OcWikiDisc : a Corpus of Wikipedia Talk Pages in Occitan

Miletic Haddad, Aleksandra

COLING
2022-10

Miletic Haddad , A & Scherrer , Y 2022 , OcWikiDisc : a Corpus of Wikipedia Talk Pages in þÿ O c c i t a n . i n Y S c h e r r e r , T J a u h i a i n e n , N L j u b e a i , P N a k o v , J T i e d e (eds) , Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects : The 29th International Conference on Computational Linguistics . International conference on computational linguistics , no. 19 , vol. 29 , COLING , Gyeongju , pp. 70-79 , Workshop on NLP for Similar Languages, Varieties and Dialects , Gyeongju , Korea, Republic of , 16/10/2022 .

http://hdl.handle.net/10138/350301

OcWikiDisc: a Corpus of Wikipedia Talk Pages in Occitan

Aleksandra Miletić
Department of Digital Humanities
University of Helsinki
aleksandra.miletic@helsinki.fi

Yves Scherrer
Department of Digital Humanities
University of Helsinki
yves.scherrer@helsinki.fi

Abstract

This paper presents OcWikiDisc, a new freely available corpus in Occitan, as well as language identification experiments done as part of the corpus building process. Occitan is a regional language spoken mainly in the south of France and in parts of Spain and Italy. It exhibits rich diatopic variation, it is not standardized, and it is still low-resourced, especially when it comes to large downloadable corpora. In an effort to remedy this lack, we created OcWikiDisc, a corpus extracted from the talk pages associated with the Occitan Wikipedia. The version of the corpus with the most restrictive language filtering contains 8K user messages for a total of 618K tokens. The language filtering is performed based on language identification experiments with four off-the-shelf tools, including HeLI (Jauhiainen et al., 2022) and a new fasttext-based language identification model from Meta AI's No Language Left Behind initiative (Costa-jussà et al., 2022).

1 Introduction

This paper provides two main contributions: we present OcWikiDisc, a new, freely available corpus in Occitan, and report results of language identification experiments executed as part of the corpus-building process. Occitan is a Romance language, mainly spoken in the south of France and in parts of Spain and Italy. It is considered a regional language in France but doesn't have the status of an official language. Despite recent efforts to endow it with various NLP tools, it still remains low-resourced, especially when it comes to large, freely available corpora. Our OcWikiDisc corpus aims to remedy this lack by relying on user-generated content available on the Web: we extract the corpus from the talk pages associated with the Occitan Wikipedia. Thus, OcWikDisc contains messages posted by users, typically in direct user-to-user interactions. As such, it offers interesting possibilities for research not only in NLP, but also in corpus-based dialectology and wider linguistic studies. To the best of our knowledge, it is the first such corpus for Occitan. It can be downloaded through Zenodo[1].

Since the extracted content contains a significant proportion of messages written in languages other than Occitan, we perform language identification (LID) experiments. We test four off-the-shelf tools: langid (Lui and Baldwin, 2012) and its Python 3 implementation, py3langid [2], developed by A. Barbaresi; HeLI (Jauhiainen et al., 2016, 2022); and the fasttext language identification models, both the original (Joulin et al., 2017) and the most recent (Costa-jussà et al., 2022), published as part of Meta AI's No Language Left Behind Initiative.[3] We identify optimal LID strategies based on the desired outcome (optimizing for precision vs recall) and use them to filter the extracted corpus. These results also offer useful pointers for LID of Occitan in general.

The remainder of the paper is organized as follows. In Section 2, we give a brief description of the main linguistic properties of Occitan. Section 3 offers more details on available NLP tools and resources for Occitan. In Section 4, we describe our corpus extraction process and present the initial corpus. Section 5 is dedicated to language identification experiments, leading to several filtered versions of the corpus, which are presented in Section 6. Finally, in Section 7, we give our conclusions and directions for future work.

2 Occitan: Linguistic Properties and Dialectological Situation

Occitan is a Romance language spoken in the south of France, in parts of Piedmont in Italy and in Val d'Aran in Spain. It does not have the status of an

[1] https://doi.org/10.5281/zenodo.7079580
[2] https://github.com/adbar/py3langid
[3] https://ai.facebook.com/research/no-language-left-behind/

Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 70–79

Figure 1: Occitan dialects

(1)

T’	aviái	laissat	un	messatge	totara
you.DAT	have.1SG.IMPF	leave.PST.PTCP	a.SG.M	message	just

’I had just left you a message’

official language in France, and as many such linguistic varieties, it is not standardized. Currently, two main spelling norms are in use: one close to medieval troubadours’ spelling (often referred to as *classical*) and another one closer to the French language spelling conventions (often referred to as *mistralian*) (Sibille, 2002). Furthermore, Occitan has a rich system of dialects organized in six main groups: Auvernhat, Gascon, Lemosin, Lengadocian, Provençau and Vivaroaupenc (see Figure 1) (Bec, 1995). Diatopic variation can be seen on the lexical, phonological, morphological or syntactic level. For an illustration of each of these types of variation, see Miletic et al. (2020b).

Some of the main linguistic properties are shared by most dialects. For example, Occitan is a null subject language with tense, person and number inflection marks on finite verbs for each person. Many dialects exhibit number and gender inflection on all components of the noun phrase. Unlike contemporary French, Occitan maintains the use of the preterite (*passat simple*), which contrasts with the perfect tense (*passat compausat*), and the use of the imperfect subjunctive, even in informal language. Example 1, extracted from the OcWikiDisc corpus, illustrates some of these properties.

3 Occitan and NLP

Until recently, Occitan belonged to the group of under-resourced languages. This situation was due to a combination of factors. First, the linguistic situation described above, compounding strong diatopic variation, absence of standardization, and use of multiple spelling norms, contributed to data sparsity. This was coupled with insufficient recognition on the institutional level, leading to a lack of human and financial resources available for NLP of Occitan. This situation is currently evolving for the better: in France, regional languages have been recognized as part of the country’s cultural heritage by the constitutional amendment Article 75-1 published in 2008. Since then, they have benefited from national and European initiatives to revitalize regional languages and help them enter the digital era. This has led to the creation of initial resources and tools for Occitan.

An electronic lexicon in Lengadocian (Bras et al., 2020; Vergez-Couret, 2016) (850K entries), an online corpus of 3,4M words called BaTelÒc (Bras and Vergez-Couret, 2016) and a PoS- tagged corpus of 12K tokens (Bernhard et al., 2018) were created as part of the RESTAURE project[4] (2016-2018). During the LINGUATEC project,[5] a 20K-token treebank following Universal Dependencies annotation guidelines was created (Miletic et al., 2020a). The existence of annotated training corpora led to initial experiments in PoS-tagging and parsing Occitan (Vergez-Couret and Urieli, 2014; Miletic et al., 2019). A first neural text-to-speech model has also

[4] https://restaure.unistra.fr/en/presentation/

[5] https://linguatec-poctefa.eu/fr/projet/

been created (Corral et al., 2020).

All of these resources represent important steps forward for Occitan. Nonetheless, the language still remains low-resourced, especially when it comes to large, freely available corpora. The annotated corpora cited above are downloadable for research purposes, but they are fairly small, whereas BaTelÒc, the largest currently available corpus in Occitan, is not downloadable due to copyright limitations. A popular solution in this type of situation is to turn to the linguistic content available on the internet. This typically consists in crawling the top-level domain of the given language and transforming it into a corpus (see, *e.g.* Ljubešić and Klubička, 2014). However, as pointed out by Barbaresi (2013), such an approach can be ill-suited for low-resourced languages and linguistic varieties. Many of them (including Occitan) do not have a dedicated top-level domain, which makes the identification of URL targets for crawling more challenging, and reliable LID systems more crucial in the process. Moreover, low-resourced languages can also have a limited presence on the Internet compared to more widely used languages. To illustrate, the latest version of the OSCAR corpus (Ortiz Suárez et al., 2019)[6], based on the CommonCrawl from November/December 2021, only contains 31K tokens in Occitan, compared to, *e.g.* 41G tokens in French. We therefore turn to a more targeted solution: extracting content from Wikipedia.

4 Extracting a Corpus from Wikipedia Talk Pages

Wikipedia content in Occitan has been extracted and used as a corpus in previous research. For example, it is mentioned as part of the training material for the transformer-based multilingual language model mBERT (Devlin et al., 2019), but also for the LID tools fasttext (Joulin et al., 2017), langid (Lui and Baldwin, 2012) and HeLI (Jauhiainen et al., 2016). To the best of our knowledge, these training corpora have not been distributed.

Wikipedia content in Occitan is also present in parallel corpora extracted from Wikipedia such as WikiMatrix (Schwenk et al., 2021). It would be possible to derive a monolingual Occitan corpus from it, but the resulting corpus would contain individual sentences that would appear without their linguistic context, and would be accompanied by limited metadata.

Our goal, as stated in Section 1, is to create a corpus suitable for a wider range of research: NLP applications, computational and corpus-based dialectology, as well as corpus-based linguistics. We therefore aim to preserve the linguistic integrity of the content in the corpus and to provide as much metadata as possible. Also, we choose to focus on the talk pages rather than the encyclopedia pages themselves. Wikipedia talk pages are dedicated to discussions between users, typically about article content and editing policies. These are direct user-to-user interactions on a variety of topics, but in general they share the same goal: improving the quality of the Wikipedia content. They often combine elements of dialogue with elements of argumentative writing (Ho-Dac et al., 2016). Given the nature of their content, the talk pages are a novel source of linguistic material for Occitan.

4.1 Data Extraction Process

As the starting point of the extraction, we use a Wikimedia data dump containing the current version of Wikipedia pages and the associated metacontent.[7] The basic data structure of the archive is encoded in XML, but the content of each page is rendered in wikitext, a text-based encoding convention that can mark some further structure (thread headings, comments), indicate hyperlinks (username mentions, internal or external page addresses) or allow for some formatting (headings, bulleting, emphasis).

Our global workflow is organized into two main steps: extraction and filtering. The extraction starts by selecting XML elements in an XML namespace dedicated to discussions. For each such discussion, the text content is extracted and individual posts are identified. We also extract some metadata encoded in the XML: contributor, timestamp, namespace and discussion title. However, these pieces of metadata are available at the discussion level, and the corresponding discussion can contain multiple messages, or even multiple threads of messages. Therefore, we also extract the header of the thread in which a given message was posted, along with the username and the timestamp present in the post's signature. All of these pieces of information are preserved as metadata associated with the message in the output.

In the second step, the text of each identified

[6] https://oscar-corpus.com/post/oscar-v22-01/

[7] Dump date: 01 May 2022.

message is cleaned for formatting commands written in wikitext and various types of non-linguistic content, such as snippets of JavaScript or HTML code.

4.2 Initial Extraction Result

The extracted corpus is formatted as a simple CSV file, in which each line represents a message extracted from the corpus. The line contains the message itself and all the extracted metadata associated with it.

Some basic quantitative information about the resulting corpus is given in Table 1. In order to provide token counts, we perform tokenization on whitespace and punctuation marks (including apostrophes). This rudimentary solution was chosen to accommodate the fact that the corpus content is multilingual (see below).

Messages	11,025
Tokens	1,186,239
Tokens/Message	107.60
Users	522
Messages/User	17.07

Table 1: OcWikiDisc: initial extraction

The building process for Web-based corpora typically includes a deduplication step, in which identical (or near-identical) texts are eliminated from the corpus. Currently, this operation is not done on the OcWikiDisc corpus. Given the structure of the data, it should not be possible for the same message with the same metadata to appear multiple times in the archive (each discussion being represented exactly once in the XML file). Some near-identical system messages were present in the initial extraction result, but these are systematically in English and can therefore be eliminated through LID (described below). There are also messages in Occitan that could be classified as near-duplicates, which typically contain demands for article validation, birthday and New Year's wishes. However, these were not produced by bots, but by the contributors, and as such, they represent genuine linguistic material. Furthermore, they are often part of message threads, and excluding them automatically could compromise the integrity of the content.

The Occitan content in the corpus is fairly uniform when it comes to the spelling norm: the community strongly recommends the use of the classical norm in the articles in order to facilitate searches, and this seems to be respected almost systematically in the discussions too. When it comes to the use of dialects, there is an incentive to preserve the identity of each individual dialect and especially to avoid writing in "pan-Occitan", an improvised standard. An initial exploration of the data shows Lengadocien as the most widely used dialect in OcWikiDisc, followed by Gascon and Provençau (see also Section 5.3.1).

However, an important part of the messages contain linguistic material in languages other than Occitan. We therefore perform language identification experiments in order to identify the optimal approach to filter the corpus content. In this first set of experiments on OcWikiDisc, we focus on identifying messages containing Occitan and leave the identification of individual dialects for future work.

5 Language Identification Experiments

Language identification is an NLP task which consists in automatically identifying the language of a given text (Jauhiainen et al., 2019). In order to perform this task on the extracted corpus, we first evaluate four off-the-shelf tools that integrate models for Occitan. Each of them is briefly presented below.

5.1 Language Identification Tools

langid (Lui and Baldwin, 2012) uses a multinomial naïve Bayes model with feature selection based on an information gain measure. The features are not complete words, but character n-grams (1 to 4 characters). It is specifically designed to control for genre differences and bias towards better resourced varieties. In addition to the original tool, we also test its Python 3 implementation, **py3langid**, developed by A. Barbaresi [8]. Both tools were trained on the same set of 97 languages.

HeLI (Jauhiainen et al., 2016, 2022) uses language models consisting of single words and character n-grams of length 1 to 6. During training, the models are created by attributing each word or n-gram a score based on its relative frequency in the given language. During language identification, for each word of the text to be classified, the tool first calls upon the word-level models. If the word is found in none of them, the tool backs off to n-gram models, going from longest to shortest, until at least one match for the word is found. The scores of all languages identified in a given instance are

[8] https://github.com/adbar/py3langid

averaged to obtain the final score for each of them. HeLI integrates models for 200 languages.

fasttext (Joulin et al., 2017) was designed as a general text classification model, but its LID models have been widely used. It implements a language representation based on bag of words and bag of n-grams. It uses a linear classifier combined with a rank constraint, supposed to improve the generalisation for classes with small numbers of instances. We test both the LID model distributed with the original version of the tool as well as a more recent one, released in July 2022 as part of Meta AI's No Language Left Behind initiative (Costa-jussà et al., 2022). This initiative being specifically aimed at low-resourced languages, we wish to evaluate the tool's performances on Occitan. The original model was trained on 176 languages, while the most recent one integrates 204.

5.2 Baseline Evaluation on Existing Occitan Data

We perform an initial LID evaluation on a test set containing only Occitan. The sample is derived from the four-dialect treebank presented in Miletic et al. (2020b) by transforming each treebank sentence into a test instance. The sample contains 1,520 instances, 73% of which are in Lengadocian, 17% in Gascon, and 5% in Provençau and Lemosin each. However, for the purposes of this experiment, all dialects were merged.

We report accuracy scores for each tool in Table 2. The more recent fasttext LID model (fasttext2) achieves the best result at 93.22%, with an improvement of almost 30 percentage points over the previous version of the model (*fasttext1*). The only other tool scoring above 90% is HeLI, with langid at and py3langid at 66.64% and 70.00% respectively.

Given these results, we keep fasttext2 and HeLI for some further experiments: we test using the top-2 predictions from each tool (heli_top2 and fasttext2_top2), and then using the union of the top prediction from each of them (fasttext2_heli). We scored the prediction as true if the list of labels contained Occitan. As shown in the section *Strategies* of Table 2, relying on two labels from HeLI achieves the same score as using the top prediction from fasttext2. Using the top 2 labels from fasttext2 improves accuracy for almost 2%, but using HeLI's top prediction instead brings a small additional improvement, equivalent to another 3 correct predictions on this dataset. This is the best overall result in this part of our evaluation.

Individual tools	
Tool	**Accuracy (%)**
fasttext1	62.30
langid	66.64
py3langid	70.00
heli	90.70
fasttext2	93.22
Strategies	
Strategy	**Accuracy (%)**
heli_top2	93.22
fasttext2_top2	95.00
fasttext2_heli	95.20

Table 2: LID results on all-Occitan dataset

As mentioned above, a concern when attempting LID on low-resourced languages is that they will be confused with better resourced closely related linguistic varieties. We can therefore expect the tools to encounter difficulties in distinguishing Occitan from other Romance languages. The confusion matrices based on the classification produced by fasttext2 and HeLI seem to confirm this. Table 3 shows the ten most frequent erroneous labels produced by the two tools.

For both tools, 7 out of 10 most frequently confused languages are from the Romance family. In the case of HeLI, Interlingua[9] and Haitian can also claim closeness to the Romance languages. On the other hand, the remaining languages for fasttext2 are somewhat surprising: there seems to be no straightforward linguistic argument for confusing Occitan with Vietnamese or Standard Malay.

This evaluation allowed us to quickly identify potentially useful strategies for LID on our corpus. However, since the initial test set only contains Occitan, it is not possible to evaluate the tools' precision in a satisfactory manner. We therefore proceeded to an evaluation on a sample extracted from OcWikiDisc in order to further test the tools in a context closer to their intended use. For these experiments, we select fasttext2 and HeLI as the most reliable systems.

[9]Interlingua is a constructed language whose vocabulary and grammar are largely based on Romance languages. See, e.g., (Gode and Blair, 1951; Gode et al., 1952)

fasttext2		heli	
Catalan	23	Catalan	38
French	11	Spanish	11
Vietnamese	10	Interlingua	7
Portuguese	8	Lombard	6
Spanish	6	French	6
English	5	Extremaduran	5
Asturian	5	Piemontese	4
Galician	5	Portuguese	4
Standard Malay	4	Haitian	3
Italian	4	Pfälzisch	3

Table 3: Top-10 erroneous labels for fasttext2 and HeLI

5.3 Evaluation on OcWikiDisc

As stated above, the content of OcWikiDisc is not written exclusively in Occitan. The content in other languages can appear as a monolingual post, or as a part of a multilingual message. These multilingual examples can also include Occitan. This has important implications both for LID itself and for our evaluation setup.

LID in multilingual and, in particular, code-switching data is a challenge for LID systems (Jauhiainen et al., 2019). One of the central issues is the need to determine how many labels need to be attributed to each classification instance. This often implies determining a threshold for the classification score and accepting all predictions that score above it to contribute to the prediction.

When it comes to the evaluation, this type of material raises questions about the manual annotation guidelines. For instance, if a message contains only a toponym (cf. *He lives in Teste de Buche*), a metalinguistic use of a word (cf. *the word 'caval' means 'horse'*), or a salutation (cf. *Bonjorn, I would like to participate in writing this article*) in a different language, should it be labelled as multilingual? We address these questions below.

5.3.1 Building a Multilingual Evaluation Sample

For the purposes of this evaluation, we create a test set of 100 messages extracted from the corpus. Roughly a third of the instances contain no Occitan (but can contain several other languages), a third contains only Occitan, and a third contains Occitan and at least one other language. The sample was manually annotated by a single annotator. For each post, the annotator indicated all languages appearing in it, even if one of them was only instantiated in a single word. Out of the 100 test instances, 58 are monolingual, with the average number of labels per instance at 1.49. The maximum number of labels per instance is 4.

The sample was also annotated with dialect and spelling norm information. The Occitan content in this sample systematically follows the classical spelling norm. As for the dialects, out of 68 messages containing Occitan, 36 were in Lengadocian, 6 in Gascon and 5 in Provençau, whereas for the remaining 21 it was impossible to specify the dialect. However, this information was not used in the experiments described in the following section, which focus solely on language identification.

Some factors should be borne in mind while considering the evaluation results presented below. In the current annotation all labels are presented equally: there is no means of knowing how the content of the post is distributed between different languages. It is also worth mentioning that this was not a trivial task for the human annotator: she reported uncertainty about a part of the languages in the test set and had to rely on help from other linguists to identify some of them.

5.3.2 Evaluation on a Multilingual Sample

We frame our evaluation as a task in identifying Occitan content in the corpus. We therefore focus our attention on the tools' performance relative to this language, at the expense of their global results.

In order to determine the number of labels from each tool to be evaluated, we first considered using a threshold on the classification scores. However, this proved problematic with fasttext2. The tool's second-best predictions are associated with an important drop in probability, with 75% of them scoring at <0.021. A meaningful threshold would therefore favour outputting only one label from fasttext2. Yet our initial evaluation suggests that additional labels would be useful for the task at hand. We therefore opted for a different approach: we base our evaluation on top-2 and top-5 labels from each tool. This, of course, affects the global precision scores, since it automatically produces incorrect labels for monolingual posts. However, as noted above, our aim is to optimize the detection of Occitan, and not the global LID scores.

The evaluation results are presented in Table 4. We evaluate tools individually on their top-2 and top-5 labels, but also on two ensemble strategies, combining the top prediction and the top-2 pre-

	Occitan			Global		
	Precision	Recall	F1-score	Precision	Recall	F1-score
fasttext2_top2	84.75	73.53	78.74	**56.50**	75.84	**64.76**
heli_top2	**93.33**	61.76	74.34	51.00	68.46	58.45
fasttext2_top5	79.49	**91.18**	84.93	26.00	**87.25**	40.06
heli_top5	88.06	86.76	**87.41**	25.41	83.22	38.93
fasttext2_heli_top1	**100.00**	57.35	72.90	**89.09**	65.77	**75.68**
fasttext2_heli_top2	85.00	**75.00**	**79.69**	42.80	**77.85**	55.24

Table 4: Evaluation results on OcWikiDisc sample

	Messages	Tokens	Tokens/Message	Users	Messages/User
ocwikidisc_precision	8149	618,153	75.86	206	33.69
ocwikidisc_balanced	9032	756,922	83.80	323	23.19
ocwikidisc_recall	9394	804,959	85.69	347	22.39
ocwikidisc_unfiltered	11025	1,186,239	107.60	522	17.07

Table 5: OcWikiDisc: filtered corpora

	Total languages	Top 11
ocwikidisc_precision	54	Occitan, Catalan, French, English, German, Spanish, Portuguese, Lombard, Romanian, Piemontese, Galician
ocwikidisc_balanced	124	Occitan, Catalan, Extremaduran, Lombard, Spanish, Interlingua, French, Galician, Piemontese, Portuguese, Lingala
ocwikidisc_recall	114	Occitan, Catalan, French, Spanish, Galician, Portuguese, Lombard, Italian, Asturian, Korean, Romanian
ocwikidisc_unfiltered	155	Occitan, Catalan, French, Spanish, Portuguese, Galician, Italian, Korean, Lombard, English, Asturian

Table 6: Overview of languages detected in different versions of the corpus

dictions from each tool. We report the results on Occitan, and include global evaluation scores for the sake of completeness.

In strategies combining output from fasttext2 and HeLI, we use the union of the labels produced by each tool. This yields an average of 1.1 labels per instance when using the top prediction from each, and 2.7 labels per instance on average when using the top 2 predictions.

In all scenarios, we evaluate the tools in terms of precision, recall and F1-score. For the global evaluation results, the scores are micro-averaged.[10]

First, let us comment briefly the global evaluation results. As expected, the scenarios with a higher number of labels achieved the best recall, but had significantly lower precision scores, leading to lower F1 scores. The best precision was obtained with the combination of the top prediction from fasttext2 and HeLI, which also shows the best F1 score. This could therefore be considered as a sound option for optimizing the LID results on all languages.

When it comes to the identification of Occitan, the results are more surprising. Unlike what we saw in the initial evaluation, combining the two tools does not seem to improve over the best individual results. The highest F1-scores were achieved by HeLI using the top-5 predictions (87.41) and fasttext2 (84.93) in the same setup. HeLI also displays balanced precision and recall scores in this setup, which recommends it as a reliable global solution for our task. Using the combination of the top predictions from fasttext2 and HeLI achieves perfect

[10]For the evaluation on Occitan only, we evaluate recall based on all manually annotated messages that are labelled as containing Occitan, whereas the precision takes into account all predictions that contain the label for Occitan.

	Users with >1 message	Messages from top 10	Tokens from top 10
ocwikidisc_precision	120 (58%)	5,173 (63%)	399,530 (65%)
ocwikidisc_balanced	166 (51%)	5,346 (59%)	435,345 (58%)
ocwikidisc_recall	188 (54%)	5,392 (57%)	456,839 (57%)
ocwikidisc_unfiltered	257 (49%)	5,757 (52%)	552,669 (47%)

Table 7: Distribution of content across users

precision on our sample, but it is coupled with a significant drop in recall (57.35). Unsurprisingly, the best recall was achieved when using the highest number of labels (fasttext_top5 and heli_top5).

Based on these results, we choose the following strategy for our corpus-building process. We annotate the corpus both with fasttext2 and with HeLI, outputting the top-5 labels from each. We create three filtered versions, favouring precision (using fasttext2_heli_top1), recall (using fasttext2_top5) and F1-score (using heli_top5), respectively. Each of the filtered versions is presented below. Through this approach, we hope to produce resources adapted to different types of applications and research. An unfiltered version of the corpus is also made available.

6 Filtered Corpus

In this section, we present the complete LID-annotated corpus and its three filtered versions. The basic information about them is available in Table 5, whereas the detected languages in each version of the corpus are presented in Table 6. To facilitate comparison, we repeat the same information for the unfiltered version of the corpus, initially given in Section 4.2.

As expected, the version of the corpus favouring precision (ocwikidisc_precision) is the most restricted, with 8K messages and 618K tokens. This represents roughly half of the unfiltered corpus (in tokens). The difference between the corpus favouring recall (ocwikidisc_recall) and the one favouring F1-score (ocwikidisc_balanced) is relatively small for all reported measures. It remains to be seen if there is a qualitative difference in their content.

It is important to note that the distribution of content across users is heavily skewed in all four versions of the corpus, both in terms of the number of messages and in terms of the number of tokens. The full distribution of messages across users is shown in Figure 2. As illustrated in Table 7, more than half of the content in each filtered version comes from the 10 most active users, and only 50-60% of users have produced more than one message. While this affects the representativeness of the corpus, it offers an interesting possibility for dialect identification: if the dialect of each of the most active users can be reliably identified manually, this information can be propagated onto all of their messages, thus annotating an important part of the corpus for dialect information. This direction will be explored in our future work.

The information on detected languages in Table 6 is based on the predictions of the strategy used to filter a given corpus. Note that the 10 most frequent languages after Occitan in each corpus predominantly belong to the Romance family. This could simply be the result of shared interests or collaboration efforts on Wikipedia, but it could also be an indicator of difficulties with the identification of closely related languages. We will be looking into this issue in the future.

7 Conclusions and Future Work

In this paper, we presented OcWikiDisc, a new corpus in Occitan extracted from Wikipedia Talk pages. The version of the corpus with the most restrictive language-based filtering contains 618K tokens. Along with its extracted content, it also contains metadata about users, time of posting and discussion subjects, as well as language annotation produced using LID tools. To the best of our knowledge, it is the largest downloadable corpus for Occitan. It can be downloaded from Zenodo.

We also presented LID experiments aimed at identifying Occitan content in the initial extracted corpus, which is multilingual. We tested four off-the-shelf LID tools. In an initial experiment on an all-Occitan sample, the best results were achieved by the new LID model from the fasttext tool and by HeLI. On a test sample extracted from OcWikiDisc, fasttext's new model had the highest recall score, whereas HeLI achieved the most balanced precision and recall. Combining the two tools optimized the

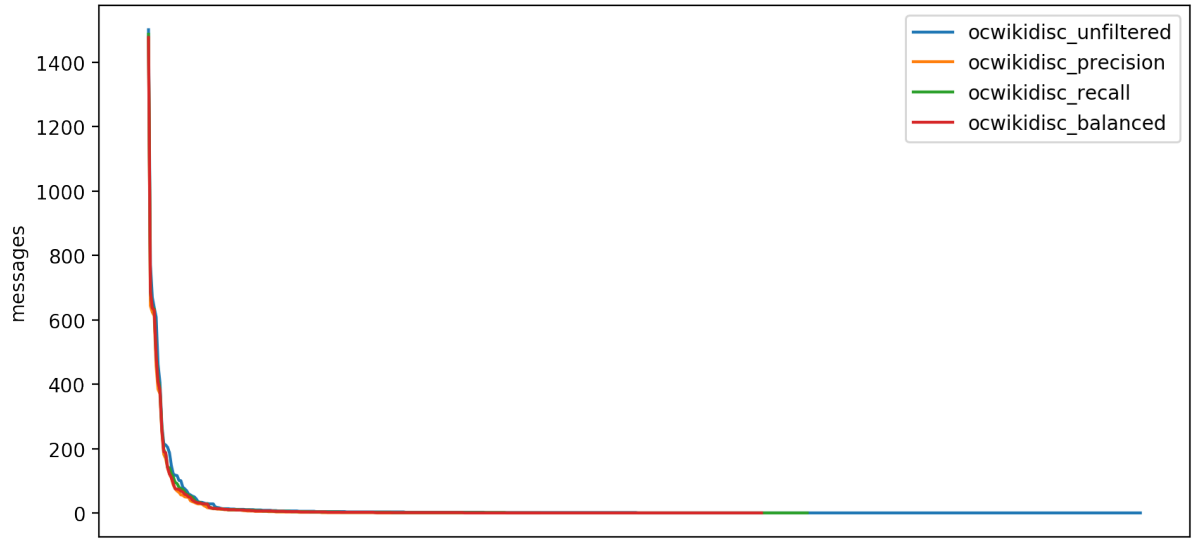

Figure 2: Number of messages per user across the four versions of the corpus

precision.

In the future, we will investigate making the LID on the corpus more fine-grained. Currently, we perform LID at message level. Given the amount of multilingual messages observed in our data, it could be beneficial to do it rather at sentence level, or even at word level. We will also examine the annotation of the Romance languages found in the corpus, since a certain amount of confusion arising from the closely related languages in the corpus can be expected.

Acknowledgements

The authors would like to thank Jean Sibille (University of Toulouse), Myriam Bras (University of Toulouse) and Marianne Vergez-Couret (University of Poitiers) for their help with the manual annotation of the test sample.

This work has been supported by the Academy of Finland through project No. 342859 “CorCoDial – Corpus-based computational dialectology”.

References

Adrien Barbaresi. 2013. Challenges in web corpus construction for low-resource languages in a post-bootcat world. In *6th Language & Technology Conference, Less Resourced Languages special track*, pages 69–73.

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

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 *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3917–3924, Miyazaki, Japan. European Language Resources Association (ELRA).

Myriam Bras and Marianne Vergez-Couret. 2016. BaTelÒc : a text base for the Occitan language. In *Language Documentation and Conservation in Europe*, pages 133–149. Honolulu: University of Hawaï Press .

Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benazet Dazéas. 2020. Lofiòc : Lexic obèrt flechit occitan. In *Fidélités et dissidences (Actes du XIIe congrès de l’Association Internationale d’Études Occitanes)*, pages 141–156, Albi. Centre d’Etude de la Littérature Occitane.

Ander Corral, Igor Leturia, Aure Séguier, Michael Barret, Benaset Dazéas, Philippe Boula de Mareüil, and Nicolas Quint. 2020. Neural text-to-speech synthesis for an under-resourced language in a diglossic environment: the case of Gascon Occitan. In *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop «Language Resources and Evaluation Conference–Marseille–11–16 May 2020»*, pages 53–60. European Language Resources Association (ELRA).

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume

Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint: https://arxiv.org/abs/2207.04672.

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.

Alexander Gode and Hugh E Blair. 1951. *Interlingua: a grammar of the international language*. Storm Publishers.

Alexander Gode, Hugh E Blair, and Forrest F Cleveland. 1952. Interlingua-english, a dictionary of the international language and interlingua, a grammar of the international language. *American Journal of Physics*, 20(6):382–382.

Lydia-Mai Ho-Dac, Veronika Laippala, Céline Poudat, and Ludovic Tanguy. 2016. French Wikipedia talk pages: Profiling and conflict detection. In *4th Conference on CMC and Social Media Corpora for the Humanities*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3912–3922.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2016. HeLI, a word-based backoff method for language identification. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 153–162, Osaka, Japan. The COLING 2016 Organizing Committee.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Aleksandra Miletic, Myriam Bras, Louise Esher, Jean Sibille, and Marianne Vergez-Couret. 2019. Building a treebank for Occitan: what use for Romance UD corpora? In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 2–11, Paris, France. Association for Computational Linguistics.

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

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).

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

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. 2016. Description du lexique Lofl`oc. Research report, CLLE-ERSS.

Marianne Vergez-Couret and Assaf Urieli. 2014. Postagging different varieties of Occitan with single-dialect resources. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 21–29.