OcWikiDialects: A Wikipedia Dataset With Rich Metadata for Occitan Dialect Identification

Oriane Nédey **Rachel Bawden** **Thibault Clérice** **Benoît Sagot**
Inria, Paris, France
{firstname.lastname}@inria.fr

Abstract

Occitan is a Romance language spoken mostly in the South of France and characterised by rich dialectal variation, which can pose problems for certain NLP tools. This shortfall is largely attributable to the scarcity of dialect-annotated corpora, in a context where linguistic classification within the Occitan dialect continuum is still debated and major nomenclatures, such as ISO 639, fail to provide granular codes for varieties below the generic "Occitan" label. In this paper, we introduce OcWikiDialects, a new dataset comprising articles from the Occitan Wikipedia. The corpus features rich metadata, including dialect labels, and is segmented at both paragraph and sentence levels. Combined with previously released datasets, we explore approaches for Occitan dialect identification by training three types of model on up to 8 labels: linear SVM classifiers based on word and character n-grams, FastText classifiers based on pretrained vectors, and BERT-based neural classifiers adapted through fine-tuning. Evaluations across in- and out-of-domain test sets demonstrate the substantial impact of our new dataset for the task. However, a peak macro-averaged F1 score of 58.15 underscores persistent challenges for underrepresented Occitan varieties, supported by our per-dialect analysis. Code, dataset and models are available: https://github.com/DEFI-COLaF/OcWikiDialects.

1 Introduction

Current NLP technologies offer very good performance in particular with large language models (LLMs), including in some low-resource settings (Pomerenke et al., 2025). However, most available tools consider the supported languages as standardised monolithic entities, thus hiding many aspects of variation that occur in natural languages (Bird, 2022). Ignoring these aspects in the NLP development process has an impact on speakers of under-represented and less standardised varieties, as revealed in studies comparing performance between a standard variety (e.g. Standard American English or Modern Standard Arabic) and related non-standardised variants (e.g. African-American English or Moroccan Arabic) (Khondaker et al., 2023; Okpala and Cheng, 2025; Gupta et al., 2025).

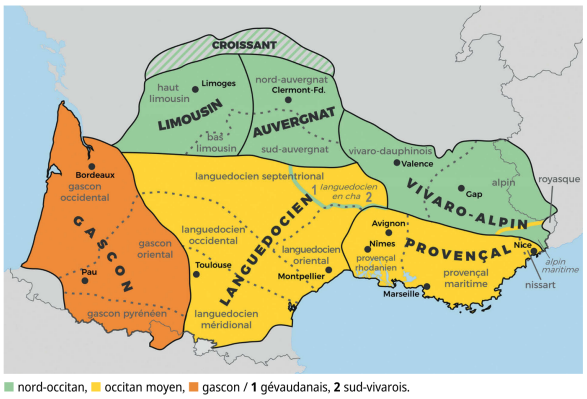


Figure 1: The dialect classification for Occitan proposed in (Sibille, 2024).

In this paper, we focus on Occitan, a dialect continuum spoken mainly in the south of France and supported as a monolithic entity in several NLP models such as pretrained multilingual BERT (Devlin et al., 2019) and the machine translation model NLLB (Costa-jussà et al., 2024). Existing Occitan NLP datasets increasingly cover diverse aspects of internal variation, including dialectal (Miletic et al., 2020), orthographical (Poujade et al., 2024) and other diastratic (Nédey et al., 2025) factors. However, to the best of our knowledge, there are no models available for the identification of the Occitan variety of a text, although such tool could be crucial for building and evaluating dialect-aware NLP systems.

Our contribution is twofold: (1) we introduce a new dialect-labelled dataset OcWikiDialects, and (2) we use it in combination with other datasets to train baseline models for the Occitan dialect identification (OCDI) task. Our dataset is made up

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 45–57
March 29, 2026.

of articles from the Occitan Wikipedia and comprises metadata about dialects and users, aggregated across revisions. Our OCDI experiments involve three types of model: linear SVM based on words and character n-grams, FastText (Joulin et al., 2017) with pretrained vectors, and BERT (Devlin et al., 2019) with pretrained models. By analysing results across three model types, five datasets and up to eight dialects, we demonstrate the impact of the OcWikiDialects dataset and emphasise the need to develop more resources for underrepresented Occitan varieties.

2 Related Work

2.1 Occitan Dialects and Datasets

Occitan is a dialect continuum that spans across the south of France and some border regions of Spain and Italy (see Figure 1). Linguistic studies based on isoglosses and dialectometry experiments led to a classification into dialects within the continuum (Esher and Sibille, 2024) that are sometimes used by speakers when referring to their language. Dialectal variation of Occitan occurs at multiple linguistic levels including phonetic, morphologic, syntactic and lexical. While the so-called and most popular *classical* spelling convention tends to reduce dialect variation in writing, some features remain visible (see some examples from our dataset OcWikiDialects in Table 1), such as the vocalisation of [l] into [w] in final positions (e.g. *aquel* vs. *aqueu* for 'this'), the presence of the enunciative particle *que*, or the use of definite plural articles *li*, *lu*, or *lei* (vs. *los* or *las*).

Next to large textual datasets using the monolithic label "Occitan" (Schwenk et al., 2021; Miletic and Scherrer, 2022; Costa-jussà et al., 2024; Penedo et al., 2025), smaller Occitan corpora include dialect labels: Tolosa Treebank (Miletic et al., 2020) is a corpus of literary texts from four Occitan varieties, manually annotated with morphosyntactic information. Similarly, CorpusArièja (Poujade et al., 2024) focuses on morphosyntactic annotation of literary texts from a transitional area between the Languedocian and Gascon varieties. The organisation *Lo Congrès permanent de la lenga occitana* compiled bilingual sentences (Occitan and French or other standardised languages) from their websites (Séguier and Lo Congrès, 2023a, 2024) and from various translated tools (Séguier and Lo Congrès, 2023b) into corpora where the document-level dialect label is available for each sentence. The organisation also released the ReVoc dataset (Lo Congrès, 2024), which contains not only sentences and corresponding dialect labels, but also information about the speakers of the related speech data collection campaign. The inclusion of sociological metadata was also emphasised in Nédey et al. (2025) for the dataset ForumOccitania, which is made of posts from an online forum, accompanied by user-declared information such as dialect, geographical location and age.

Despite the existence of these datasets, the ratio of dialect annotated data remains very low, and the distribution of dialects is very uneven, with a large over-representation of the Languedocian variety while some others are barely represented (e.g. Auvergnat, Vivaroalpine).

2.2 Language Identification for Similar Varieties and Dialects

While the most popular tools for language identification cover a large number of languages (Grave et al., 2018; Kargaran et al., 2023; Costa-jussà et al., 2024), many dialectal variants remain unsupported. This is partly due to their reliance on major language nomenclatures such as ISO 639, which typically lack granular codes for varieties below the level of macro languages, for instance Brazilian and European Portuguese. Works published during VarDial workshops reveal that the identification of similar languages and dialects remains challenging (Aepli et al., 2023), especially in the context of dialect continuum where established linguistic categories have overlapping and occasionally contested boundaries (Aepli et al., 2022).

For such varieties, a text can often be valid for several labels, motivating recent work in similar language identification (Keleg and Magdy, 2023; Bernier-colborne et al., 2023; Chifu et al., 2024; Fedorova et al., 2025) to emphasize a shift toward multi-label classification. Although FastText is the most popular architecture to discriminate between many high-resource languages, its usage seems less popular in the context of discriminating between similar languages (Fedorova et al., 2025), where the most common approaches are based on statistical models such as Naïve Bayes and SVM, using n-gram-based features, and on pretrained language models derived from BERT (Devlin et al., 2019). The limited amount of data and lack of diversity (Cahyawijaya et al., 2023) that is common in annotated corpora for these languages make the models (especially statistical ones) more likely to over-rely

on named entities or other semantic aspects of the corpora (e.g. names of cities or languages, topics) instead of linguistic features. Sousa et al. (2025) address this bias by creating a multi-domain corpus for two national varieties of Portuguese and by randomly masking named entities and replacing words with their part of speech tag.

There are no publicly available systems to discriminate between varieties of Occitan, although some works on classification have been carried out: Seguier (2015) explores statistical approaches based on word and character n-gram frequencies, as well as manually defined grapheme-based and grammatical-based features, resulting in an almost perfect accuracy on 45 samples in 8 Occitan varieties when using only the character n-gram features. More recently, Nédey et al. (2025) carry out unsupervised topic modelling experiments that result in a classifier able to distinguish between 4 Occitan varieties with a macro F1 score of 85.50, when evaluated on 9.5k in-domain test samples.

3 Dataset Creation

We fetch the clean Markdown text of the Occitan Wikipedia[1] articles from the FineWiki dataset[2] (Penedo, 2025), and we use the XML dump version with complete history[3] and texts with wiki markup (Wikitext) to extract more detailed metadata about the dialects used, the article revisions, and the contributors. We retrieve the dialect label of each article from a tag in the Wikitext when it exists, and drop the article from our dataset otherwise. From the revision history, we extract user IDs and timestamps. We also parse user pages to extract declared Occitan language proficiency levels[4] and dialects, and to mark bot users.

We derive additional metadata from the revision history, such as timestamps at creation and latest update, Occitan level of the first contribution, and highest Occitan level. We use the Wikitext version of each revision to rank users by number and size[5] of contributions on the article, and to aggregate the number and size of contributions by Occitan level.

[1] https://oc.wikipedia.org

[2] https://huggingface.co/datasets/HuggingFaceFW/finewiki

[3] https://dumps.wikimedia.org/ocwiki/20250901/. Since FineWiki was built from the HTML dump of 20250820, we parse the history only up to that date.

[4] See scale in Appendix A.

[5] Absolute difference in bytes with the previous contribution, where negative differences are divided by two, with a minimum value of one.

The clean articles are split into paragraphs (based on empty lines) and into sentences (using NLTK (Bird and Loper, 2004)).

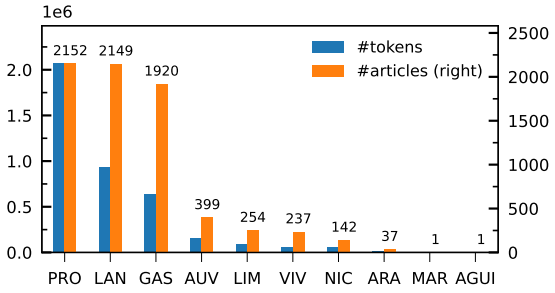


Figure 2: Dialect distribution in OcWikiDialects.

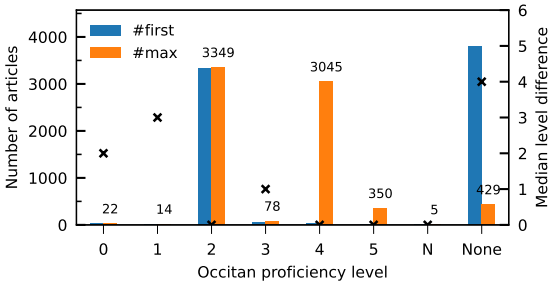


Figure 3: Distribution of self-declared Occitan proficiency levels in OcWikiDialects articles. As #first, we report the number of articles per user-declared Occitan level when considering the first non-empty contribution, and as #max, the number when considering the maximum level over the article's history. Bar labels correspond to #max.

The resulting OcWikiDialects dataset contains 7,292 articles with a dialect label, comprising approximately 57k paragraphs, 290k sentences, and 4M tokens.[6] Dialect labels have ten possible values, six of them corresponding to the usual high-level Occitan varieties (Auvergnat (AUV), Gascon (GAS), Limousine (LIM), Languedocian (LAN), Provençal (PRO) and Vivaroalpine (VIV)), two corresponding to local, very distinct varieties (Aranese (ARA), linguistically closer to Gascon, and Niçard (NIC), linguistically closer to Provençal), and the last two corresponding to transitional varieties between Occitan and *langues d'Oïl*[7] (Aguianese (AGUI) and Marchese (MAR)). The dialect distribution presented in Figure 2 reveals that PRO and LAN are the

[6] Counts after splitting on whitespace and removing punctuation (more frequent with the Markdown format). Paragraphs and sentences containing only punctuation are ignored.

[7] A continuum of northern Romance varieties, encompassing standard French and multiple regional varieties.

most frequently assigned labels (ca. 2k articles), closely followed by GAS. Notably, PRO articles are substantially longer on average than those in the other classes, resulting in PRO being the most represented variety in terms of token count, with over 2M tokens. Moreover, even if dialects AUV and VIV are under-represented in OcWikiDialects, Table 2 shows that the number of samples for these dialects surpasses by far that of other datasets, enabling more reliable evaluations on these varieties.

Out of 1249 active users, only 35 declared a dialect in their user page (see Figure 5 in Appendix B.1), with values only for the five most represented dialects in the dataset. Yet, we observe that over 90% of articles labelled as PRO were edited at least once by one of the 8 active users declaring this dialect.

Our analysis of user-declared Occitan levels (see Figure 3) shows that almost all articles were created by users with a level either unknown or intermediate. However, when considering the maximum level across contributions for each article, a clear shift towards high levels is visible, especially as almost 47% of articles were edited by a user of level 4 (near-native ability) or above, whereas the ratio of articles with an unknown level drops from 52% (#first) to 6% (#max).

Additional analyses in Appendix B.2 show that bots are frequent authors of small contributions, suggesting a limited impact on the textual contents of the dataset. Nevertheless, a seemingly important proportion of articles in OcWikiDialects concern municipalities with very similar patterns and templates (see excerpts in Table 1). This lack of topic, style and syntactic diversity could have a negative impact on downstream NLP applications, as observed by Lambrecht et al. (2022) in Machine Translation experiments on dialect-labelled articles of the Alemannic Wikipedia.

4 Occitan Dialect Identification (OCDI) Experiments

4.1 Methodology

We train OCDI model baselines using three complementary modelling approaches, each providing benefits relevant to specific application scenarios.

SVM We combine n-grams of characters (2-5) and full words (unigrams), vectorised with TF-IDF, to train a linear SVM classifier using `scikit-learn` (Pedregosa et al., 2018).[8] The learned features provide the most straightforward basis for interpretability.

FastText We choose this type of model as it usually provides very good results as well as the fastest inference (Fedorova et al., 2025; Suarez et al., 2026). We train a classifier based on FastText embeddings pretrained on Occitan data. For the embeddings, we compare results obtained when using existing vectors trained on Occitan data from Common Crawl and Wikipedia (CC)[9] (Grave et al., 2018), or when training a new embedding model.

BERT We fine-tune mBERT-cased[10] and mBERT-uncased[11] models (Devlin et al., 2019), multilingual Transformer (Vaswani et al., 2017) encoders that were pretrained partly on Occitan data, and oc-mBERT,[12] the result of continued pretraining mBERT-cased specifically on Occitan data (Hopton and Aepli, 2024).

Hyperparameters for each type of model are described in Appendix D.

We evaluate models using accuracy, recall, precision and F1 metrics, for each dialect and macro-averaged per dialect class. As we train and evaluate separately on datasets that contain differing sets of labels (see Table 2), at test time we group predicted labels that are absent from the test set under a single "other" label, to avoid over-penalising these predictions when computing macro-average scores.

To enable comparison in settings where source datasets differ in both class sets and sample distributions, we report the theoretically expected performance of a random-label baseline. This baseline assigns labels uniformly at random to samples, resulting in a fixed per-class recall equal to the inverse of the number of classes, while the precision for each class corresponds to its empirical prevalence in the test set (i.e., the number of samples of that class divided by the test set size).

4.2 Data

We use the following dialect-labelled datasets to train and evaluate the models on the OCDI task:

[8] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[9] https://fasttext.cc/docs/en/crawl-vectors.html

[10] https://huggingface.co/google-bert/bert-base-multilingual-cased

[11] https://huggingface.co/google-bert/bert-base-multilingual-uncased

[12] https://huggingface.co/zhopto3/oc_mbert

LAN	Menèrba es una comuna lengadociana situada dins lo departament d'Erau.
PRO	Barcilona es una comuna provençala situada dins lo departament **deis Aups** d'**Auta** Provença.
NIC	Lo Puget Tenier es una comuna d'Occitània, dins lo departament **dei Aups** Maritims.
VIV	Chastelnòu de Bordeta es una comuna occitana de Daufinat situ**aa** dins lo departament de Droma.
AUV	Mauriac **z-**es **'na** comuna d'Auvèrnhe ; **z-**es administrada pel departament de**lh Ch**antal.
LIM	Lo dolmen situat entre lo vilatge de la Valada e la D 979 fuguet fortament **'b**imat en 1862 e completament destru**ch** en 1902.
GAS	Biriato **qu'ei ua** comuna de la província tradicion**au** de Labord, administrada **peu** departament **deus** Pirenèus Atlantics **dens** la region de **Navèra** Aquitània.
ARA	Les **ei ua** vila e municipi **dera** Val d'Aran, **en eth** terçon de Quate Lòcs, **ath** marge **deth** riu Garona.

Table 1: Excerpts from OcWikiDialects with some elements of dialectal variation in bold, as compared to LAN. Translations are available in Appendix C.

Dialect	WIKI			TTB			CONGRES			SOFT			FORUM			CONCAT		
	Tr	D	Te	Tr	D	Te	Tr	D	Te	Tr	D	Te	Tr	D	Te	Tr	D	Te
LAN	14575	507	501	615	61	437	9592	500	500	72119	500	500	1410	381	250	98311	1949	2188
GAS	13383	467	500	134	18	103	6821	500	500	103	102	204	1594	242	250	22035	1329	1557
PRO	17985	509	508	45	0	32	59	60	119	666	500	500	704	86	250	19459	1155	1409
LIM	1723	500	503	36	0	41	67	68	135	61	61	122	3619	291	250	5506	920	1051
AUV	2787	512	501	0	0	0	11	12	24	39	39	78	0	0	0	2837	563	603
NIC	127	145	274	0	0	0	0	0	0	861	500	500	0	0	0	988	645	774
VIV	223	223	414	0	0	0	9	10	18	39	40	78	0	0	0	271	273	510
ARA	57	58	122	0	0	0	15	16	32	0	0	0	0	0	0	72	74	154
CONCAT	50860	2921	3323	830	79	613	16574	1166	1328	73888	1742	1982	7327	1000	1000	149479	6908	8246

Table 2: Number of samples in each dataset and split (Tr = train, D = development, Te = test), with distributions per dialect label.

- *OcWikiDialects* (WIKI),[13] 8 dialect labels. Article-level labels were projected onto paragraph-level samples.
- *Tolosa Treebank*[14] (TTB) (Miletic et al., 2020), 4 dialect labels. Document-level labels were projected onto sentence-level samples.
- *ForumOccitania* (FORUM) (Nédey et al., 2025), 4 dialect labels. User dialects were projected onto post-level anonymised samples.
- CONGRES: the deduplicated concatenation of corpora *Lo Congrès Websites*[15] and *Lo Congrès News*[16], 7 dialect labels. [17] Document-level labels were projected onto sentence-level samples.
- *SoftwaresOccitanTranslations*[18] (SOFT), 7 dialect labels. Document-level labels were projected onto sentence-level samples.

We use the existing train/dev/test splits for TTB and FORUM and create splits for the other corpora. The dev and test sets are built by iterating over dialect subsets without exceeding 500 or 50% of samples for each dialect, in order to produce fairly balanced test sets, in terms of dialects and domains. We use paragraph-level samples for OcWikiDialects, and do not mix articles between splits. We concatenate all sources (WIKI, TTB, FORUM, CONGRES and SOFT) with their splits into CONCAT.

Our FastText embedding model is trained on the concatenation of CONCAT-train and the following datasets: NLLB[19] (Costa-jussà et al., 2024), OcWikiDisc[20] (Miletic and Scherrer, 2022), FineWiki[21] (Penedo, 2025), FineWeb2[22] (Penedo et al., 2025), and Tatoeba.[23]

In order to assess the importance of preprocessing on the OCDI task, we train our models separately on raw text and on preprocessed texts. The chosen preprocessing steps take into account the presence of User-Generated Content in ForumOccitania. First, we remove URLs and emails, then we convert accents to their ASCII equivalent. Sequences of three or more of the same character

[13] AGUI and MAR were excluded due to insufficient data. The version of the dataset used for the experiments does not include articles created after 2025-08-01.

[14] From UD 2.17. https://universaldependencies.org/treebanks/oc_ttb/index.html

[15] https://zenodo.org/records/12192029

[16] https://zenodo.org/records/8411197

[17] Cisalpine was excluded due to insufficient data.

[18] sic. https://zenodo.org/records/8411351

[19] https://huggingface.co/datasets/allenai/nllb. We use the Occitan samples with a LID score ≥ 0.8.

[20] https://zenodo.org/records/7079580

[21] https://huggingface.co/datasets/HuggingFaceFW/finewiki

[22] https://huggingface.co/datasets/HuggingFaceFW/fineweb-2

[23] https://opus.nlpl.eu/Tatoeba/oc&fr/v2023-04-12/Tatoeba (Tiedemann, 2012)

are normalised into a single one. Finally, we remove punctuation and numbers, and we turn the remaining text to lowercase.

4.3 Results and Discussion

Results in Table 3 show that our pretrained FastText vectors increased the overall macro-F1 by 4.48 points. Similarly, our BERT classifier fine-tuned from oc-mBERT performed better than the original mBERT-cased pretrained model, and also slightly better than the mBERT-uncased model (see Table 4).

While preprocessing is usually recommended to prevent overfitting on irrelevant features when training language identification models (Fedorova et al., 2025; Sousa et al., 2025), our experiments on domain-specific and multi-domain data resulted in lower or similar performance when using it, as shown in Table 5. The rest of our analysis will therefore focus only on models trained and evaluated on non-preprocessed datasets, and adapted from the pretrained FastText vectors and BERT model that resulted in the best performance scores.

FastText vectors	F1	Recall	Precision	Accuracy
CC	47.11	45.30	69.25	62.70
Ours	51.59	49.30	70.39	65.95

Table 3: Results of FastText OCDI models depending on the pretrained vectors used. Training set: CONCAT-train. Test set: CONCAT-test. Except for accuracy, scores are macro-averaged over dialect classes.

Pretrained model	F1	Recall	Precision	Accuracy
mBERT-cased	54.51	50.78	75.99	65.69
mBERT-uncased	57.89	53.66	77.64	67.89
oc-mBERT	58.15	54.27	78.49	66.61

Table 4: Performance depending on the pretrained BERT model, fine-tuned on CONCAT-train and tested on CONCAT-test. Except for accuracy, scores are macro-averaged over dialect classes.

Model type	w/ prep	w/o prep
SVM	53.47	55.45
FastText	51.54	51.59
BERT	54.84	58.15

Table 5: F1 scores macro-averaged, comparing approaches with and without text preprocessing, fine-tuned on CONCAT-train and tested on CONCAT-test.

Model type	F1	Recall	Precision	Accuracy	Time
Random baseline	11.32	12.50	12.50	12.50	-
SVM	55.45	52.12	72.32	**67.05**	3
FastText + our vectors	51.59	49.30	70.39	65.95	**0.01**
oc-mBERT	**58.15**	**54.27**	**78.49**	66.61	192

Table 6: Performance of models trained on CONCAT-train and evaluated on CONCAT-test, without preprocessing. Recall, precision and F1 scores are macro-averaged. Best score for each column is indicated in **bold**. Time corresponds to the average runtime per sample, expressed in milliseconds, measured on an 13th Gen Intel Core i7-1370P CPU. For oc-mBERT, time is measured on 1000 random samples from CONCAT-test instead of the full test set.

Performance scores in Table 6, obtained from the evaluation on CONCAT-test, report a macro-average F1 score of 11.32 for the random baseline, which is surpassed by all models trained on CONCAT-train. The best F1 score of 58.15 is achieved with the BERT model, which corresponds to an accuracy of 66.61. Performance using the SVM classifier is lower with 55.45 F1, and the lowest using FastText, with 51.59.

While these scores are low, we expect our evaluation methodology to underestimate the systems' performance, due to the projection of single labels from original documents onto the samples, which hides situations where a same text could be valid in more than one variety, especially for shorter texts.

A more in-depth analysis of results indicates performance disparities between dialects and models (see Table 7). Dialects that are under-represented in the training data (ARA, VIV, NIC) tend to have higher a precision (up to 91.76 for ARA) and a lower recall (only 19.51 for NIC) than bigger classes. As shown in Figure 4, this is due to the over-prediction of the larger classes, especially LAN, which gets a 91.86 recall, and only a precision of 59.10 with oc-mBERT. In usage cases where recall is preferred over precision (e.g. data mining), we recommend using the best-performing model in terms of recall per Occitan variety, rather than a single model across dialects: SVM for AUV, GAS, NIC, and VIV, FastText for GAS and LIM, and oc-mBERT for ARA, LAN and PRO.

The analysis of dialect-specific features ranked by the SVM weights (see Table 8) reveals that some numbers and punctuation are incorrectly overweighted, and that the model is biased with features related to names of languages (like 'aran' for class ARA). However, some important features of each

Model	ARA		AUV		GAS		LAN		LIM		NIC		PRO		VIV	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
Random baseline	1.87	12.50	7.31	12.50	18.88	12.50	26.53	12.50	12.75	12.50	9.39	12.50	17.09	12.50	6.18	12.50
SVM	85.29	18.83	62.30	**37.81**	76.49	**83.82**	60.22	89.03	**81.83**	70.69	88.16	**27.91**	**60.11**	67.49	64.12	**21.37**
FastText	86.36	12.34	66.21	32.17	75.61	**83.82**	**60.49**	90.40	79.06	**71.84**	86.38	23.77	57.18	66.15	51.82	13.92
oc-mBERT	**91.76**	**50.65**	**79.69**	33.83	**77.37**	82.15	59.10	**91.86**	76.50	66.60	**88.82**	19.51	58.22	**68.35**	**96.43**	21.18

Table 7: Precision (P) and recall (R) per dialect for each model trained on CONCAT-train and evaluated on CONCAT-test. Best score for each column is indicated in bold.

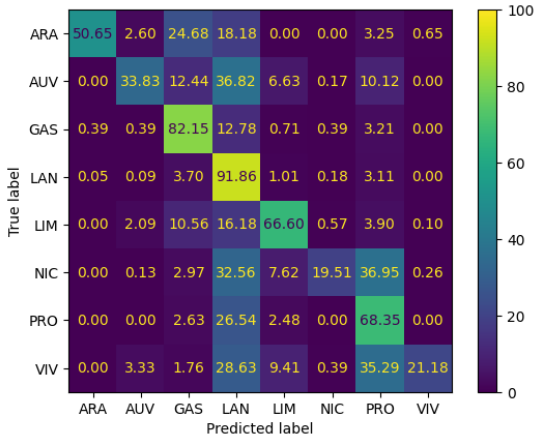
dialect stand out in the learned features, such as the lexical variant 'dab' (*with*) specific to GAS, the non-vocalised 'l' at word ending in LAN (as in 'sul' *on the*), the aphaeresis of [a] in LIM (e.g. 'quò' *this*), the muting of intervocalic [d] in VIV (as in 'poètz' *you can*), or the variant 'lu' for the masculine plural definite article in NIC. However, we notice that the character n-gram features 'ou' and 'ì' correspond to the so-called *mistralian* writing norm. While this norm seems to be used only for quotes and explanations in the Occitan Wikipedia, their over-weighted association to NIC is problematic as this writing norm is not restricted to only this Occitan variety; in fact it is also used frequently by PRO speakers.

Dialect	Features
ARA	aran determinants referéncies der comandar mès lesan galin telediario substantius
AUV	delh cercar pueis 3,45 z-es 982 dreita 26,43 favier 20,17
GAS	dens dab ei shens mei ua au deu dreta 2019
LAN	sense del sul ∅,∅ l∅ ambe ∅, aude ièr iè
LIM	maires daus quò queu 'na ente quo quel 833 quela
NIC	ou à∅ lu caracter adressa ligna statut recerca da ì∅
PRO	lei deis ò leis dau dei direccions #∅ premiera ∅1
VIV	internes recèrcha poètz aa navigaor pechon predefinia segua mas mesmes

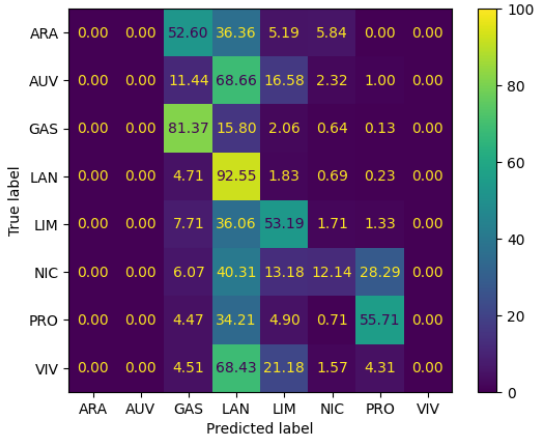
Table 8: Top 10 learned features per dialect for the SVM model without preprocessing. Word features are highlighted in blue, and character n-gram features in green. Word boundaries are indicated with ∅.

From the results of models trained on a single data source (cf. Table 9), the best macro-F1 scores are obtained by models trained and evaluated on only four classes, i.e. TTB and FORUM. Out of the two, all three models trained on FORUM exhibit larger performance gaps on TTB-test compared to the TTB models evaluated on FORUM-test, suggesting potential overfitting of the FORUM models, even if scores on TTB-test are within the range of scores with the same-source model. Models trained on CONGRES and SOFT (with similar sets of seven labels) display a low performance on the OCDI task in general, and a clear drop of performance out-of-domain on the test sets with seven or eight labels, with some scores that are close to the random baselines, or even lower for two out of the three models trained on SOFT and evaluated on WIKI. Possible reasons for these drops are the very specific and technical domain of SOFT, a large class imbalance in the training datasets, and for CONGRES, a very low number of samples apart from the LAN and GAS classes. Models trained on WIKI obtain the best results when evaluated in-domain, and often surpass performance of other single-source models in the out-of-domain evaluation settings. Scores on SOFT-test are similar to those obtained with the same-source models, and for CONGRES-test, WIKI models reached a better performance than the same-source ones. Training on the concatenated datasets results in higher scores across test sets, as the CONCAT models obtained the best scores on three test sets (CONCAT, CONGRES and SOFT), and second or third rank otherwise. Also, the macro F1 scores on CONCAT-test are higher for the CONCAT models than for the WIKI ones, showing that augmenting both the quantity of data and the diversity of domains has an impact on the OCDI task.

In order to assess the impact of our dataset OcWikiDialects on the OCDI task, we performed an ablation study comparing models trained on the full concatenation of available training data (CONCAT) with models trained on the same data excluding WIKI-train (CONCATNOWIKI). Evaluation was performed on the CONCAT-test set in both settings. As shown in Table 10, the inclusion of OcWikiDialects results in substantial performance gains, with improvements of up to 23.48 macro F1 points for the oc-mBERT model. This impact is particularly pronounced for the rarer classes, as shown when comparing the confusion matrices in Figure 4: while the oc-mBERT model trained with-



(a) CONCAT-train



(b) CONCATNOWIKI-train

Figure 4: Confusion matrices for the oc-mBERT models fine-tuned from CONCAT-train (top) or CONCATNOWIKI-train (down), without preprocessing, and evaluated on CONCAT-test.

Trainset and model \ Testset	CONCAT	WIKI	CONGRES	SOFT	TTB	FORUM
Random baseline	11.32	12.07	10.59	12.52	19.08	25.00
CONCAT-train						
SVM	55.45	56.15	42.74	42.33	60.57	68.98
FastText	51.59	51.27	39.27	39.45	62.42	67.85
oc-mBERT	**58.15**	57.72	**59.15**	**43.91**	56.79	70.22
WIKI-train						
SVM	53.61	**60.21**	44.81	31.99	44.16	57.61
FastText	49.23	49.86	43.01	34.71	57.67	66.06
oc-mBERT	48.12	53.18	50.45	31.67	45.99	52.37
CONGRES-train						
SVM	20.45	15.98	38.33	14.21	33.29	36.77
FastText	24.01	16.63	37.46	15.60	37.51	43.14
oc-mBERT	20.71	18.48	33.24	16.66	37.66	48.32
SOFT-train						
SVM	20.57	11.52	15.00	39.35	25.60	20.02
FastText	19.42	9.06	19.86	30.18	38.52	38.41
oc-mBERT	24.44	14.21	26.16	31.17	42.22	45.70
TTB-train						
SVM	20.92	17.56	28.27	17.48	66.66	47.36
FastText	26.00	19.71	32.86	25.23	**73.19**	68.85
oc-mBERT	15.48	11.23	22.73	14.20	46.38	36.36
FORUM-train						
SVM	29.11	26.90	36.36	22.21	53.82	88.93
FastText	32.66	28.91	41.63	31.35	71.38	**89.01**
oc-mBERT	29.22	26.11	37.29	22.14	63.95	86.62

Table 9: F1 scores on individual test sets for each model based on its training dataset and type. Best score for each test set is marked in bold. We warn the reader about the non-comparability of scores between most columns, as the number of labels differs between the test sets.

	$\overline{W}$	$C - \overline{W}$			
	F1	F1	Recall	Precision	Acc.
SVM	36.94	+18.51	+15.52	+16.46	+11.70
FastText	35.70	+15.89	+13.20	+16.63	+10.38
oc-mBERT	34.67	+23.48	+17.39	+17.92	+9.25

Table 10: Performance gaps between models trained on CONCAT-train (C) or on CONCATNOWIKI-train ($\overline{W}$), and evaluated on CONCAT-test. Except for accuracy (Acc.), scores are macro-averaged over dialect classes.

out OcWikiDialects never predicts the rarer classes ARA, AUV and VIV, incorporating the WIKI samples to the training leads to more frequent predictions of these labels, and results in higher recall across all classes except LAN, while also reducing over-prediction of the LAN label (i.e. improving precision).

5 Conclusion

In this paper, we introduce a new corpus OcWikiDialects, comprising over 7k articles (4M tokens) from the Occitan Wikipedia, segmented into paragraphs and sentences. In doing so, we contribute to increasing the amount of dialect-labelled Occitan data, necessary to increase the visibility of Occitan variation as a non-standard language in datasets for NLP. Metadata from revision histories and user pages included in the dataset enabled us to label articles with ten Occitan varieties and allowed us to carry out in-depth analyses of user-declared proficiency levels and bot contributions.

We used this dataset in combination with previously published dialect-labelled corpora to train Occitan dialect identification (OCDI) models in single-domain and multi-domain settings to discriminate between up to 8 dialects. Our approaches include n-gram based SVM classifiers, FastText classifiers based on pretrained vectors, and fine-tuned BERT models. Results favour the fine-tuned oc-mBERT model without text preprocessing, with a macro-average F1 score of 58.15 and fairly balanced scores across domains. However, performance differs between dialect classes, with low recall scores for less represented dialects, and low precision scores for the over-represented classes

which tend to be over-predicted, especially Languedocian. Future work should therefore prioritise the development of additional resources for underrepresented Occitan varieties.

We release models from all three main approaches (SVM, FastText, BERT), as they have different advantages, in particular interpretation of predictions, inference speed and overall accuracy, and we hope that our new dataset and OCDI models will foster more dialect-aware NLP research and applications for the Occitan language.

Limitations

The models presented in this paper are meant as baselines for future research on the topic of Occitan Dialect Identification, with several limitations that may restrict their suitability in certain settings, especially when a high confidence is required.

Beside a limited performance described in this paper, in particular for underrepresented Occitan varieties (Section 4.3), we did not train our models to predict a class 'other' when prompted with texts from other varieties, closely related to Occitan (e.g. Croissant transitional dialects) or not (e.g. French or Italian).

Furthermore, the datasets used in this study only contain texts in the classical spelling, except for the Niçard subset of SoftwaresOccitanTranslations which contains both samples in classical and mistralian norms. Therefore, using our models on texts written with the mistralian norm might lead to biased predictions towards the Niçard label.

Assessing performance in a multi-domain setting enhances the reliability of results, however the test samples have been selected automatically from the available datasets without removing irrelevant samples such as bibliography entries with titles in languages other than Occitan. Manual annotation of the test samples by Occitan specialists could be particularly relevant to improve the quality of test sets, and also to move towards a multi-label classification task where the samples might be valid in multiple varieties.

Acknowlegments

The authors would like to thank Zachary Hopton for releasing the oc-mBERT model upon our request, the reviewers for their valuable feedback, and the CLEPS infrastructure from Inria Paris for providing computational resources. This work was partly funded by Rachel Bawden and Benoît Sagot's chairs in the PRAIRIE institute, funded by the French national agency ANR, as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by Benoît Sagot's chair in its follow-up, PRAIRIE-PSAI, also funded by the ANR as part of the "France 2030" strategy under the reference ANR23-IACL-0008.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945,

Nusa Dua, Bali. Association for Computational Linguistics.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. Publisher: Nature Publishing Group.

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.

Louise Esher and Jean Sibille, editors. 2024. *Manuel de linguistique occitane*. De Gruyter.

Mariia Fedorova, Jonas Sebulon Frydenberg, Victoria Handford, Victoria Ovedie Chruickshank Langø, Solveig Helene Willoch, Marthe Løken Midtgaard, Yves Scherrer, Petter Mæhlum, and David Samuel. 2025. Multi-label Scandinavian language identification (SLIDE). In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 179–189, Tallinn, Estonia. University of Tartu Library, Estonia.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Kevin Zhu, Austen Liao, and Sean O'Brien. 2025. EnDive: A cross-dialect benchmark for fairness and performance in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16830–16855, Suzhou, China. Association for Computational Linguistics.

Zachary Hopton and Noëmi Aepli. 2024. Modeling Orthographic Variation in Occitan's Dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 78–88, Mexico City, Mexico. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas 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, Valencia, Spain. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Louisa Lambrecht, Felix Schneider, and Alexander Waibel. 2022. Machine translation from Standard German to alemannic dialects. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 129–136, Marseille, France. European Language Resources Association.

Lo Congrès. 2024. ReVoc Corpus.

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 Miletic and Yves Scherrer. 2022. OcWikiDisc: a corpus of Wikipedia talk pages in Occitan. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–79, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Oriane Nédey, Juliette Janès, Rachel Bawden, Thibault Clérice, and Benoît Sagot. 2025. ForumOccitania: a Corpus of User-Generated Content for Multiple Occitan Varieties.

Ebuka Okpala and Long Cheng. 2025. Large Language Model Annotation Bias in Hate Speech Detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1389–1418.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine Learning in Python. *arXiv preprint*. ArXiv:1201.0490 [cs].

Guilherme Penedo. 2025. FineWiki.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language.

David Pomerenke, Jonas Nothnagel, and Simon Ostermann. 2025. The AI Language Proficiency Monitor – Tracking the Progress of LLMs on Multilingual Benchmarks. *arXiv preprint*. ArXiv:2507.08538 [cs] version: 1.

Clamenca Poujade, Myriam Bras, and Assaf Urieli. 2024. CorpusArièja: Building an annotated corpus with variation in Occitan. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 66–71, Torino, Italia. ELRA and ICCL.

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, Online. Association for Computational Linguistics.

Eve Seguier. 2015. Reconnaissance automatique des dialectes occitans à l'écrit. Master's thesis, Université Toulouse Jean Jaurès, Toulouse, France.

Jean Sibille. 2024. 16 Les dialectes occitans. In Louise Esher and Jean Sibille, editors, *Manuel de linguistique occitane*, pages 423–471. De Gruyter.

Hugo Sousa, Rúben Almeida, Purificação Silvano, Inês Cantante, Ricardo Campos, and Alípio Jorge. 2025. Enhancing portuguese variety identification with cross-domain approaches. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, volume 39 of *AAAI'25/IAAI'25/EAAI'25*, pages 25192–25200. AAAI Press.

Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Verrah Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, and 78 others. 2026. CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data. *arXiv preprint*. ArXiv:2601.18026 [cs].

Aure Séguier and Lo Congrès. 2023a. Occitan Corpus from Lo Congrès news.

Aure Séguier and Lo Congrès. 2023b. SoftwaresOccitanTranslations corpus.

Aure Séguier and Lo Congrès. 2024. Lo Congrès websites Corpus.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Scale of Occitan Proficiency Levels

The Occitan language proficiency levels used in OcWikiDialects correspond to the Babel scale defined by Wikimedia Commons.[24], with the following definitions to which we add level 'None':

- 'None' means that no Occitan level was found in the user page
- 0 indicates someone who does not understand the language.
- 1 stands for basic knowledge: the ability to understand and answer simple questions in the language.
- 2 stands for intermediate knowledge.
- 3 stands for advanced or fluent knowledge: the ability to correct spelling and grammar errors in the language.
- 4 stands for near-native ability.
- 5 stands for professional proficiency.
- N stands for native language.

[24] https://commons.wikimedia.org/wiki/Commons:Babel Consulted on 2025-12-30.

B Additional Statistics from Dataset OcWikiDialects

B.1 Dialects Distribution

Figure 5 shows the distribution of dialect tags declared in articles and in user pages.

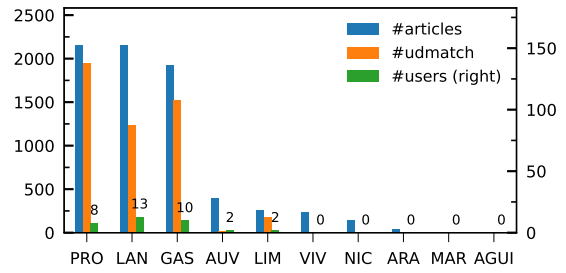


Figure 5: Distribution of article dialect tags (#articles) and user-declared dialects (#users, only active users) in OcWikiDialects, and number of articles per dialect for which at least one contribution was made by a user of the same dialect (#udmatch).

B.2 Bot Contributions

Regarding the role of the 135 active users detected as bots, they have been used to create 13% of the articles in the dataset and 41% of the total contributions. However, this ratio drops to 14% when considering only contributions with at least +100 bytes (based on the difference before and after), and to 8% for the first non-empty contributions[25] authored by a bot.

C Excerpts from OcWikiDialects

Table 11 shows a few excerpts from the dataset OcWikiDialects with their translations into English.

D Hyperparameters

For SVM models, the impact of class imbalance is reduced by setting the `class_weight` option to "balanced". Features occurring in only one sample (`min_df`) or in more than 90% of the training data (`max_df`) are ignored.

For FastText, our embedding skipgram model has the same dimension of 300 than the CC vectors, and is trained for 5 epochs with a learning rate of 0.05. The classification layer is trained for 20 epochs with a hierarchical softmax loss and a learning rate of 0.1. In preliminary experiments we found that using word unigrams for the OCDI task results in performance that is better than or similar to using word bigrams.

For BERT models, we use a learning rate of 1e-5 with linear scheduling and AdamW optimiser, an effective batch size of 64. Each model is trained for 50 epochs with 0.2% of the total steps (i.e. 1% of the training set) used for warmup and early stopping with a patience of 10.

[25] For each article, we consider the earliest revision with at least 100 bytes.

LAN	Menèrba (Minerve en francés) es una comuna lengadociana situada dins lo departament d'Erau e la region d'Occitània, ancianament de Lengadòc-Rosselhon. *Menèrba (Minerve in French) is a Languedocian municipality located in the Hérault departement and the Occitania region, formerly Languedoc-Roussillon.*
PRO	Barcilona o benlèu Barciloneta (Barcelonnette en francés) es una comuna provençala situada dins lo departament deis Aups d'Auta Provença e la region de Provença-Aups-Còsta d'Azur. *Barcilona or maybe Barciloneta (Barcelonnette in French) is a Provence municipality located in the Alps of Upper Provence departement and Provence-Alpes-Côte d'Azur region.*
NIC	Lo Puget Tenier (Lou Puget-Teniés en nòrma mistralenca; Puget-Théniers en francés) es una comuna d'Occitània, dins lo País Niçard e lo departament dei Aups Maritims. *Lo Puget Tenier (Lou Puget-Teniés in the mistralian norm; Puget-Théniers in French) is a municipality of Occitania, in the County of Nice and the departement of Maritime Alps.*
VIV	Chastelnòu de Bordeta (Châteauneuf-de-Bordette en francés) es una comuna occitana de Daufinat situaa dins lo departament de Droma e la region d'Auvèrnhe-Ròse-Aups, ancianament de Ròse-Aups. *Chastelnòu de Bordeta (Châteauneuf-de-Bordette in French) is an Occitan municipality of Dauphiny located in the Drôme departement and the Auvergne-Rhône-Alpes region, formerly Rhône-Alpes.*
AUV	Mauriac (Mauriac en francés) z-es 'na comuna d'Auvèrnhe ; z-es administrada pel departament delh Chantal de la region d'Auvèrnhe-Ròse-Aups, ancianament d'Auvèrnhe. *Mauriac (Mauriac in French) is an Auvergnat municipality ; it is administrated by the Cantal departement of Auvergne-Rhône-Alpes region, formerly Auvergne.*
LIM	Lo territòri de la comuna de Sent Deunis fuguet 'bitat desjà la Preïstòria. Ne'n tesmonha lo dolmen situat entre lo vilatge de la Valada e la D 979, qui fuguet fortament 'bimat en 1862 e completament destruch en 1902. *The territory of Sent Deunis municipality was already inhabited during Prehistory. This is testified by the dolmen located between the village of La Valada and the D 979 [road], which was strongly damaged in 1862 and completely destroyed in 1902.*
GAS	Biriato (Biriatu en basco, Biriatou en francés) qu'ei ua comuna de la província tradicionau de Labord, administrada peu departament deus Pirenèus Atlantics dens la region de Navèra Aquitània, ancianament d'Aquitània. *Biriato (Biriatu in Basque, Biriatou in French) is a municipality of the traditional province of Labord, administrated by the Pyrénées Atlantiques department in New Aquitaine region, formerly Aquitaine.*
ARA	Les (en catalan: Lés) ei ua vila e municipi dera Val d'Aran, en eth terçon de Quate Lòcs, ath marge deth riu Garona. *Les (in Catalan: Lés) is a city and municipality of the Val d'Aran, in the district of Quate Lòcs, along the Garonne river.*

Table 11: Excerpts from OcWikiDialects with their translation into English.