OcWikiAnnot: Annotated Wikipedia Corpus in Occitan

OcWikiAnnot is a corpus of Wikipedia content in Occitan that is tokenized, PoS-tagged and lemmatized. The corpus contains 100 000 sentences for a total of 2 037 723 tokens.

1. Original Corpus

OcWikiAnnot is based on the downloadable Wikipedia corpus in Occitan that is part of the Leipzig Corpora Collection (Goldhahn *et al.* 2012). In the original corpus, the content is split into sentences, which are ordered alphabetically and associated with a sentence ID. The largest freely downloadable version of the corpus (under the CC BY license) contains 100 000 sentences.

2. Processing

In the OcWikiAnnot corpus, the sentences were reshuffled, tokenized, part-of-speech tagged and lemmatized. The annotated corpus follows the Universal Dependencies CoNLL-U format. The sentence ID from the original corpus is preserved in the *# sent_id* metadata line that heads each sentence.

2.1 Tokenization

Tokenization was done using a rule-based tokenizer. The tokenizer is based on the tokenization specifications for the Occitan tokenizer available here. Unlike the original tokenizer, the one used for OcWikiAnnot does not keep multiword expressions as a single token, but separates them into individual tokens.

2.2 Annotation

The part-of-speech tagging and lemmatization were done automatically. A MaChAmp (van der Goot *et al*. 2021) model was trained on the Tolosa Treebank (Miletic *et al.* 2020) and ensembled with a lexicon (Bras *et al.* 2020) at prediction time. Part-of-speech tagging is done using the Universal Dependencies tagset. More information on the annotation process is available in Miletić and Siewert (2023).

NB: the tokenization step in the current version **does not** split the fused PREP+DET forms (e.g. *dels → de+ los)* as expected by the UD guidelines on tokenization. This will be dealt with in the future versions of the corpus.

3. Corpus Statistics

Some basic corpus statistics are available below.

Tokens: 2 037 723
Types: 147 068
Lemmas: 111 656

Number of occurrences of different PoS tags

Tag	Occurrences	Tag	Occurrences
ADJ	135 760	NUM	45 116
ADP	306 798	PART	4 610
ADV	76 063	PRON	65 916
AUX	55 023	PROPN	92 554
CCONJ	56 173	PUNCT	225 596
DET	317 677	SCONJ	20 831
INTJ	3 589	VERB	191 342
NOUN	434 306	X	1 654

4. Contact

The corpus was annotated by Aleksandra Miletić (Department of Digital Humanities, University of Helsinki).
Contact: firstname.lastname@helsinki.fi

5. Acknowledgements

This work was supported by the Academy of Finland through project No. 342859 CorCoDial - Corpus-based computational dialectology.

6. License

7. Citing

If you reuse the corpus, please cite the following publication:

Aleksandra Miletić and Janine Siewert. 2023. Lemmatization Experiments on Two Low-Resourced Languages: Occitan and Low Saxon. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (to appear).* Association for Computational Linguistics.

8. References

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

Dirk Goldhahn, Thomas Eckart and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

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).