ForumOccitania: a Corpus of User-Generated Content for Multiple Occitan Varieties

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

Abstract

We introduce ForumOccitania, a new Occitan corpus of posts from an online forum, covering a range of topics and dialects. While some existing datasets for this low-resource language include labels of varieties within the dialect continuum, we go one step further by providing metadata pertaining to sociolinguistic factors of language variation (dialect, geographical location, age, proficiency), extracted from self-declared user profiles. We carry out statistical and qualitative analyses, as well as preliminary experiments on unsupervised dialect identification. Our results show that (i) most of the contents is written in Occitan, with the classical spelling conventions, and by young speakers, (ii) posts display a strong presence of dialectal features from four major Occitan varieties (*Lemosin*, *Lengadocian*, *Gascon*, *Provençau*), and (iii) a simple topic modelling approach introduced by Kuparinen and Scherrer (2024) effectively detects salient features of these four varieties, but also reveals finer-grained diatopical variation tendencies.

Keywords: corpus, Occitan, low-resource language, dialect variation, language identification, user-generated content

1. Introduction

As increasing number of languages are supported by NLP systems (Costa-jussà et al., 2024; Kudugunta et al., 2023), research is also turning towards the processing of closely related languages, including non-standardised language varieties (Joshi et al., 2025). The term *dialect continuum* is attributed to some of these languages that are characterised by important internal variation at a diatopical (i.e. geographical) level (Chambers and Trudgill, 1998).[1] Current NLP systems tend to lack robustness to dialectal variation (Faisal et al., 2024), and tasks like variety identification are of particular importance for the development of tools supporting these languages (Scherrer et al., 2025). In this paper, we focus on Occitan, a dialect continuum that extends through a large area in the south of France as well as some neighbouring valleys of Spain and Italy. While linguistic variation operates rather continuously at the lowest geographical level, linguistic studies have shown the presence of larger groupings of localities through the use of isoglosses, leading to precise delimitations of varieties within this continuum, as in Figure 1.

Previous work on Occitan has led to the creation of several datasets and tools for NLP (Miletic and Scherrer, 2022; Vergez-Couret et al., 2024; Hopton and Aepli, 2024; Morcillo et al., 2024). While some corpora include dialect annotations (Miletić et al., 2020), they are characterised by a large class imbalance, and only use labels corresponding to coarse-grained dialectal groupings (except for three very specific, localised varieties), thereby limiting the possibilities of research to tackle challenges related to the Occitan linguistic continuum.

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

The main contribution of this paper is the introduction of ForumOccitania,[2] a novel Occitan corpus annotated with dialect labels and additional metadata, with the purpose of enabling both corpus-based linguistic studies and dialectal NLP research to account for multiple sociolinguistic dimensions in projects with Occitan data. The corpus is com-

[1] While we are aware that the distinction between *language* and *dialect* is mostly a political one, for clarity reasons, we will preferably use the word *language* to refer to the largest linguistic entities (French, Occitan), and the words *dialect* or *variety* (interchangeably) to refer to regional or local linguistic entities within a dialectal continuum.

[2] Source code associated to the paper: https://github.com/DEFI-COLaF/forum-occitania

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3220–3233
11–16 May 2026.

posed of threads from an online discussion forum,[3] active between 2007 and 2012, covering multiple themes (news, music, linguistics, food, technology, books, etc.) and mainly involving young speakers of Occitan from diverse places in the Occitan-speaking area (and beyond). Through a mapping to the user profiles, we include post-level annotations of dialect, age, Occitan proficiency, and geographical location. Moreover, as multiple languages are present in the forum (including code-switching), we provide language predictions. In order to preserve user privacy in the final dataset, we apply an anonymisation pipeline to both user metadata and post content.

We evaluate the dataset quality both with a manual analysis focusing on the presence of dialectal features and with preliminary experiments on unsupervised dialect identification. They show that it is possible to effectively detect some salient features of a few Occitan varieties using a simple topic modelling approach, and that the dialect subsets as declared by users closely match the computed dialect clusters.

2. Related Work

The Occitan language is a dialect continuum with a long history and various denominations. Besides unspecified denominations such as *patois*, or regional names coined for political reasons like *Provençau* (Lieutard, 2024), linguists have proposed several internal classifications over time. The most commonly used diatopic classification distinguishes between six broad areas of the continuum, with the *Gascon* (GAS), *Lengadocian* (LAN), *Provençau* (PRO), *Lemosin* (LIM), *Auvernhat* (AUV) and *Vivaroaupenc* (VIV) varieties.[4]

With a long-standing oral and written tradition dating back to the Middle Ages, the language has undergone profound changes over the past century, creating large differences in usage between older native speakers and newer generations who acquired it outside the family (Brennan, 2024). Its standardisation is an ongoing process, currently evolving towards pluricentric standards (Lamuela, 2024). In the absence of an officially recognised standard outside Spain, multiple spelling systems coexist, sometimes even for the same Occitan variety, with varying degrees of divergence. They are mainly based on works from the late 19th and early 20th centuries that have led to the two major spelling conventions used in France today (and at the time when the forum was active): the *mistralian* norm, which places greater emphasis on pronunciation and shows influence from French orthographic conventions, and the more widely adopted *classical* norm, grounded in medieval Occitan orthography and designed to reduce dialectal variation in writing.

2.1. Datasets for Occitan

Various projects have helped increase the presence of contemporary Occitan in NLP datasets (Miletić et al., 2020; Penedo et al., 2025). However, the amount of data is still low, and the geographical coverage of the continuum is biased towards the *Lengadocian* variety.

Most of the Occitan data available for NLP purposes is included in very large, multilingual, automatically mined datasets, such as NLLB (Costa-jussà et al., 2024) and FineWeb-2 (Penedo et al., 2025). However, the quality of these datasets is questionable as they contain many samples erroneously labelled as Occitan.

Other significantly large corpora have their origins in Wikipedia: Wikipedia Monthly (Omar Kamali, 2025) contains dumps of plain text articles, WikiMatrix (Schwenk et al., 2019) is the product of automatically aligned parallel translations, OcWikiDisc (Miletic and Scherrer, 2022) consists of talk pages messages with predictions from language identification models–and is therefore the first dataset of user-generated content (UGC) in Occitan, before ours–, and OcWikiAnnot (Miletić, 2023) contains sentences with automatically predicted part-of-speech tagging and lemmatisation annotations.

However, few Occitan-specific research projects include dialect information in their datasets. Tolosa Treebank (Miletić et al., 2020) includes train and test splits for each of four varieties, which were used to discover dialectal biases in morphosyntactic tagging experiments (Miletić, 2023). Poujade et al. (2024) provide annotations on both dialect (two varieties) and spelling system used in each document, on top of word-level morphosyntactic annotations. Various datasets with dialect labels were also published by the Occitan organisation Lo Congrès. *Occitan Corpus from Lo Congrès news* (Séguier and Lo Congrès, 2023a) contains bilingual data from the news section of the organisation, with sentence-level labels for six varieties. SoftwaresOccitanTranslations [sic.] (Séguier and Lo Congrès, 2023b) contains translations of many open source tools into Occitan, with sentence-level labels for the six major varieties as well as *Niçard* (NIC), a localised variety that is classified by Sibille (2024) as *Provençau* with some characteristics of *Vivaroaupenc*. *Lo Congrès websites Corpus* (Séguier and Lo Congrès, 2024) contains bilingual data from all websites belonging to the organisation, with sentence-level labels for the six major varieties as well as *Aranés* (the variety spoken in Spain, linguis-

[3] https://occitania.forumactif.com/

[4] We use the endonym of each variety to reflect the designations used on the forum and extend this naming convention to additional varieties not available in the forum's profile options.

tically related to *Gascon*) and *Cisaupenc* (spoken in the north of Italy).

The distribution of dialect labels in existing datasets is largely imbalanced, with an overrepresentation of *Lengadocian*, while most other varieties are either very scarcely resourced or almost (if not completely) absent (e.g. *Cisaupenc*). Our contribution addresses this gap as the FORUMOCCITANIA corpus contains both a high amount of self-declared *Lemosin* and *Provençau* data (largely under-represented varieties in the existing datasets), and a seemly balanced distribution between the represented varieties, according to our predictions in Section 6. Also, annotations in existing datasets pertaining to the internal variation within the Occitan language are mostly limited to the dimension of dialect variation, whereas other linguistic (e.g. spelling system) and sociological (e.g. age, geolocation) factors could help conduct more in-depth computational linguistics analyses, but also build NLP applications that are more representative of the diversity of Occitan linguistic usages.

2.2. Anonymising User-Generated Content

Anonymisation of unstructured data typically involves de-identifying personally identifiable information (PII) (Pilán et al., 2022). The detection of direct or quasi-identifiers often combines manual and automated techniques, such as regular expressions and named-entity recognition (NER) models (Adams et al., 2019). In defining which entities to anonymise, several works on UGC datasets (Adams et al., 2019; Çetinoğlu and Schweitzer, 2022; Riabi et al., 2024) distinguish between the names of famous and non-famous people, and choose to leave famous people non-anonymised. Once detected, entities are anonymised using labelled placeholders, or category-consistent replacements (e.g. substituting one name for another) (Çetinoğlu and Schweitzer, 2022). Pilán et al. (2022) and Riabi et al. (2024) highlight the importance of risk assessment, in order to choose the right balance between user privacy and data utility, usually measured with recall and precision metrics.

2.3. Dialectometry

There is an important tradition of dialectology and dialectometry research for Occitan, with studies based on linguistic atlases (Léonard et al., 2024). However, to the best of our knowledge, our corpus-based dialectometry experiments–while not aimed at a detailed linguistic analysis– are the first of the kind to be conducted on Occitan.

Among previous corpus-based approaches, Hovy and Purschke (2018) propose to train Doc2Vec embeddings on online posts in the German-speaking area, before applying a combination of clustering and dimensionality reduction techniques to obtain dialect clusters. Kuparinen and Scherrer (2023) extract clusters of dialectal Finnish and Norwegian by applying principal component analysis (PCA) to speaker ID embeddings that are learned jointly with a dialect-to-standard neural machine translation model.

Our experiments, however, are based on the clustering method described in (Kuparinen and Scherrer, 2024), who train topic models using phonetically transcribed interviews of dialectal Finnish, Norwegian and Swiss German.

3. Dataset Creation

3.1. Data Collection

We scraped the forum website to collect all posts' full contents, the forum structure and user information. On the forum, each post is displayed next to a user block, containing a summary of their profile. Users can voluntarily declare information based on the available fields, all of which are optional. We extract the following metadata: age (at the time of parsing, as computed by the website), geographical location (free text field), dialect (from the options *Gascon*, *Lemosin*, *Lengadocian*, *Provençau*, *Vivarés*, or *Auvernhat*), and Occitan proficiency level.[5] The forum is structured into categories, subcategories and topics. For each post, we extract the contents in HTML format, the position of the post in the topic, the parent category and subcategory, as well as the date and time of publication.

3.2. Derived Metadata

Based on the parsed metadata, we normalise the self-declared age (grouped into buckets to reduce the risk of identification) and geographical locations, additionally completing values corresponding to the country, administrative region (before and after the administrative fusions of 2016), cultural region,[6] French *département* and locality.[7]

We also provide estimated metadata: predictions of (i) languages and (ii) Occitan varieties used within the dataset.

For language identification (LID), we use FastText (Joulin et al., 2017)[8] and the model from (NLLB Team et al., 2022) to predict the main language(s)

[5] See options in Appendix A.

[6] An area below the administrative regional level, defined by local culture, geography, or history.

[7] We only share localities with over 30k inhabitants.

[8] https://huggingface.co/facebook/fasttext-language-identification

used by each user, and present in each post, restricting predictions to those found in a predefined list of plausible languages: Occitan, French, Catalan, English and Spanish. LID is applied to all posts of five words or more, after removal of HTML markup, quoted contents, URLs and emails. For each post, we predict a maximum of two languages, always selecting Occitan if it is in the top five model predictions and only selecting another language if it is the top prediction. For each user, we process all posts individually and select a maximum of three most frequently predicted languages, associated with their relative frequency. For posts with two predicted languages, the second one counts for half in the label frequencies.

For Occitan dialect prediction, we use the results of our clustering and dialect identification experiments presented in Section 6 for each post associated with a high Occitan LID score.

3.3. De-identification

We carry out de-identification to reduce the risk of disseminating direct and quasi-identifiers (the number of online active Occitan speakers around 2010 was probably low), especially for its use in downstream NLP applications.[9] Our de-identification pipeline detects PII entities (usernames, names (except public figures), phone numbers, email addresses and postal addresses) in each post, and replaces all occurrences with category-specific placeholders.

We use a combination of regular expressions and a GLiNER model[10] (Zaratiana et al., 2024) specialised in the detection of PIIs. We first use regular expressions and word lists to detect usernames (from all usernames on the forum), email addresses, phone numbers, and names (given a list of French and Occitan first names that can be extended to the right with a capitalised last name). We then apply GLiNER to truecased[11] texts to detect postal addresses (validated with a regular expression) as well as additional person names. Longer posts are split into smaller chunks with a sliding window before being passed to the model.

Detected names (from regular expressions and GLiNER) are filtered in three steps to reduce false positives. First, predicted entities shorter than three characters are discarded. Next, a lexicon is used to discard entities containing only words that are not proper nouns. We build the lexicon by concatenating all words except for proper nouns, punctuation and numbers from Lofl`oc (Vergez-Couret et al., 2024) and Universal Dependencies (Zeman, 2025) corpora in Occitan, French, Spanish and Catalan.[12] Third, the Wikipedia API[13] is used to identify and discard public figures who should not be anonymised.

We evaluate the pipeline by manually annotating 200 posts, extracted randomly with weights based on the inverse number of posts by each user. Inspired by the recommendations of Pilán et al. (2022), we compute the precision, recall and F2-score on exact matches,[14] by PII category, as well as the macro-averages for each metric.

Results presented in Table 1 show a macro-average F2-score of 75.76. Scores are high or even perfect across all categories except names, where only half of the annotated entities were detected by our pipeline, and the number of wrong predictions remains very high despite the post-filtering steps. These errors include a high number of public figures, but also common Occitan words with several variants, which we suspect is related to the limited coverage of Occitan in GLiNER.

PII	#Gold	#Err	P	R	F2
address	3	0	100.00	66.67	71.43
email	4	0	100.00	100.00	100.00
name	30	100	12.71	50.00	31.51
phone	1	0	100.00	100.00	100.00
username	30	3	88.00	80.00	75.86
Totals	68	103	80.14	78.00	75.76

Table 1: PII detection result based on 200 manually annotated samples. The totals row indicates the sum for columns with absolute values, and average for columns with relative values. *#Err* is the number of spans that do not exist in the golden annotation.

3.4. Dataset Versions and Splits

The dataset exists in two formats: TEI and JSONL. The TEI dataset version contains detailed information about each participant, reproduces the forum structure (main categories, subcategories, topics and posts in their original order), and is faithful to the original HTML contents of the posts. We defined TEI guidelines[15] to encode the web forum and

[9] Since the forum remains online, the dataset cannot be fully anonymised, as a web search of an excerpt may reveal the original content.

[10] https://huggingface.co/urchade/gliner_multi_pii-v1

[11] We train a custom model using the Moses truecaser (Koehn et al., 2007) on the concatenation of the following datasets: ForumOccitania, LoCongresNews, OcWikiDisc and part of NLLB fr-oc (LID > 0.8).

[12] UD v2.17. Occitan: Tolosa Treebank. French: Sequoia. Spanish: AnCora (dev+test). Catalan: AnCora (dev+test)

[13] Specifically the Occitan, French and English versions.

[14] Using Python package nervaluate (v1.1.0)

[15] https://defi-colaf.github.io/metadata/TEI/ODD.html

the fine-grained, non-standard language varieties (Janès et al., 2025). We use Glottolog identifiers (Nordhoff and Hammarström, 2011) to declare languages used throughout the document, including references to specific varieties of Occitan. Converting the original HTML posts to TEI involves steps such as paragraph detection from empty lines, list detection via regular expressions (since forum lists rarely use HTML markup), preservation of image/emoji/video/audio references, and simplification of style formatting. Also, quoted replies are processed in order to retrieve the original post being quoted.

The JSONL version is meant for use in NLP applications. We parse the TEI version to extract the raw text from posts, except for quotes, along with user and LID metadata. Line breaks are added in the raw texts where indicated in the TEI, paragraphs are separated by an empty line, and emoji images are converted to a textual representation consisting of their *alt* description enclosed in double colons (e.g. `::plan urós::` 'very happy')

We also provide train, development, and test splits for building and evaluating downstream, dialect-aware NLP systems. The splits are based only on samples with a high Occitan LID score (above 0.9) and a dialect label, except for *Vivarés* and *Auvernhat* labels that are incorrect as shown in Section 5. The test and development sets contain 1000 samples each, leaving 7327 samples for the training set. Our splitting strategy maximises divergence between each subset by minimising shared authors and forum categories, and ensures a balanced dialect label distribution in the test set.

4. Dataset Statistics

The dataset contains 20,005 posts (approximately 1.15M tokens), published between 2007 and 2012. They are organised in 14 main forum categories, 25 subcategories,[16] and 2,583 topics (discussion threads). The names of forum categories relate to various themes: news and media, music, linguistics, education, computers and technical support, events, books, sports, movies, tourism, food, jobs. The category with the most data (*Estanquet* 'bar') is not thematic and entails a subcategory where many users introduced themselves.

Most of the posts are short between 12 and 59 tokens. However there are some exceptions, with posts up to 6,372 tokens, and 393 posts with 0 tokens (they originally contained only a combination of images and quoted posts).

Among the 292 extracted users, 66% provided information in at least one profile field (dialect, Occitan language level, age, geographical location), and 37% completed all fields (see Figure 2).

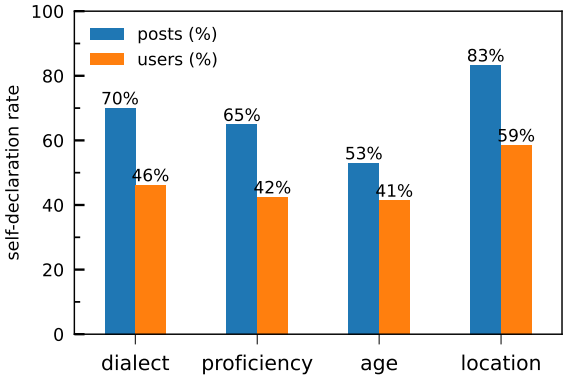

Figure 2: Proportion of declared profile fields across users and posts.

Age and Proficiency Age declarations show a majority of young users between 20 and 35 years old,[17] although a few older users were also active. As shown by Figure 3, users who declared a higher language level tend to be older. However, the oldest users only declared a low to medium level of Occitan.

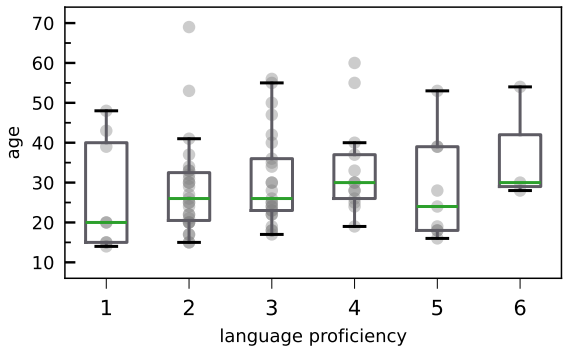

Figure 3: Distribution of age grouped by language proficiency level, converted to numerical levels from 1 (low) to 6 (high). See Appendix A for corresponding textual labels. Green lines represent the median age of users for each language level.

Dialects The self-declared language varieties encompass all six major families of Occitan (see Figure 4). *Lemosin* is the most represented variety in terms of numbers of posts, with a small number of users contributing a large volume of posts. *Gascon* and *Lengadocian* are also well represented, with *Lengadocian* represented by the largest number of users. The forum also includes a non-negligible

[16] Some posts belong directly to a main category, without a subcategory.

[17] We compute age statistics based on the age of all users in 2009, i.e. the median year of post publication.

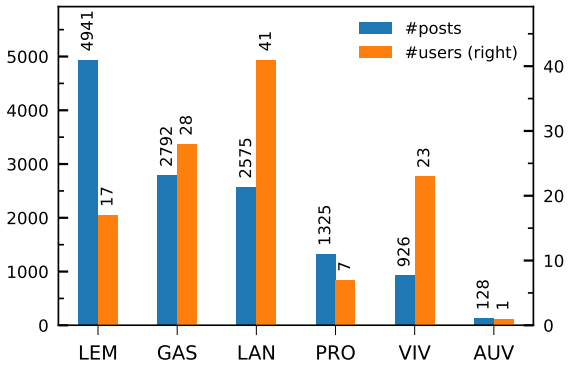

Figure 4: Users and posts per declared dialect.

number of posts for the *Provençau* and *Vivarés* dialects. Finally, the *Auvernhat* dialect, which is very rare in other Occitan datasets, is represented by a single, prolific user.

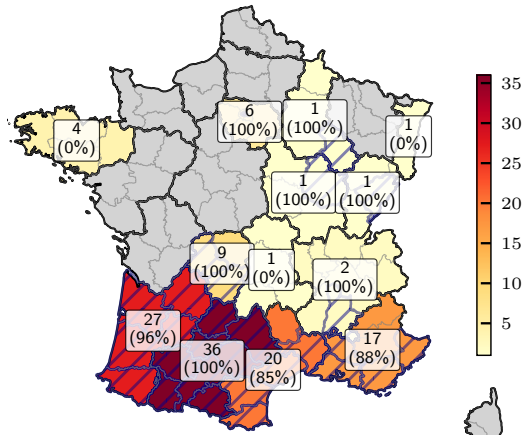

Figure 5: Geographical distribution of users. Region labels show the number of users who declared a location. Percentages in parentheses indicate the share of declared locations matched both to the region and to a *département*. Hatchings mark *départements* represented by at least one user.

Geographical Locations Visualisations of the users' normalised geographical locations (see Figure 5) reveal that the data of ForumOccitania covers a large portion of the dialectal continuum area, including Spanish Val d'Aran (one user, among a few others in Spanish Catalonia). All French administrative regions belonging to the continuum are represented in the dataset with at least one user, however the distribution of users places most of them in the western regions – historically associated to the *Gascon* and *Lengadocian* varieties –, while the northern regions – especially those corresponding to the lower-resourced *Auvernhat* and *Vivaroaupenc* varieties – are represented by very few users. When looking into more geographical details, we notice that the coverage decreases at the *département*-level, leaving most of the northern and eastern *départements* unrepresented.

On par with our observations of the distribution of posts per users, the geographical distribution of posts reveals a striking imbalance, with most of the data located near three cities: Limoges (*Lemosin* variety), Bordeaux (*Gascon* variety) and Toulouse (*Lengadocian* variety).

Finally, some users declared locations outside of the historical Occitan dialect continuum, both elsewhere in France and beyond.

LID Predictions reveal that most of the dataset is in Occitan. 78% of users and 89% of posts are associated with Occitan, which is the top prediction in most cases (84% of posts), or as second prediction. 11% (2233) of posts have no LID label (corresponding to empty or very short posts, or because of predictions outside of our predefined list), and other languages are also present: French (4% or 788 posts), Catalan (141 posts), English (23 posts) and Spanish (20 posts).

5. Qualitative Analysis

All the posts we observed use the *classical* spelling system. The few references we found of the *mistralian* spelling were negative opinions about it.

General as well as detailed observations of the posts organised by subsets per dialect led to the conclusion that most of the posts with labels *Gascon*, *Lemosin*, *Lengadocian* and *Provençau* present a high number of distinctive dialectal features corresponding to their label (see Table 2 for a few examples). For the *Provençau* subset, some of the features we observed correspond more specifically to the *Niçard* variety. However, the subset of posts labelled as *Auvernhat* features multiple *Gascon* characteristics, and we did not identify any *Vivaroaupenc* characteristic in the *Vivarés* subset, which seems to contain only posts written in *Lengadocian*.[18]

We noticed frequent co-occurrences of multiple languages–mostly between Occitan and French. This corresponded to some cases of code-switching, but seemingly more often to quotes of external sources of contents such as inline citations or long portions of documents, or when providing bilingual announcements for local events.

[18]This absence may be related to a bug following a website update that reportedly reassigned dialect labels in some user profiles, according to users in a 2009 forum discussion.

Dialect	Selection of dialect features	Sample excerpt and translation
Lengadocian	*amb* “with” *dels* plural indefinite article	Apreni l’occitan a Montpelhièr sol **amb dels** livres e un pauc **amb dels** cors a l’Universitat Paul Valery (sens èsser estudiant en occitan). *I’m learning Occitan in Montpellier, by myself with books and a little with lessons at University Paul Valery [in Toulouse] (without being a student of Occitan studies).*
Gascon	*ua* “a”/“one” (feminine) *har* “make” *perpausar* “propose”	Un o **dus** còps lo mes, vos **perpausi ua** fotò o un text, e vos mandi de vos en inspirar per **har** un **tribalh** d’invencion. *Once or twice per month, I propose to you a photo or a text, and I ask you to draw inspiration from it to make a work of invention.*
Lemosin	*’quò* “that” *charjar* “load” *pita* “small”	[S]e **um** ajosta a **’quo-’qui** dessubre que **’na** pagina MySpace quand **um** n’ai que **’na pita** conexion internet, ben **’quo** te planta l’ordinator e **’quo charja** pendent 3 o 4 minutas... *If one adds to this, moreover, that a MySpace page, when one only has a weak Internet connection, well that freezes your computer and it loads for 3 or 4 minutes...*
Provençau	*lei* “the” (plural) *emé* “with” *dau* “of the”	Per **lei** que conoisson l’istòria **dau** hip hop, savètz qu’en 1989 un disco sortèt **emé** dessus la fina flor **dau** rap new yorkés *For those who know the history of hip hop, do you know that in 1989 a record came out with the finest selection of New York rap[?]*

Table 2: Examples of dialectal features found in the ForumOccitania dataset.

6. Clustering Experiments

We carry out an initial series of experiments to assess the potential of the ForumOccitania dataset for both linguistic analysis and NLP applications, with a particular focus on Occitan dialect identification. The dataset contains dialect annotations, though these are not independently verified, and it also provides geographical information. To evaluate the usability of these labels, we employ an unsupervised approach, analysing the learned features and comparing them with linguistic descriptions, user-declared dialects, and (normalised) locations. We leave supervised dialect classification to other works, including the recently published Nédey et al. (2026) who compare performance across three types of supervised approaches and several annotated datasets, including ForumOccitania. In this paper, we adapt the topic modelling approach proposed by Kuparinen and Scherrer (2024). Our experiments examine whether dialect clusters in the dataset align with known descriptions of Occitan variation, how precisely they capture distinctions along the continuum, and what topic distributions on labelled and unlabelled samples reveal about dialect distribution.

6.1. Methodology

Modelling Approach We use non-negative matrix factorisation (NMF) (Paatero and Tapper, 1994) to perform topic modelling on Occitan datasets with features including full words and character n-grams, vectorised using TF-IDF. We explore results when varying the number of topics between 4 and 15.

We preprocess all samples by lowercasing, removing accents, substituting repeated characters with a single one (when there are three or more, typical to UGC style), as well as removing punctuation, URLs, email addresses, anonymisation-related placeholders and words corresponding to language names from a predefined list (e.g. “occitan”, “lengadocian”) that could lead to irrelevant topics as well as biased predictions, as mentioned in (Sousa et al., 2025).

Data We train NMF models on ForumOccitania non-anonymised samples, with and without dialect labels, filtered by Occitan LID score ($\geq$0.9) and by number of tokens ($\geq$30 tokens). For evaluation, we take samples filtered by LID score only. To calculate the performance, we use the samples with a dialect label, except for labels *Auvernhat* and *Vivarés*, which are mislabelled, as shown in Section 5. Unlabelled samples are used separately and in combination with labelled ones for other analyses. Table 3 shows some basic statistics of the data passed to the NMF preprocessing and training pipeline.

Evaluation To evaluate, we predict the dominant topic of each sample of the test data. The subset with self-declared dialect labels is used to compute automated metric scores, but we also use the predictions on unlabelled samples for a manual analysis based on a map visualisation.

For the automated evaluation, we follow Rabus and Scherrer (2025) and use the following popular clustering metrics: homogeneity (used to measure

	AUV		GAS		LAN		LIM		PRO		VIV		unk		Total	
	#s	#t	#s	#t	#s	#t	#s	#t	#s	#t	#s	#t	#s	#t	#s	#t
Training	81	13k	915	85k	897	72k	2402	240k	531	46k	413	39k	2653	276k	8478	820k
Test - all	109	13k	2120	104k	2093	90k	4199	269k	1056	55k	788	45k	4845	312k	16281	945k
Test - w/ dialect	0	0	2120	104k	2093	90k	4199	269k	1056	55k	0	0	-	-	9468	518k
Test - w/o dialect	-	-	-	-	-	-	-	-	-	-	-	-	4845	312k	4845	312k

Table 3: Statistics on the training and test data used during our clustering experiments. Each cell shows the number of samples (#s) and tokens (#t), for each subset based on self-declared dialect. Values abbreviated with 'k' are in thousands.

to what extent all predictions of the same topic belong to samples of the same dialect subset), completeness (used to measure to what extent all samples of the same dialect subset have the same topic prediction), and V-score (the harmonic mean between the two metrics above). Since the dialect classes are very imbalanced in the test sets, we duplicate the predictions of under-represented classes when computing the scores above, so that the difference of each class to the most represented one is below 5%. We compute these metrics to compare predicted subsets per dominant topic with reference subsets from self-declared dialect labels, focusing on the V-score value. We also use the completeness metric to compare the same predicted subsets with subsets by user ID, to ascertain to what extent all samples of the same user are assigned to the same topic.

6.2. Results

Our NMF clustering model trained for 4 topics (NMF-t4) resulted in the best balanced V-score of 64.43, completeness score of 64.54 and second-best homogeneity score of 64.32 when evaluated on dialect labels, and a clear one-to-one topic-to-dialect mapping (see Figure 6). With this model, users are generally assigned a low number of topics: the completeness score in comparison to user IDs is 72.74. Scores when varying the number of topics are presented in Appendix D.

These scores are relatively low compared to the results reported in (Kuparinen and Scherrer, 2024) and (Rabus and Scherrer, 2025). However, the experimental settings differ in several aspects: those studies use normalised transcriptions of long speech interviews, whereas our samples are short user-generated posts. Moreover, in our setting, speakers of different varieties interact, which may lead authors to adopt more mutually intelligible or shared conventions, or to incorporate features associated with other varieties, due to accommodation and dialect contact mechanisms (Dragojevic et al., 2015; Britain, 2017).

Nevertheless, several of the top features per topic from our NMF-t4 model (see Table 4) clearly relate to salient features of the dialects they were

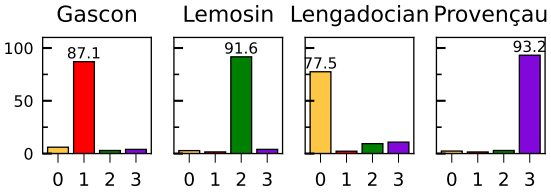

Figure 6: Predicted NMF-t4 topic distribution for each labelled dialect subset.

ID	Dialect	Feats
0	*Lengadocian*	la que a del e es lo d amb una pas se un per l al me mas en aquo
1	*Gascon*	ei mes que ua qu qui pas deu e n a mei dab la lo ne dens un en com
2	*Lemosin*	quo mas es dau daus pas per la que un a los lo las e l i en qu na
3	*Provençau*	lei que es mai l lo la si a per d e mi en un pas una ben dei li

Table 4: Top 10 features per topic for model NMF-t4. The dialect column reflects the observed one-to-one topic–dialect mapping.

mapped to. For instance, in the *Lengadocian* topic, features such as *del*, *al* and *l* could relate to the [l] phone in final position, which has been maintained throughout *Lengadocian*'s evolution from Latin, in contrast to most other varieties where it is frequently vocalised to [w]–as in the topic features such as *deu* (*Gascon*) and *dau*/*daus* (*Lemosin*). As another example, the most weighted feature that maps to *Provençau*–*lei*–relates to the specific form of the definite plural article used in southern *Provençau* varieties, as opposed to *los*/*las*, more common in other varieties.

The map of topics distribution per *département* (see Figure 7) reveals that the topics associated with each dialect broadly follow the linguistic maps of the Occitan continuum, with *Provençau* predictions in the South-East, *Lengadocian* in the central South, *Lemosin* in the North-West, and *Gascon* in the South-West. Looking in more detail, we notice that *départements* in the *Provençau* area that have borders in contact with other varieties (*Lengadocian* and *Vivarés*) display a lower ratio of the *Provençau* topic among the predicted samples. However we do not observe such a tendency in

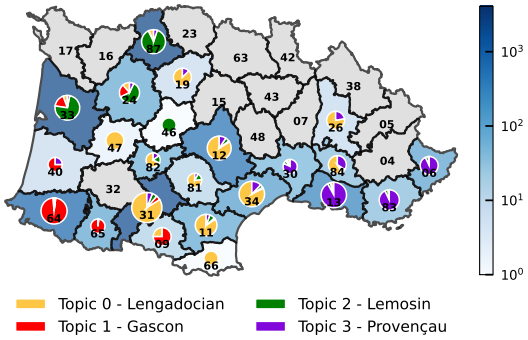

Figure 7: Map with the distribution of NMF-t4 topics per *département*, in or close to the Occitan-speaking area (France only). Background colours indicate the number of samples predicted in that *département* (log scale). The size of each pie chart is proportional to the number of users represented in that *département*'s samples, also on a log scale.

the *Lengadocian*-speaking area, although it is described in (Léonard et al., 2024). Zooming in on the *Gascon* area, we observe that the ratio for this topic differs between *départements*,[19] in a way that is similar to the description of what is called *gasconity gradient field* in (Sibille, 2024), which shows a maximal expression of *Gascon* dialectal features in the Pyreneans (as we can see in *départements* 64 and 65) and on the contrary a very low expression in the north of the area (compatible with the ratio of *Gascon* topic predictions in *département* 33, near Bordeaux). Our experiments with NMF models trained for at least eight topics reflect this internal variation of the *Gascon* variety with more details, as an additional topic emerges with top features that clearly relate to some of the most important features of the Pyrenean *Gascon*, such as the use of the masculine singular article *eth* (see Table 6, Appendix D).

When used as a classification model, our NMF-t4 model has a good potential for dialect identification applications as the macro-average F1-score of dialect predictions compared to reference labels is 85.50. While the distribution of dialect labels in the labelled data is imbalanced, with an over-representation of *Lemosin*, the predictions on the unlabelled samples[20] are characterised by a strong under-representation of this variety, which results in a fairly balanced distribution of the four major varieties within the dataset, as shown in Figure 8, frequencies ranging from 18% (*Provençau*) to 30% (*Lengadocian*).

[19]We ignore *départements* 31 and 09 as we suspect them to include both *Gascon* and *Lengadocian* samples, as most isoglosses between these varieties are crossing these *départements*.

[20]still with LID threshold $\geq$0.9 ; no length filtering

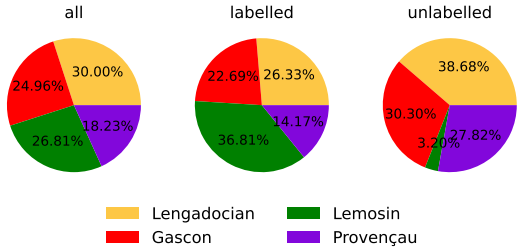

Figure 8: Distribution of dialect predictions using NMF-t4 across FORUMOCCITANIA (samples with a high Occitan LID score, with or without a dialect label), where the most frequent topic per dialect was renamed with this dialect (one-to-one mapping).

7. Conclusion

In this paper, we introduced the FORUMOCCITANIA corpus, containing authentic UGC texts in Occitan written using the classical spelling, and accompanied by user-provided metadata and LID predictions.

Statistical and qualitative analyses, along with preliminary dialectometry experiments, show that the dataset has strong potential for linguistic research and NLP, owing to rich sociolinguistic metadata (dialect, age, proficiency, location). Most forum users who produced posts are relatively young, although some are older fluent speakers.

The corpus spans a large portion of the Occitan-speaking area. Although *Lemosin* is over-represented among self-declared dialects, clustering on both labelled and unlabelled posts results in a fairly balanced distribution across four major varieties (*Lengadocian*, *Lemosin*, *Gascon*, *Provençau*). Our best NMF-based topic model, trained on four topics, reached a balanced V-score of 64.43 and a classification F1-score of 85.50. Furthermore, geographical analysis of clusters further reveals fine-grained variation, e.g. gradual shifts within *Gascon*.

The dataset will be made available in XML-TEI and JSONL formats for research purposes, including train/dev/test splits made for dialect-aware NLP experiments.

Limitations

Although our dataset helps increase the representation of under-represented Occitan varieties (especially *Lemosin* and *Provençau*), several others remain largely absent from NLP datasets, including *Auvernhat* and *Vivaroaupenc* for which the labels in FORUMOCCITANIA were found to be unreliable. The inclusion of language learner productions offers potential insights from linguistic and didactic perspectives. However, because many forum users did not report advanced Occitan proficiency, our

dataset might not be suitable for NLP tasks such as language generation or fluency evaluation.

In our experiments, dialect clustering was evaluated using self-declared labels. Manual validation or correction of these labels at least in the test set could result in more reliable performance scores. Likewise, our training data was filtered only by text length and LID scores; additional filtering based on self-reported or externally validated Occitan proficiency might further enhance model performance.

Finally, although our unsupervised approach confirmed the general alignment between self-declared dialects, broad geographical areas, and dialectal features, the limited number of users per location constrains the interpretability of results at finer geographical scales, such as individual cities or cultural regions.

Ethical Statement

The ForumOccitania corpus contains user-generated content from a publicly accessible online forum, along with metadata from the public profiles of active users. Some of the contents might be considered as sensitive, e.g. political opinions. In order to collect, analyse, and release the data, certain legal and ethical aspects need to be taken into account, in particular regarding intellectual property and user privacy.

The data collection, analyses and experiments were conducted on the basis of article 3 of the European Directive 2019/790 on copyright and related rights in the Digital Single Market, which creates an exception that allows research organisations, for the purposes of scientific research, to scrape and otherwise collect protected works, to store copies of those works, and to carry out text and data mining on the collected materials. In France, this exception was incorporated into national law (Article L122-5-3 of the French Intellectual Property Code), where it is stated that *opt-out* mechanisms (such as instructions in a *robots.txt* file) do not override the exception for scientific research.

Given the potential for future research in Occitan linguistics and NLP, we will make the anonymised dataset accessible for research purposes, subject to conditions detailed in the dataset repository.

Acknowledgements

This work was funded by Inria under the "*Défi*"-type project COLaF. It was also 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.

We are grateful to Djamé Seddah and Arij Riabi for their helpful advice on anonymisation, and would also like to thank Aina Garí Soler for her insights on the samples predicted as Catalan, Sarah Bénière for her comments to help improve the TEI version of the corpus, and the reviewers for their constructive feedback.

8. Bibliographical References

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A Toolkit for Anonymizing Unstructured Chat Data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.

Sarah C. Brennan. 2024. 21 La situation sociolinguistique de l'occitan du début du XXe siècle à nos jours. In Louise Esher and Jean Sibille, editors, *Manuel de linguistique occitane*, pages 593–621. De Gruyter.

David Britain. 2017. Dialect Contact and New Dialect Formation. In *The Handbook of Dialectology*, pages 143–158. John Wiley & Sons, Ltd.

J. K. Chambers and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press. Google-Books-ID: 9bYV43UhKssC.

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

Marko Dragojevic, Jessica Gasiorek, and Howard Giles. 2015. Communication Accommodation Theory. In *The International Encyclopedia of Interpersonal Communication*, pages 1–21. John Wiley & Sons, Ltd.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECT-BENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. 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.

Dirk Hovy and Christoph Purschke. 2018. Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Juliette Janès, Rachel Bawden, Thibault Clérice, Rasul Dent, Lucence Ing, Oriane Nédey, and Benoît Sagot. 2025. Encoding language diversity in TEI: a description of regional and non-standard languages in multilingual diachronic corpora. In *TEI Conference 2025*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural Language Processing for Dialects of a Language: A Survey. *ACM Comput. Surv.*, 57(6):149:1–149:37.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Olli Kuparinen and Yves Scherrer. 2023. Dialect Representation Learning with Neural Dialect-to-Standard Normalization. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, Dubrovnik, Croatia. Association for Computational Linguistics.

Olli Kuparinen and Yves Scherrer. 2024. Corpus-based dialectometry with topic models. *Journal of Linguistic Geography*, 12(1):1–12.

Xavier Lamuela. 2024. 20 Codification et élaboration linguistiques. In Louise Esher and Jean Sibille, editors, *Manuel de linguistique occitane*, pages 563–589. De Gruyter.

Hervé Lieutard. 2024. Provençal et niçard : des configurations exemplaires des tendances intégratives et séparatrices en domaine occitan. In *Les noms des variantes de langue minoritaire Études de cas en France et en Russie*, presses universitaires de bordeaux edition, Diglossi@_2, pages 99–117 [en ligne]. Presses universitaires de Bordeaux; PUB, Pessac. ISBN: 9791030008395 Section: Langues.

Jean Léo Léonard, Guylaine Brun-Trigaud, and Flore Picard. 2024. 17 Atlas linguistiques et perspectives dialectométriques. In Louise Esher and Jean Sibille, editors, *Manuel de linguistique occitane*, pages 473–520. De Gruyter.

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.

Aleksandra Miletić. 2023. Outiller l'occitan : nouvelles ressources et lemmatisation. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 217–231, Paris, France. ATALA.

Aleksandra Miletić, 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).

Iñigo Morcillo, Igor Leturia, Ander Corral, Xabier Sarasola, Michaël Barret, Aure Séguier, and Benaset Dazéas. 2024. Automatic Speech Recognition for Gascon and Languedocian Variants of Occitan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1969–1978, Torino, Italia. ELRA and ICCL.

NLLB Team, 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:2207.04672 [cs].

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/Langdoc: Defining Dialects, Languages, and Language Families as Collections of Resources. In *Proceedings of ISWC 2011*, Bonn, Germany.

Oriane Nédey, Rachel Bawden, Thibault Clérice, and Benoît Sagot. 2026. OcWikiDialects: A Wikipedia Dataset with Rich Metadata for Occitan Dialect Identification. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco.

Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

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. ArXiv:2506.20920 [cs].

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

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.

Achim Rabus and Yves Scherrer. 2025. Dialects, Topic Models, and Border Effects: The Rusyn Case. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 38–43, Vienna, Austria. Association for Computational Linguistics.

Arij Riabi, Menel Mahamdi, Virginie Mouilleron, and Djamé Seddah. 2024. Cloaked Classifiers: Pseudonymization Strategies on Sensitive Classification Tasks. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 123–136, Bangkok, Thailand. Association for Computational Linguistics.

Yves Scherrer, Rob van der Goot, and Petter Mæhlum. 2025. Findings of the VarDial Evaluation Campaign 2025: The NorSID Shared Task on Norwegian Slot, Intent and Dialect Identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–8, Abu Dhabi, UAE. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. ArXiv:1907.05791 [cs].

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*, AAAI'25. AAAI Press.

Marianne Vergez-Couret, Myriam Bras, Aleksandra Miletić, and Clamença Poujade. 2024. Loflòc: A

Morphological Lexicon for Occitan using Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10716–10724, Torino, Italia. ELRA and ICCL.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Özlem Çetinoğlu and Antje Schweitzer. 2022. Anonymising the SAGT Speech Corpus and Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5557–5564, Marseille, France. European Language Resources Association.

9. Language Resource References

Omar Kamali. 2025. *Wikipedia Monthly*. Omneity Labs. HuggingFace datasets, 1.0.

Séguier, Aure and Lo Congrès. 2023a. *Occitan Corpus from Lo Congrès news*. Lo Congrès permanent de la lenga occitana. Lo Congrès, distributed via Zenodo: DOI 10.5281/zenodo.8411197, 1.0.

Séguier, Aure and Lo Congrès. 2023b. *SoftwaresOccitanTranslations corpus*. Lo Congrès permanent de la lenga occitana. Lo Congrès, distributed via Zenodo: DOI 10.5281/zenodo.8411351, 1.0.

Séguier, Aure and Lo Congrès. 2024. *Lo Congrès websites Corpus*. Lo Congrès permanent de la lenga occitana. Lo Congrès, distributed via Zenodo: DOI 10.5281/zenodo.12192029, 1.0.

Vergez-Couret, Marianne and Bras, Myriam and Miletić, Aleksandra and Poujade, Clamença. 2024. *Loflòc: A Morphological Lexicon for Occitan using Universal Dependencies*. 1.0.

Zeman, Daniel. 2025. *Universal Dependencies 2.17*. Universal Dependencies Consortium, 2.17.

A. Language Proficiency Levels

In our analyses of ForumOccitania, we converted the text levels of Occitan proficiency to numerical levels, as used in Figure 3, as follows (translations in parentheses):

- 1: *"Comprèni, parli pas"* (I understand, but I can't speak it),
- 2: *"Parli un pauc"* (I speak it a little),
- 3: *"Parli plan, fau sonque de pichòtas dècas"* (I speak it well, I only make small mistakes),
- 4: *"Parli corentament"* (I speak it correctly),
- 5: *"Soi confirmat, dins mai d'un dialècte"* (I am proficient, and in more than one dialect),
- 6: *"Parli naturalament"* (I speak it fluently).

B. Additional Analyses

The following plots (Figure 9 and Figure 10) give additional insights into the distribution of the corpus.

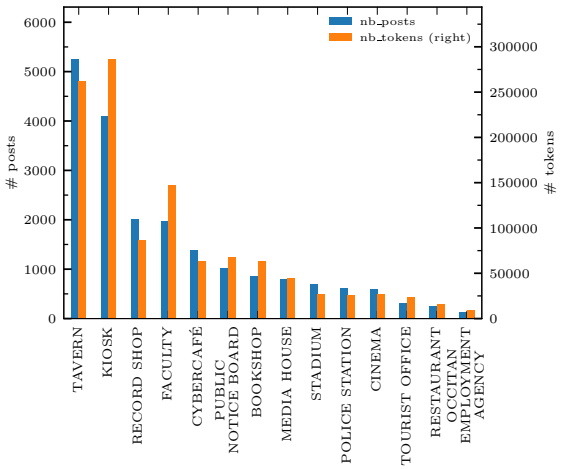

Figure 9: Distribution of posts and tokens across main forum categories. Forum titles have been translated to English (originally in Occitan).

C. Labels *Auvernhat* and *Vivarés*

As mentioned in Section 5, we consider the user-declared dialects *Auvernhat* and *Vivarés* to be mislabelled, potentially due to a bug following a website update. Table 5 presents excerpts of posts associated with those labels in the dataset. The excerpt labelled as *Auvernhat* presents several distinctive *Gascon* characteristics, such as the enunciative *que*, the verb *deishar* 'to leave' and the form *ua* for the feminine indefinite article. *Vivaroaupenc* is linguistically closer to *Lengadocian*, and Sibille

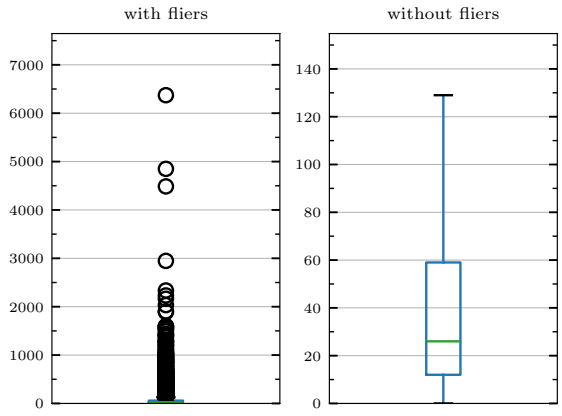

Figure 10: Distribution of the number of tokens per post, with fliers (left) and without (right). Fliers are defined as posts with more tokens than *1.5 * IQR (IQR = Q3 - Q1)*.

(2024) describe a geographically complex diatopic situation with diverging isoglosses and many subregions exhibiting opposing characteristics. Nevertheless, the excerpt labelled as *Vivarés* (as well as the rest of the subset) shows neither widespread features such as the palatalisation of [k] (*causas* > *chausas* 'things') or the indicative present first-person singular ending in *-o* (*compreni* > *compreno* 'I understand'), nor more localised features such as variants of the final /l/ (*cal* > *chau*).

Dialect	Sample excerpt and translation
Auvernhat	que'vs deishi aqui un extracte d'ua entervista que trobaratz suu siti). *Here I leave you an excerpt from an interview that you will find on the website.*
Vivarés	Bah! cal pas estre trop dur amb aquela musica, que compreni qu'agrade a d'unas personas. Cal tojorn veser lo bon costat de las causas, aquo pot menar d'unis a s'interessar a la (vertadiera)musica occitana. *Bah! One must not be too harsh with that music; I understand that it pleases some people. One must always see the good side of things: this can lead some to take an interest in (true) Occitan music.*

Table 5: Excerpts of seemingly mislabelled *Auvernhat* and *Vivarés* posts in the FORUMOCCITANIA dataset.

D. Varying the Number of Topics

Figure 11 reports the evolution of performance scores when varying the number of topics of the NMF model. Table 6 displays the top ten features per topic when training for eight topics. As mentioned in Section 6.2, we can observe that features of the *Gascon* variety are spread across topics 1 and 7, with the topic 7 giving important weight to the masculine singular article *eth*, which is an important characteristic of Pyrenean *Gascon*.

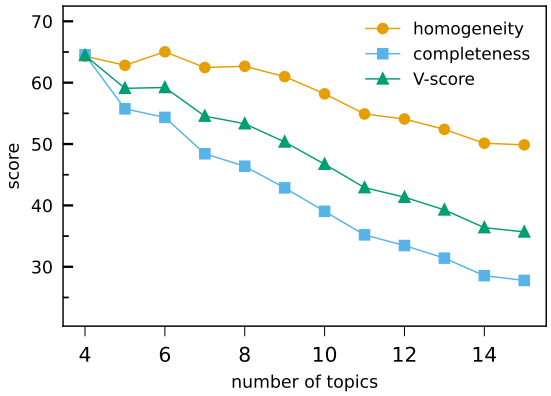

Figure 11: Evolution of scores as the number of topics varies (hyperparameter of the NMF training pipeline). Scores are computed in comparison with user-declared dialects, on the test split of FORUMOCCITANIA.

Topic ID	Features
0	es pas que mas una un d aquo amb a
1	ei que mes qu ua pas deu mei qui dens
2	quo dau mas es daus pas que un per lo
3	lei que si mai mi es lo l dei ben
4	las los la son ion cio cion l ci acion
5	a la lo e grop sus musica un l vos
6	cedric te amistats faire aprep me radio que ieu se
7	eth era eths ei mes eth th qu eth ua

Table 6: Top 10 features per topic for the model NMF trained for eight topics. Duplicates are possible as word and character n-gram features are mixed, and word boundaries are not shown here.