Rebelòt: Datasets and Token-Level Language Identification for Lombard-Italian-English Code-Mixing

Edoardo Signoroni*, **Emma Bednaříková***, **Pavel Rychlý**
Faculty of Informatics - Masaryk University
Botanická 68a 60200 Brno Czechia
e.signoroni@mail.muni.cz, 536251@mail.muni.cz, pary@fi.muni.cz

Abstract

Lombard is an endangered and under-resourced Gallo-Italic language variety that exists with Standard Italian. As with other language varieties of Italy, code-switching and code-mixing is common between Lombard and Italian in everyday conversation and with English, online. This linguistic complexity, and the lack of a unified written standard, poses challenges for Natural Language Processing tools. We introduce *Rebelòt*, a novel multi-domain, token-level annotated dataset for Lombard-Italian-English code-mixing. Furthermore, we develop and evaluate three variants of a token-level Language Identification (LID) tool based on a pre-trained encoder architecture, fine-tuned using both authentic data from our corpus and synthetically generated code-mixed text. Our evaluation demonstrates that the optimal model variant achieves an accuracy of over 0.99 on token-level prediction, and substantially outperforms widely used off-the-shelf LID baselines at sentence-level.

Keywords: Code-Mixing, Language Identification, Lombard, Low-Resource Languages

1. Introduction

Language identification is a widely practiced task within the field of natural language processing (NLP). The applications of language identification are numerous, including content filtering on social media or web searches, with the purpose of providing users with data in their native or preferred language. Further applications include data collection and labeling for subsequent processing or identification of the language of text given as a source to a machine translation system (Jauhiainen et al., 2019).

In many cases of language identification, it is sufficient to provide information about the language on the document- or sentence- level. Nevertheless, in certain scenarios, the identification on a lower level may be preferable. For instance, in the context of training machine translation systems, a common practice is to obtain parallel data from movie subtitles (Lavecchia et al., 2007). Such data is typically in the form of transcripts and translations of informal language speech, where code-mixing is likely to be present (Gumperz, 1977). Another application of token-level language identification arises in text-to-speech (TTS) synthesis. Since many systems rely on language-specific synthesizers, identifying the language of each token is necessary to select the correct pronunciation rules. Without this step, code-mixed input risks being rendered with unnatural or inaccurate speech.

While language identification is mostly done at sentence level, the actual usage cases may be more nuanced and complex. Bilingualism and multilingualism are the norm throughout the world (Wardhaugh, 2006), thus code-switching and code-mixing are widespread linguistic phenomena. Muysken (2000) defines **code-switching** as the succession of several languages in a single speech event (inter-sentential code-switching), and **code-mixing** as the insertion of lexical items or grammatical features from more languages in one sentence (intra-sentential code-switching).

This is especially true for Italy, which represents one of the most, if not the most, linguistically diverse areas of Europe (Maiden and Perry, 1997). In this area, the alternation of the Italian language and the language varieties coexisting with it is a common phenomenon in everyday conversation, especially after the Second World War, when spoken Standard Italian has spread to all social classes (Berruto, 1997).

Due to complex historical and socio-political factors, the local language varieties were often subjected to marginalization, legal ambiguity, and social stigmatization, leading to their declined use (ISTAT, 2026). Nowadays, more than 30 of these are listed as endangered by the UNESCO (Moseley, 2010). Since their natural generational transmission is broken, these varieties will fade away with incalculable cultural and linguistic loss (Maiden and Perry, 1997).

One such language variety is Lombard, spoken in the area centered around the Italian region of Lombardy.

This paper makes some key contributions:

- ***Rebelòt***[1] a multi-domain corpus of Lombard-

* Equal Contribution

[1] *Rebelòt* [re.be'lɔt]: Confusion, disorder, chaos.

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 253–262
11–12 May 2026.

Italian-English Code-Mixing.

- three variants of **a novel LID tool** operating on the token level. Our tools were found to outperform off-the-shelf LID tools, such as FastText and GlotLID.

Both the datasets and the language identification models are available online.[2]

The remainder of the paper is structured as follows: Section 2 gives an overview of Lombard and its current situation; Section 3 describes some related work for language identification and code-mixing; Section 4 outlines the methodology for the creation of the corpus and the token-level language identification models; Section 5 discusses their evaluation and results; and finally Section 6 gives some conclusions.

2. The Lombard Language

Lombard is a Western Romance language belonging to the Gallo-Italic (Cisalpine) group, spoken by around 3.5 million people in the Italian region of Lombardy, the Piedmontese provinces of Novara and Verbano-Cusio-Ossola, and the Swiss cantons of Ticino and Graubünden. Lombard constitutes a linguistic continuum where, despite internal phonetic, lexical, and morphosyntactic differences, varieties remain largely mutually intelligible (Coluzzi et al., 2018; Bonfadini, 2010; Loporcaro, 2009; Coluzzi et al., 2021). Scholars typically categorize these into two to four main branches, the most prominent being Western Lombard (provinces of Milan, Monza, Varese, Como, Lecco, Sondrio, Lodi, and Pavia, eastern Piedmont and Canton Ticino) and Eastern Lombard (provinces of Bergamo, Brescia, and Cremona).[3]

UNESCO's "Atlas of the World's Languages in Danger" (Moseley and Nicholas, 2010) lists Lombard as a "Definitely endangered" language.[4] According to other metrics, such as EGIDS (Extended Graded Intergenerational Disruption Scale) (Lewis and Simons, 2010), Lombard is between grades 6b "threatened" and "moribund". The status of Lombard varies by area. In Switzerland Lombard varieties enjoy higher social prestige and institutional support, notably through the research and preservation efforts of the "*Centro di dialettologia e di etnografia*" (CDE) in Bellinzona. In Italy, the situation is quite different: local languages like Lombard are frequently stigmatized and colloquially labeled as *dialetti* ("dialects"). This term often carries a pejorative connotation, inaccurately implying that these varieties are subordinate derivatives of Standard Italian rather than independent linguistic developments. Consequently, the term "language varieties" is preferred in contemporary scholarship as a more neutral designation that avoids prestige-based bias (Ramponi, 2024). In Italy, Lombard holds limited recognition, and only at the regional level: Lombardy's Regional Law 25/2016 states that "the Region promotes the revitalization, the valorization, and diffusion of all the local variants of the Lombard language, as they are meaningful expressions of immaterial culture".[5]

Today, Lombard exists in a state of *dilalìa* (Berruto, 1987) with Standard Italian (Ramponi, 2024). While Italian dominates formal and official domains, Lombard is increasingly restricted to informal contexts, where it frequently overlaps and mixes with the national language. According to official data (ISTAT, 2026), in 2024 only 4.2% of the population in Lombardy spoke exclusively the local variety at home, while 18.1% used both it and Italian. These figures signal a decline in use, especially mixed use, from the previous report about 2015 (ISTAT, 2017), where the percentages were 5.6% and 26.1%.

Despite a rich history of local literary traditions, modern Lombard remains predominantly an oral medium. Even if some proposals for a standardized orthography were put forward,[6] the language lacks a codified, unified written form, with most speakers employing ad-hoc phonetic approaches when writing the language, especially for artistic purposes, such as poetry or theatrical plays.

3. Related Work

3.1. Language Identification

Language identification (LID) is a well-studied task, and widely used tools achieve strong performance in high-resource settings. However, many of these systems are trained on dozens of languages and

[2] https://github.com/edoardosignoroni/rebelot

[3] Due to 19th and 20th-century migratory flows, a variant of Eastern Lombard (Bergamasque) is also present in parts of southern Brazil (Paganessi, 2017).

[4] A language that "is no longer being learned as the mother tongue by children in the home. The youngest speakers are thus of the parental generation. At this stage, parents may still speak their language to their children, but their children do not typically respond in the language."

[5] From Article 24. https://normelombardia.consiglio.regione.lombardia.it/Accessibile/main.aspx?exp_coll=lr002016100700025&view=showdoc&iddoc=lr002016100700025&selnode=lr002016100700025 (in Italian, own translation).

[6] E.g. on the Lombard Wikipedia: https://lmo.wikipedia.org/wiki/Wikipedia:GrafCat (in Lombard)

often provide limited coverage for low-resource languages. To address this limitation, Kargaran et al. (2023) introduced GlotLID, a language identification tool based on the FastText architecture that covers more than 1,600 languages.

Another challenge that most well-known LID tools do not sufficiently address is the identification of languages in code-mixed texts, where tokens from multiple languages occur within the same sentence. The importance of token-level language identification in such settings has been widely acknowledged in recent research.

An annotated corpus of German–English mixed-language data, Denglisch (Osmelak and Wintner, 2023), has been developed to support training of statistical classifiers. This system utilizes Conditional Random Fields (Lafferty et al., 2001) to provide the classification.

Another study on German–English multilingual usage was presented by Sterner and Teufel (2023), who introduced a rule-based method for identifying language alternation between the two languages. Their approach was used to annotate a corpus of 25.6M tweets containing German–English mixing. The resulting corpus was subsequently used to pre-train a neural language model, which was fine-tuned for token-level language identification.

Numerous other tools and datasets have been introduced for bilingual mixed-language scenarios. Sabty et al. (2021) developed a token-level language identification system tailored to English–Arabic mixed texts. Similarly, Nayak and Joshi (2022) and Dey and Fung (2014) focus on Hindi–English multilingual usage within single texts.

While these works concentrate on bilingual settings, other research has extended token-level language identification to multilingual contexts. For example, Zhang et al. (2018) proposed a model capable of token-level identification across 100 languages. Their approach combines a feed-forward classifier with a decoder incorporating global constraints.

From a neural perspective, Santy et al. (2021) examined the impact of different types of training data for token-level identification in mixed-language texts. Their findings indicate that training on authentic multilingual data yields the strongest performance.

3.2. Code-Mixing and Code-Switching in the Italian Context

To our knowledge, no recent computational linguistics or NLP work targeted Lombard code-mixing and code-switching. Most, if not all the work, is conducted on speech from a sociolinguistics point-of-view.

Among these, some recent contributions on Lombard varieties include Cerruti (2018), which deals with grammatical aspects of code-switching in a corpus of Italian and *Bresciano* speech data; and Andreoli (2022), who focuses on code-switching and code-mixing between the local variety and Italian in the community of speakers of Poggiridenti, in the Lombard province of Sondrio.

There is further work on the interaction of Italian and some other language varieties. Alfonzetti (2015) deals with intergenerational patterns of code-switching between Italian and Sicilian both in a transcribed speech corpus and a written corpus of e-mails, text messages, and social networks posts. Frighetto (2025) examines code-mixing between Italian and Venetian dialect in digital communication on Facebook groups and private conversations on WhatsApp.

Still relevant to the Italian linguistic landscape, some effort has been directed to the study of code-switching and code-mixing between Italian and some of the regional languages recognized by law. For example, Fiorentini (2017), focuses on the use of Italian discourse markers by Ladin speakers in the three valleys (Fassa, Badia, Ghardena) of the Ladin area of Trentino-South Tyrol. Geographically close, Dal Negro and Ciccolone (2018) (and other works related to the *Kontatto*[7] corpus), investigates language contact patterns between Italian and the local Bavarian German variety in South Tyrol.

4. Methodology

4.1. The Rebelòt Dataset

Rebelòt is a multi-domain dataset of Lombard-Italian-English code-mixing. It was built with two main aims. First, as a resource to investigate Lombard code-mixing and code-switching from a computational linguistics point-of-view. Second, as a training dataset for NLP tools that can handle such complex situations, such as language identification models. This paper focuses on the latter application. The focus of the dataset is Eastern Lombard, mainly its Brescian and Bergamasque varieties.

Collection and Sources The data for the corpus were collected from different sources and mediums. While all of the data comes from the web, some originally belong to physical mediums (e.g. books).

The **books** text is from the book "Mondo Popolare in Lombardia - Fiabe Bergamasche" ("Folk World in Lombardy - Bergamasque Tales")(Anesa and Rondi, 1981). A compilation of Bergamasque tales, gathered from native speakers in interviews and transcribed.[8] The digitalized version of the

[7] https://kontatti.projects.unibz.it/

[8] The authors use a transcription form specifically devised for this series of books with the aim to preserve

book is accessible on the website of the "Archivio di Etnografia e Storia Sociale della Regione Lombardia" (Archive of Ethnography and Social History of Lombardy).[9] We focused on pages 29 to 37, relating the interviews about how the speakers learned their tales and their context. The text is structured with the interviewer question, the answer in Bergamasque and the Italian translation later on. This can be regarded as a code-switching situation, where the interviewer text is in Italian, and the answers are in Lombard.

The text for the **news** domain are from the "Giornale di Brescia", the local newspaper of Brescia and its province. The articles are freely accessible after logging in with an email. We collected the articles about the Brescian variety[10]. They discuss idioms, etymology,[11] and cultural concepts of Brescian, and thus they are mostly in Italian, with Lombard sentences mixed within. This is closer to typical code-mixing.

sayings is the smallest domain, and its text was collected from a website about the *Dialèt de Brèsa* ("Dialect of Brescia").[12] While the website contains also a dictionary and a grammar description, the sayings page consists of a single page of common sayings and their Italian translation.

The **socials** domain and its text may be the most interesting and challenging from both a linguistics and a computational point of view. It consists of the manually collected text from the post of a page about Brescian.[13] The average post presents a common idiom from Brescian and its English translation, followed by a short sentence describing its typical use or provenance. This section combines and mixes Lombard, Italian, and English, with creative aims thus presenting an high degree of code-mixing.

Table 1 gives one text sample for each domain of the corpus.

Processing and Annotation The text in the *books* domain needed preliminary processing steps due to its different medium. First, we run the PNG files of the relevant pages through `tesseract` OCR. Some brief experimentation shows that using both Italian and German models (`ita+deu`) leads to slightly better results than using just the Italian one.[14] The resulting text was then manually corrected for errors.

The text was collected and manually annotated in plain text format by a native speaker of Eastern Lombard, proficient in all three languages in the dataset. Italian was assumed by default, Lombard parts where marked with `<lmo> some_lmo_text </lmo>`, similarly the English parts were annotated with `<eng> some_eng_text </eng>`. Non-linguistic text, such as punctuation, was annotated with the previous language span.

The text was then anonymized by substituting personal names with `"Nome Cognome"` ("Name Surname"), and usernames with `@utente` (@user). We then converted all the text in vertical format, where each sentence is identified by a unique ID in the form `overall-index_domain_domain-index`, e.g. `512_socials_296`. For each sentence, tokens are listed in order with one line consisting of `token_index token lang_annotation`, e.g. `319 sègn lmo`. For the experimental phase, we handled the text in JSONL format.

Composition and Statistics The *Rebelòt* corpus consists of 704 lines and documents, for a total of almost 80k words divided in four domains. Three quarters of the data (217 lines and ~60k words) comes from *news*, these are longer documents averaging 276 words per line. They contain ~90% Italian text, intermixed with Lombard sentences or expressions. The *socials* domain is the second in size, with 413 lines and ~15k words: one fifth of these are in Lombard, while the rest is almost evenly split between Italian and English, for which this domain is the almost exclusive source. The text from *books* accounts for 58 longer lines of paragraph length, for ~5k words, evenly split between Italian and Lombard. Lastly, *saying* are the smallest domain of the corpus, with 16 short lines for less than 400 words, two thirds of which are in Italian. In creating the `train`, `dev`, `test` splits, we mix all the domains. Overall, the text in the corpus is ~79% Italian (63k words), ~14% Lombard (11.5k words), and ~7% English (6k words). Table 2 reports some

as best as they can the orality and peculiarities of each speaker and variant. This transcription differs from the orthographies currently proposed and currently used (see Signoroni and Rychlý (2026) for a more detailed description of Lombard orthographies).

[9] https://aess.regione.lombardia.it/da/viewer/?volume=011-01

[10] https://www.giornaledibrescia.it/tag/dialetto%20bresciano

[11] Especially in texts dealing with this topic, some other languages, even historical ones such as Latin or Gothic, were present. Since these instances were not numerous, at this stage we choose to not annotate them.

[12] https://sites.google.com/site/dialetdebresa/proverbi-bresciani-antichi-e-moderni

[13] https://www.instagram.com/bresciadice/?hl=en

[14] We used `tesseract` because was readily available and enough for the purposes of this work. A broader and systematic evaluation of OCR for Lombard and its varieties is out of the scope of this paper and demanded to future work.

books:	*Del « <lmo>Refenìstola</lmo> », mitica figura di vecchio ambulante che veniva da molto lontano, ci vengono presentate anche le condizioni di vita. [...]* Of the « <lmo>Refenìstola[a]</lmo> », mythical figure of the old peddler for far away, we are shown also the living conditions. [...]
news:	*<lmo>«Arda chi gh'è, stét bé?</lmo> Stai bene?»* *Il barista compensa la bassa statura con l'alto volume della voce. [...]* <lmo>"Look who's there, how are you?</lmo> How are you?" The barman compensates for his short stature with the high volume of his voice. [...]
sayings:	*<lmo>Quand che l'amùr al gh'è, la gamba la tira 'l pè.</lmo>* *Quando c'è l'amore, la gamba trascina il piede.* *Tutto va da da sé, quando si è innamorati.* <lmo>When there is love, the leg pulls the foot.</lmo> When there is love, the leg pulls the foot. Everything goes by itself, when one is in love.
socials:	*<lmo>'Gho gnà na lira'</lmo> • <eng>have no penny, lit.</eng>* *<eng> If your friends suggest to have a two-week holiday in Ibiza but you can't afford, you say</eng> <lmo>(vecio) gho gnà na lira!</lmo>* *Quanti in vacanza sul balcone quest'estate? #BresciaDice <eng>#BrixianSays</eng> #Ibiza #Brescia #Vacanze <lmo>#GhoGnàNaLira</lmo>* <lmo> I don't have even one penny! </lmo> <eng>have no penny, lit.</eng> <eng> If your friends suggest to have a two-week holiday in Ibiza but you can't afford, you say</eng> <lmo> (old one) I don't have even one penny! How many are having holidays on the balcony this summer? #BresciaDice <eng>#BrixianSays</eng> #Ibiza #Brescia #Vacanze <lmo>#GhoGnàNaLira</lmo>

[a]The name comes from the shout the peddler made while announcing his arrival: *"Rèf e nìstola!"* (it. "Refe e fettuccia!", eng. "Yarn and ribbon!")

Table 1: Partial samples for each domain in the dataset.

Domain	Lines	Words	Chars	Vocab	Avg W/L	Avg W.Len	<eng>	<ita>	<lmo>
books	58	4674	21960	1472	80.59	3.71	0	2456	2223
news	217	59905	357554	17690	276.06	4.97	13	53876	6543
sayings	16	388	2041	269	24.25	4.29	0	271	128
socials	413	14640	95252	4559	35.45	5.53	5967	6603	2763
Split	Lines	Words	Chars	Vocab	Avg W/L	Avg C/W	<eng>	<ita>	<lmo>
train	563	63916	384093	18931	113.53	5.02	4733	50904	9301
dev	70	7358	43289	3339	105.11	4.89	655	5425	1384
test	71	8333	49388	3687	117.37	4.94	592	6877	972
all	704	79607	476807	22284	113.08	5.00	5980	63206	11657

Table 2: Some statistics about the corpus. For each domain and split, the table gives the number of lines, the total amount of white-spaced words, the amount of characters and the size of the vocabulary. It also gives the average number of words per line and the average characters for word. The last three columns report the number of words for each language.

statistics about the composition of the corpus.

4.2. Token-Level Language Identification

To enable the identification of Lombard in code-mixed texts, such as those contained in the *Rebelòt* dataset, we developed a tool for token-level language identification. Our approach consisted of constructing a code-mixed dataset with token-level annotations and fine-tuning a pre-trained encoder-architecture language model on this data. This strategy follows a similar paradigm to the development of the Langtok model (Bednaříková and Rychlý, 2025), another tool for token-level language identification. Additionally, we aimed to investigate the feasibility of applying pre-trained models

dataset	entries	LMO	ITA	ENG	XXX
cm	563	8,925	47,826	4,627	18,420
synthetic	18,795	107,198	81,370	91,376	60,516
cm-synthetic	19,358	116,760	129,196	96,003	78,299

Table 3: Overview of the training datasets. The values represent the number of entries in each dataset and the support for the included token classes.

entry type	entries	LMO	ITA	ENG	XXX
code-mixed	71	1,002	6,449	580	2,130
Lombard	1118	44,008	-	-	10,559
Italian	1118	-	11,661	-	3,051
English	1118	-	-	12,356	2,415
total	3,425	45,010	18,110	12,936	18,155

Table 4: Statistics of the dataset used to evaluate token-level language identification performance. The values represent the number of entries in the dataset and the support for the included token classes.

to token-level identification of Lombard.

4.2.1. Synthetic Data

Although fine-tuning a pre-trained model generally requires fewer training examples than training a model from scratch, collecting a sufficient amount of code-mixed text with token-level annotations remains challenging for any language pair, especially for an under-resourced language such as Lombard. To address this limitation, we generated synthetic code-mixed data and used it for fine-tuning.

Data Sources The synthetic data were constructed from the three languages included in the *Rebelòt* dataset: Lombard, Italian, and English. The Lombard data were collected from a clean Lombard corpus from Wikipedia (Signoroni and Rychlý, 2026), while the Italian and English data were obtained from OPUS OpenSubtitles v1 (Lison and Tiedemann, 2016). The monolingual corpus used to generate the synthetic dataset comprised 11,186 sentences per language. The data were subsequently split into training, validation, and test sets.

Construction Strategy The synthetic data were constructed using a simple insertion-based strategy, similar to those proposed by Bednaříková and Rychlý (2025). The monolingual sentences were first tokenized using the `word_tokenize` function from the `nltk.tokenize` module (Bird et al., 2009). The data were then modified by extracting contiguous token spans from Lombard sentences and inserting them at random positions into English and Italian sentences. The length of the inserted spans varied between 1 and 10 tokens, and each modified sentence contained exactly one Lombard span. Admittedly, this data construction strategy does not produce syntactically well-formed sentences and therefore cannot be considered a method for generating realistic code-mixed data. However, as the results presented below suggest, it is sufficient to provide the model with enough information to learn the token-level language identification task.

4.2.2. Training Datasets

In total, we created three training datasets. The first comprised authentic code-mixed examples from the *Rebelòt* corpus. The second consisted of the synthetic data together with monolingual sentences in each of the included languages. The third combined the previous two datasets. An overview of the sizes of the training datasets is provided in Table 3. In addition to the training data, validation and test sets were created to evaluate the model's performance.

Each entry in these datasets consisted of a list of tokens representing the model input and a corresponding list of labels representing the gold annotations. In addition to the language classes *LMO*, *ITA*, and *ENG*, we introduced an additional class, *XXX*, which was assigned to tokens consisting solely of non-alphabetic characters.

The class *XXX* was introduced after inspecting the model's behavior in its absence. Without this class, the model tended to assign the label *LMO* to tokens that were neither English nor Italian, including those consisting of non-alphabetic characters. Introducing the *XXX* class mitigated this issue by encouraging the model to treat *LMO* as a distinct language label rather than a residual category for all non-English and non-Italian tokens.

The tokens in the constructed dataset correspond to the units produced by the `word_tokenize` method (Bird et al., 2009) and thus generally approximate full words. Before training, these tokens were further processed

using the tokenizer of the pre-trained model, which splits them into smaller subtoken units, and the labels were subsequently aligned to match these subtokens.

4.2.3. Choice of the pre-trained model

As the underlying model for our tool, we selected mmBERT (Marone et al., 2025). The model is based on the ModernBERT architecture (Warner et al., 2025) and was pre-trained on a multilingual corpus that includes low-resource languages such as Lombard, which is particularly advantageous for our task. It is available in two sizes: *small* (42M non-embedding parameters) and *base* (110M non-embedding parameters). In our experiments, we used the larger *base* variant.

By fine-tuning the mmBERT model on the three datasets described above, we obtained three models, which we refer to according to the datasets on which they were trained: *cm*, *synthetic*, and *cm-synthetic*.

5. Evaluation

We conducted the evaluation at both the token and sentence levels. In the following subsections, we present the datasets, methodology, and results for each evaluation level.

5.1. Token-level Language Prediction

The dataset used to evaluate the model's token-level performance consisted of code-mixing examples from the test split of the *Rebelòt* corpus (71 entries) and monolingual sentences from the test split of the corpora used to build the synthetic training data (1,118 entries per language). Detailed statistics of the test dataset are provided in Table 4. The results of the evaluation of the three different model variants are presented in Table 5, and detailed evaluation statistics of the highest-performing model are shown in Table 6.

Evaluation metrics were computed at the subtoken level, corresponding to the units produced by the model's tokenizer when splitting original tokens. As expected, the model trained on the largest dataset achieved the highest performance. However, this result may also reflect the fact that the code-mixed data in the training set and the test set share the same domain. Therefore, the dominance of the *cm-synthetic* model may not generalize to other datasets, and the *synthetic* model could prove to be a more robust solution in such scenarios.

Model	Accuracy
cm	0.907
synthetic	0.978
cm-synthetic	0.997

Table 5: Accuracy achieved by different models on token-level evaluation.

	Precision	Recall	F1
LMO	0.996	0.998	0.997
ITA	0.995	0.993	0.994
ENG	0.997	0.989	0.993
XXX	0.999	0.999	0.999
Macro	0.997	0.995	0.996
Weighted	0.997	0.997	0.997

Table 6: Detailed token-level evaluation results for the model trained on the cm-synthetic dataset.

5.2. Sentence-level Language Prediction

To assess the models' performance at the sentence level, we aggregated the token-level predictions and assigned each sentence a single language label using majority-class voting, excluding tokens labeled *XXX*. The test data consisted of the monolingual sentences from the token-level test dataset presented in Table 4.

We also compared our models with two off-the-shelf LID tools: FastText (Joulin et al., 2017a), (Joulin et al., 2017b), a widely used language identification tool, and GlotLID (Kargaran et al., 2023), which extends FastText to improve performance on low-resource languages. The results of this comparison are presented in Table 7.

As expected, FastText performed worse than GlotLID, reflecting the latter's better suitability for low-resource languages. Both tools, however, were outperformed by our models. The variant trained solely on code-mixed data performed worse than those trained on larger datasets. It is worth noting that our models have a substantial advantage over FastText and GlotLID because they predict among only four labels, whereas FastText supports over 170 languages and GlotLID supports more than 2,000.

Tool	Accuracy
FastText	0.845
GlotLID	0.970
cm	0.971
synthetic	0.996
cm-synthetic	0.997

Table 7: Accuracy achieved by different LID tools in the sentence-level evaluation.

6. Conclusions

This paper introduced *Rebelòt*, a novel multi-domain dataset manually annotated at the token level, designed to capture the complex code-mixing and code-switching dynamics between Lombard, Italian, and English. By extracting data from diverse mediums, ranging from transcribed oral folktales to modern social media posts, the corpus aims to give a glimpse into the linguistic reality and diversity of this low-resource variety. It can serve as both a resource for linguistic research and as training data for NLP applications.

The data from the *Rebelòt* corpus was subsequently utilized to develop and evaluate a novel LID tool operating on the token level. The tool was presented in three variants, the distinguishing factor of which was the amount and type of data on which they had been trained. The evaluation of the tools was conducted at the token level, with the optimal variant attaining an accuracy of over 0.99. Furthermore, an assessment was conducted to evaluate the performance of our models against two off-the-shelf LID tools (FastText and GlotLID) in sentence-level language identification, with a specific focus on the tools' capacity to identify Lombard. The performance of our custom tools was found to exceed that of the baselines, achieving an accuracy rate of over 0.99 for two of the three proposed variants.

Limitation and Future Work

A limitation of the present work lies in the representativeness of the *Rebelòt* dataset. It is now limited to just two variants of Eastern Lombard, *Bresciano* and *Bergamasco*. It is also limited in the scope of domains, with several other contexts and modalities of use left ignored. In future work, we plan to extend the corpus to ameliorate this issue and thus create a dataset more representative of code-switching and code-mixing in Lombard.

The sentence-level comparison may suggest an overly strong superiority of our models compared to FastText and GlotLID, since our tools were designed to only distinguish between Lombard, Italian, and English. In contrast, FastText and GlotLID were trained to recognize tens (hundreds) of languages. This may be considered a disadvantage when assessing their performance on a subset of the languages for which they were designed.

Ethical Considerations

We intend these resources, the corpus and the tools, as a means to aid the conservation and use of endangered, under-resourced languages, such as Lombard, even in the current age.

7. Bibliographical References

Giovanna Alfonzetti. 2015. Age-related variation in code-switching between italian and the sicilian dialect. *ATHENS JOURNAL OF PHILOLOGY*, 2:21–34.

Giulia Andreoli. 2022. In italiàn t'el dìsi. italiano e dialetto a poggiridenti (so). Master's thesis, Università degli Studi di Pavia.

Marino Anesa and Mario Rondi. 1981. *Mondo Popolare in Lombardia - Fiabe Bergamasche*, 1st edition. Silvana Editoriale, Cinisello Balsamo.

Emma Bednaříková and Pavel Rychlý. 2025. Evaluating training data construction strategies for token-level language identification. In *Recent Advances in Slavonic Natural Language Processing (RASLAN 2025)*, pages 45–53. Tribun EU.

Gaetano Berruto. 1987. Lingua, dialetto, diglossia, dilalìa. In Johannes Holtus, Günter e Kramer, editor, *Romania et Slavia adriatica. Festschrift für Zarko Muljačić*, 1st edition, pages 57–81. Buske, Hamburg.

Gaetano Berruto. 1997. Code-switching and code-mixing. In Martin Maiden and Mair Parry, editors, *The Dialects of Italy*, 1st edition, chapter 46, pages 394–400. Routledge, London.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Giovanni Bonfadini. 2010. lombardi, dialetti. In Treccani Eds., editor, *Enciclopedia dell'italiano*. Treccani.

Massimo Cerruti. 2018. Code-switching in italo-romance: a variationist study of convergence in bilingual speech. *Lingue e linguaggio, Rivista semestrale*, (1/2018):87–106.

Paolo Coluzzi, Lissander Brasca, and Simona Scuri. 2021. Revitalizing contested languages: The case of lombard. In Marco Tamburelli and Mauro Tosco, editors, *Contested Languages: The hidden multilingualism of Europe*, chapter 9, pages 163–182. John Benjamins, Amsterdam.

Paolo Coluzzi, Lissander Brasca, Marco Trizzino, and Simona Scuri. 2018. Language planning for italian regional languages: the case of lombard and sicilian. In Dieter Stern, Motoki Nomachi, and Bojan Belić, editors, *Linguistic Regionalism in Eastern Europe and Beyond: Minority, Regional and Literary Microlanguages*, pages 274–298. Peter Lang, Frankfurt am Main.

Silvia Dal Negro and Simone Ciccolone. 2018. Il parlato bilingue: Italiano e tedesco a contatto in un corpus sudtirolese. In *Lingua parlata: Un confronto fra l'italiano e alcune lingue europee*, page 23. Lang, Berlin.

Anik Dey and Pascale Fung. 2014. A hindi-english code-switching corpus. In *LREC*, pages 2410–2413.

Ilaria Fiorentini. 2017. *Segnali di contatto. Italiano e ladino nelle valli del Trentino-Alto Adige*. FrancoAngeli.

Federica Frighetto. 2025. 'mi no go parole… o forse si' - code mixing tra italiano standard e dialetto veneto nei social media. Master's thesis, Università "Ca' Foscari" - Venezia.

John J. Gumperz. 1977. The sociolinguistic significance of conversational code-switching. *RELC Journal*, 8(2):1–34.

ISTAT. 2017. L'uso della lingua italiana, dei dialetti e di altre lingue in italia. https://www.istat.it/it/archivio/207961. Accessed: 2025-08-07.

ISTAT. 2026. L'uso della lingua italiana, dei dialetti e delle lingue straniere - anno 2024. https://www.istat.it/comunicato-stampa/luso-della-lingua-italiana-dei-dialetti-e-delle-lingue-straniere-anno-2024/. Accessed: 2026-02-17.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jegou, and Tomas Mikolov. 2017a. Fasttext.zip: Compressing text classification models.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017b. 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.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Caroline Lavecchia, Kamel Smaïli, and David Langlois. 2007. Building Parallel Corpora from Movies. In *The 4th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2007*, Funchal, Madeira, Portugal.

M. Paul Lewis and Gary F. Simons. 2010. Assessing endangerment: Expanding fishmans's gids.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Michele Loporcaro. 2009. *Profilo Linguistico dei Dialetti Italiani*, 1st edition. Manuali Laterza. Editori Laterza, Bari.

Martin Maiden and Mair Perry. 1997. *The Dialects of Italy*. Routledge.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO Publishing.

Christopher Moseley and Alexandre Nicholas. 2010. *Atlas of the World's Languages in Danger*, 3rd edition, volume 19 of *Memory of Peoples*. UNESCO, Paris.

Pieter Muysken. 2000. *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Doreen Osmelak and Shuly Wintner. 2023. The denglisch corpus of German-English code-switching. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 42–51, Dubrovnik, Croatia. Association for Computational Linguistics.

Giulia Paganessi. 2017. *Brazilian Bergamasch: an Italian language spoken in Botuverá (Santa Catarina, Brazil)*. Leiden University.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Caroline Sabty, Islam Mesabah, Özlem Çetinoğlu, and Slim Abdennadher. 2021. Language identification of intra-word code-switching for arabic–english. *Array*, 12:100104.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychlý. 2026. Lombardographia: Automatic classification of lombard orthography variants.

Igor Sterner and Simone Teufel. 2023. TongueSwitcher: Fine-grained identification of German-English code-switching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–13, Singapore. Association for Computational Linguistics.

Ronald Wardhaugh. 2006. *An Introduction to Sociolinguistics*. Blackwell Publishing.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.