% of types in the Restaure corpus, 86.59% of tokens and 69.77% of types in the Tolosa Treebank corpus, and 82.25% of tokens and 30.56% of types in OcWikiAnnot. The drastic drop in the coverage of types, but a high coverage of tokens in OcWikiAnnot indicates that Loflòc indeed tends to contain high frequency wordforms.

When it comes to different dialects, for both RESTAURE and Tolosa Treebank, the dialect with the highest coverage is Lengadocian, with up to 91.57% on token level. This is, of course, expected, since the lexicon is based on this variety of Occitan. The lower coverage rate of the other dialects can be expected to be the result of phonetic, morphological and lexical variation.

Among the other three dialects, Gascon has the lowest coverage at only 67.02%, which is consistent with the fact that it is the most distinct with respect to others. In particular, it exhibits phonetic processes that do not exist in other dialects, leading to a higher amount of spelling differences.

The effect of dialectal variation is also visible in the coverage of OcWikiAnnot. An examination of the most frequent types from this corpus not covered by the lexicon show that an important portion of them represent forms of function words specific to Gascon or to Provençau, such as *eth* (Gasc. 'him'), *era* (Gasc. 'her'), *dab* (Gasc. 'with'), *ambé* (Prov. 'with'), *aqueu* (Prov. 'this (one)'), *aquelei* (Prov. 'these (ones)').

5. Cross-Dialectal Coverage: Areas of Improvement

To better examine the coverage of Loflòc on dialects other than Lengadocian, we rely on the two manually annotated corpora which are also stratified by dialect: RESTAURE and Tolosa Treebank. In this scenario, we use the manually provided annotations to precisely identify full entries that are absent from the lexicon (as opposed to *inflected forms* absent from the lexicon, accounted for in the previous section). We pay special attention to coverage gaps due to dialectal variation.

In this step, we check how many tokens from annotated corpora are present in the lexicon *with the right part-of-speech tag* (cf. Table 6). The tokens that are present in the lexicon, but not associated with the same POS tag as in the corpus are likely to be ambiguous wordforms for which not all relevant entries have been created at this point.

Across both RESTAURE and Tolosa Treebank, the PoS tags associated with the most frequent absent forms are `PART`, `ADP`, `PROPN`, `DET`, `NOUN` and `VERB`. Nearly all of the non covered forms from both corpora represent forms that do not belong to Lengadocian. Overwhelmingly, these forms belong to a single dialect, as opposed to being shared between multiple non-Lengadocian dialects.

Here too, Gascon exhibits the highest percentage of tokens not correctly covered by the lexicon. In both corpora, the major driver of this difference is the form *que*. This form is a pronoun and a subordinating conjunction in Lengadocian, and is present with these two parts of speech in Loflòc. However, in Gascon it can also be an enunciative parti-

Corpus	Dialect	# Tokens	rightPOS (%)	wrongPOS (%)
Restaure	Gascon	3,311	56.45	10.57
	Lengadocian	3,608	87.83	3.74
	Provençau	1,085	81.75	4.15
	Lemosin	1,975	72.51	3.70
	All	9,979	73.72	6.04
Tolosa Treebank	Gascon	3,465	60.49	8.89
	Lengadocian	16,192	89.22	1.84
	Provençau	1,113	83.92	3.59
	Lemosin	1,147	72.10	2.44
	All	21,917	83.52	3.08

Table 6: Token-level coverage of annotated corpora taking into account the POS annotation

cle, which should be tagged as PART. Enunciative particles exist only in Gascon. They immediately precede the main verb of the sentence and mark the status of the clause (different particles signal declarative, negative, interrogative and exclamative modalities). The occurrences of the declarative particle *que* in the Gascon subcorpora account for close to a third of the tokens that do not have the appropriate POS tag in the lexicon.

When it comes to the ADP category, the Gascon form *dab* 'with' is not in the lexicon, whereas its counterpart *amb* is. The absent DET forms are due to phonetic variation that distinguishes other dialects from Lengadocian. Most notably, the Provençau form *sei* 'their', which is not covered by the lexicon, is equivalent to the Lengadocian form *sos*. The NOUNs which are absent from Loflòc are due to lexical or phonetic variation in the other three dialects. For example, *chamin* 'road, way' in Lemosin corresponds to *camin* in Lengadocian. Absent verbal forms are mostly due to differing inflections across dialects: for example, *fuguet* ('to be', 3rd person singular preterit) in Lemosin corresponds to *fuguèt* in Lengadocian.

While the current version of Loflòc does provide a degree of coverage for other dialects, this analysis clearly shows that there is room for improvement. The lists of forms that are currently absent from the lexicon will be used to guide the extensions in future versions.

6. Conclusions and Future Work

In this paper, we presented Loflòc, the first morphological lexicon for Occitan, which contains 650 thousand entries for 57 thousand lemmas. Our analysis based on available annotated corpora shows that Loflòc has good coverage on Lengadocian, the dialect on which it is based, but also solid coverage of Provençau, Lemosin and, to a lesser extent, Gascon. The gaps in the coverage of these dialects are indeed driven by dialectal variation, both in closed and in open classes.

Our three priorities for the future of Loflòc are as follows: (1) adding detailed morphological information to the current version of the lexicon; (2) extending the coverage of other dialects by relying on existing lexicographic resources; and (3) including other spelling norms, starting with the Mistralian norm, which is widely used in Provençau.

Finally, in our experience, one of the core principles when working on low-resource languages is "Use what you have". Occitan is lucky enough to possess many dictionaries which were being digitized and encoded in XML as a first step to build an online lexicographic database [9] at the time we started working on lexical resources for NLP. This made building Loflòc relatively easy compared to other types of resources, such as annotated corpora. In terms of the classification of languages by Joshi et al. (2020) based on the amount of resources available to them, using Loflòc has allowed Occitan to pass from the category of "Scraping By"s into the category of "Hopeful"s with relative speed and ease. We are now hopeful that our experience might provide useful pointers to languages and communities in a similar situation.

[9] https://dicodoc.eu/fr/dictionnaires

7. Acknowledgements

We would like to thank Benaset Dazeas and Aure Séguier of Lo Congrès Permanent de la Lenga Occitana, who helped us in the first steps of Loflòc construction. This work has been carried out within the framework of several projects : the ANR-14-CE24-0003 RESTAURE project and the ANR-21-CE27-0004 DIVITAL project supported by the French National Research Agency; the EFA 227/16 LINGUATEC Project, financed by the POCTEFA Interreg European funds. The work of Aleksandra Miletić was funded by the Academy of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology".

8. Bibliographical References

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

Delphine Bernhard, Anne-Laure Ligozat, Myriam Bras, Fanny Martin, Marianne Vergez-Couret, Pascale Erhart, Jean Sibille, Amalia Todirascu, Philippe Boula de Mareüil, and Dominique Huck. 2021. Collecting and annotating corpora for three under-resourced languages of france: Methodological issues. *Language Documentation & Conservation*, 15:316–357.

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

Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benaset Dazéas. 2020. Loflòc : Lexic obert flechit occitan. In *Fidelitats e dissidéncias. Actes del XIIn Congrès de l'Associacion internacionala d'estudis occitans. Actes du XIIe Congrès de l'Association internationales d'études occitanes.*, Albi, France.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christian Laux. 2001. *Dictionnaire occitan-français: languedocien*. Section du Tarn de l'Institut d'estudis occitans.

Christian Laux. 2005. Dictionnaire français-occitan. *Puylaurens: IEO*.

Aleksandra Miletic, Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, and Marianne Vergez-Couret. 2019. Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan. In *26e conférence sur le Traitement Automatique des Langues Naturelles (TALN-2019) et 21e édition la conférence jeunes chercheur×euse×s RECITAL*, volume 2 of *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*, pages 427–435, Toulouse, France. ATALA.

Aleksandra Miletić, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. Building a Universal Dependencies treebank for Occitan. In *Proceedings of*

the Twelfth Language Resources and Evaluation Conference, pages 2932–2939.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. 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).

Aleksandra Miletić and Janine Siewert. 2023. Lemmatization experiments on two low-resourced languages: Low Saxon and Occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 163–173, Dubrovnik, Croatia. Association for Computational Linguistics.

Aleksandra Miletić. 2023. Outiller l'occitan: nouvelles ressources et lemmatisation. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 217–231. Association pour le Traitement Automatique des Langues.

Martin Rajman, Josette Lecomte, and Patrick Paroubek. 1997. Format de description lexicale pour le français. partie 2 : Description morpho-syntaxique. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.

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

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.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

A. Appendix A: Conversion Table from GRACE to UD tagset

GRACE tag	Meaning	Example	UD tag
Af	(qualificative adjective)	polit 'nice', triste 'sad', bon 'good', melhor 'better', màger 'biggest'	ADJ
Ao	(ordinal adjective)	primièr 'first', segond 'second', tresen 'third', darrièr 'last'	ADJ
Ak	(cardinal adjective)	dos 'two', tres 'three', quatre 'four'	NUM
Ai	(indefinite adjective)	autre 'other', quite 'very', tal 'so'	ADJ
As	(possessive adjective)	miu 'my', teuna 'your', seu 'his/her', sieuna 'his/her'	ADJ
Cc	(coordinating conj.)	e 'and', mas 'but'	CCONJ
Cs	(subordinating conj.)	quand 'when', coma 'like', que 'that', se 'if', perque 'because'	SCONJ
Da	(article determiner)	lo, la, los, las 'the', un, una 'a', de 'some'	DET
Dd	(demonstrative det.)	aquel, aqueste, aiceste, este 'dem'	DET
Di	(indefinite determiner)	cada 'each', qualque 'some', mantun 'some', mai d'un 'more than one', tot 'all', un pauc de 'a little'	DET
Ds	(possessive determiner)	mon 'my', ton 'your', son 'his/her', ma 'my.f', ta 'your.f', sa 'his/her.f', nòstre 'our', vòstra 'your.pl.f, lor 'their', lors 'their.pl'	DET
Dt	(interrogative/exclamative det.)	quin 'which', qual 'who', quun 'which', quane 'which'	DET
Dr	(relative determiner)	loqual, lasqualas 'which'	DET
Dk	(cardinal determiner)	un 'one', dos 'two', tres 'three'	NUM
Dp	(partitive determiner)	de 'of'	DET
Nc	(common noun)	ostal 'house', dròlla 'girl', cèl 'sky', flor 'flower'	NOUN
Np	(proper noun)	Maria, Aran, Tolosa	PROPN
Nk	(cardinal noun)	dos (dins « un parelh de dos ») 'two (in 'a pair of two')	NUM
Pp	(personal pronoun)	ieu 'I', tu 'you.sg', el 'he', nosautres 'we', eles 'they', ne 'partitive', òm 'one', l'òm 'one', o, ac, ba 'npro'	PRON
Pd	(demonstrative pronoun)	aquò, aquel, aqueste, çò	PRON
Pi	(indefinite pronoun)	pauc 'little', qualques unes 'some', mantuns 'many', cadun each one', quicòm 'something', degun 'nobody', totòm 'everyone', res 'nothing', cap 'not any'	PRON
Ps	(possessive pronoun)	miu 'mine', teu yours.sg', vòstra 'yours.pl'	PRON
Pt	(interrogative pronoun)	quant 'how many', quin which', qual 'who', dequé what', qué 'what', que 'what'	PRON
Pr	(relative pronoun)	ont 'where', dont 'of which', que 'that/which', lasqualas 'which', loqual 'which', lo qual 'which'	PRON
Px	(reflexive pronoun)	me 'myself', te 'yourself', se himself/herself', lor 'themselves'	PRON
Pk	(cardinal pronoun)	dos 'two', tres 'three', trenta-cinc 'thirty-five'	NUM

GRACE tag	Meaning	Example	UD tag
Rg	(general adverb)	pas ‘not’, aisidament ‘easily’, bravament ‘much’, aicí ‘here’, aquí ‘here’, defòra ‘outside’, çai-jós ‘there below’, ara ‘now’, uèi ‘today’, puèi ‘then’, non (when replaces the negative « pas »), melhor ‘better’, jamai ‘never’, òc ‘yes’	ADV
Rx	(interrogative/exclamative adv.)	quant ‘how much’, qué ‘what’, que, coma, cossí ‘how’	ADV
Rp	(particle adverb)	ne, non (when it goes with the negative « pas ») que, be, e, ja, si (Gascon enunciative particles)	PART
Rq	(intensive/quantitative adv.)	fòrça ‘much’, plan ‘much’, cap ‘none’, pus ‘no more’, brica ‘not any’, mai ‘more’, tot ‘all’, tròp ‘too much’, gaire ‘little’, aitant ‘as much’, un pauc ‘a little’, pauc ‘little’	ADV
Sp	(preposition)	per ‘for’, de ‘of’, coma ‘as’, dins ‘in’, abans ‘before’, dempuèi ‘since’, a ‘to’, sus ‘on’, jós ‘under’, en ‘in’	ADP
Spda	(preposition + article)	del, dels, al, als, pel, pels, sul, suls, jol, jols, vèl, vèls	ADP+DET
Sd	(deictic preposition)	vaquí	ADP
Vm	(main verb)	dansar ‘dance’, manjar ‘eat’, poder ‘can’, èsser ‘be’, aver ‘have’	VERB
Va	(auxiliary verb)	èsser ‘be’, aver ‘have’	AUX
I	(interjection)	zo, i, a, o, òu, flica-flaca, pam	INTJ
X	(epenthetical consonant)	n-, -n-, -z- , -s, s-, -s-	X
F	(punctuation)	, ; : . ! ? ...	PUNCT

Table 7: Conversion table from the GRACE tagset to the UD tagset used for Lofl̀oc