LombardoGraphia: Automatic Classification of Lombard Orthography Variants

Edoardo Signoroni, Pavel Rychlý
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czechia
e.signoroni@mail.muni.cz, pary@fi.muni.cz

Abstract

Lombard, an underresourced language variety spoken by approximately 3.8 million people in Northern Italy and Southern Switzerland, lacks a unified orthographic standard. Multiple orthographic systems exist, creating challenges for NLP resource development and model training. This paper presents the first study of automatic Lombard orthography classification and LombardoGraphia, a curated corpus of 11,186 Lombard Wikipedia samples tagged across 9 orthographic variants, and models for automatic orthography classification. We curate the dataset, processing and filtering raw Wikipedia content to ensure text suitable for orthographic analysis. We train 24 traditional and neural classification models with various features and encoding levels. Our best models achieve 96.06% and 85.78% overall and average class accuracy, though performance on minority classes remains challenging due to data imbalance. Our work provides crucial infrastructure for building variety-aware NLP resources for Lombard.

Keywords: Lombard, Low-Resource Languages, Orthography Classification, Language Identification, Italian Language Varieties

1. Introduction

Recent advances in Large Language Models (LLMs) have significantly improved multilingual natural language processing, enabling high-quality machine translation (MT) and other downstream tasks for many major world languages. However, the benefits of these models are unevenly distributed, with underresourced languages being left behind in both training data and model capabilities.

Italy's uniquely diverse linguistic landscape features numerous regional varieties alongside Standard Italian, which only achieved widespread adoption following the birth of mass media. Consequently, most local varieties have declined due to marginalization and social stigma. Today, over 30 Italian language varieties are endangered (Moseley and Nicholas, 2010), though scattered interest is re-emerging in NLP.

One such language is Lombard, spoken in and around the Northern Italian region of Lombardy and in parts of Switzerland. Despite having an estimated number of speakers of 3.8 million people, NLP work on Lombard and its varieties is scarce (Ramponi, 2024), barring its use in an increasingly digitalized and interconnected world. However, developing tools and resources for Lombard is not a trivial task. As a primarily spoken language existing across a continuum of intercomprehensible varieties, it lacks a common standard orthography.

While some data is already available online, a fundamental first step is to correctly recognize and classify which of the many proposed orthographies is being used in a given text, with the ultimate aim to create a substantial corpus for each variety, and for Lombard as a whole. This would enable effective and language-aware training and developing NLP models and tools for the benefit of the Lombard community, without forcing an artificial standard on the speakers or unknowingly mixing together data for different varieties, potentially hampering downstream applications. This work addresses this gap by curating both the corpus and classification models needed for variety-aware Lombard NLP.

This paper introduces LombardoGraphia (lmo_graphia)[1], a curated multi-orthography corpus, and presents the first study of automatic Lombard orthography classification. Our contributions are:

- The **LombardoGraphia Corpus**: A curated dataset of 11,186 Lombard Wikipedia samples tagged across 9 orthographic variants (MILCLASS, LOCC, LORUNIF, SL, NOL, CRES, BREMOD, BERGDUC, LSI), with train/validation/test splits suitable for supervised learning. The corpus includes detailed metadata about orthographic systems and geographic distributions.
- **Trained models for automatic Lombard orthography classification**: Training and evaluation of 24 classification models featuring both traditional and neural approaches with different feature and encoding levels, enabling variety-

[1]The corpus, trained models, and code available online at https://github.com/edoardosignoroni/lmo_graphia

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3270–3280
11–16 May 2026.

aware NLP development for Lombard, supporting applications like corpus building, language identification, and orthographic normalization, while establishing benchmarks for future work.

2. Lombard and its Orthographies

2.1. The Lombard Language

Lombard is a language variety[2] spoken in and around the Northern Italian region of Lombardy and in the Swiss Cantons of Ticino and Grisons[3] by about 3.8 million people, where it exists alongside the official language in a state of *dilalìa* (Ramponi, 2024; Berruto, 1987). Italian is used in all formal and official settings, whereas the local variety is more and more confined to informal situations, overlapping with Italian even in these domains.[4] For this reason, even if some historical literary traditions exist, Lombard varieties are primarily used in spoken and informal settings, and lack a codified written form, with the speakers improvising writing "as the words sound". When the speakers write their variety, it is often in code-switching. Lombard is promoted at a regional level in Lombardy by the Regional Law 25/2016.[6]

It belongs to the Gallo-Romance-Cisalpine group of the Western Romance family of the Indo-European languages, and it is said to have between two and four varieties, the main ones being Western (in the provinces of Varese, Como, Lecco, Sondrio, Milan, Monza, Pavia and Lodi, in addition to Novara and Verbania in Piedmont and Canton Ticino in Switzerland) and Eastern Lombard (in the provinces of Bergamo, Brescia and Northern Cremona). These varieties, even with some phonetic, lexical, and grammatical differences, can be loosely considered to be one language continuum, since they are mutually intelligible (Coluzzi et al., 2018; Bonfadini, 2010; Loporcaro, 2009; Coluzzi et al., 2021). At the present day, the language is mostly used in oral conversation and no unified orthography exists.

2.2. Lombard Orthographies

Lombard, as a primarily spoken language, does not have a written standard common to all speakers, with the majority of them either avoiding writing or using their own subjective variants. Several orthographies were proposed, such the ones used on the Lombard Wikipedia,[7] which can be divided in Pan-Lombard[8], Macro-dialectal, and Local Orthographies. Below, we will introduce the ones that are relevant for the Wikipedia dataset and this work.

2.2.1. Pan-Lombard Orthographies

Noeuva Ortografia Lombarda The *Noeuva Ortografia Lombarda* (NOL, "New Lombard Orthography")[9] is an orthography based on the writing tradition of different Lombard variants and created for the whole Lombard language. It is a polynomic system which aims to give all Lombard speakers the flexibility to write with the same rules, while allowing to keep and show local differences. It is also easy to write with a standard Italian keyboard, avoiding symbols and diacritics borrowed from foreign languages, such as the German *umlaut*.

Scriver Lombard *Scriver Lombard* (SL, "Writing Lombard")[10] is another polynomic orthography proposed by Brasca (2011). This system was inspired by older Lombard literary tradition, such as the medieval author Bonvesin de la Riva (1250-1313/15) and others. It proposes a "partially uniform" system for all speakers that still allows for minimal variations. The author states that SL has the revitalization and intergenerational transmission of Lombard as its ultimate objective.

2.2.2. Macro-Dialectal Orthographies

Urtugrafia Insübrica Ünificada The *Urtugrafia insübrica ünificada* (LOCC, "Unified Insubric Orthog-

[2] In Italy, local language varieties are often stigmatized as a sign of ignorance and lack of integration, and denoted with the negatively-charged and linguistically improper connotation of *dialetti* (dialects), implying a derivative status as "dialects of Standard Italian". As Ramponi (2024) points out, the term *language varieties* is a more neutral denomination, preventing judgment on the prestige and status of each language.

[3] Lombard, in its Bergamasque variant, is spoken also in parts of southern Brazil, being brought there by immigrants during the 19th and 20th centuries (Paganessi, 2017)

[4] In Switzerland the status of the local Lombard varieties is more vital, due to the better attitudes towards the language, with institutions such as the *CDE – Centro di dialettologia e di etnografia* ("Centre for Dialectology and Ethnography")[5] of Bellinzona doing research and maintain resources for the local Lombard variety.

[6] https://normelombardia.consiglio.regione.lombardia.it/normelombardia/accessibile/main.aspx?view=showpart&idparte=lr002016100700025ar0024a

[7] https://lmo.wikipedia.org/wiki/Wikipedia:GrafCat

[8] These are also usually polynomic, that is their structure allows for local variations to spelling, while maintaining the same pronunciation rules for all Lombard dialects.

[9] https://lmo.wikipedia.org/wiki/Noeuva_Ortografia_Lombarda; https://academiabonvesin.eu/noeuva-ortografia-lombarda/

[10] https://lmo.wikipedia.org/wiki/Scriver_Lombard; http://inlombard.eu5.net/indexLmo.html

raphy") [11] was used by the journal of the cultural association "*La Vus de l'Insübria*" ("The Voice of Insubria") for the local Lombard variant. It is a phonemic system based on Italian graphemes with diacritics to distinguish between open and closed vowels, searching for a compromise between the different local Insubric pronunciation.

Ortografía Orientàl Ünificàda The *Ortografía orientàl ünificàda* (LORUNIF, "Unified Oriental Orthography")[12] is an attempt to give one phonematic writing system to the most spoken variants of Eastern Lombard, *Bergamàsch* ("Bergamasque"), *Bresà* ("Brescian"), and *Cremàsch* ("Cremasque"). It is based on several attested traditions, in particular the system used by Canossi (1862-1943) and Melchiori (Melchiori, 1817) for Brescian, and by Zappettini (Zappettini, 1859) and the cultural association "*Dücàt de Piàsa Puntìda*" ("Duchy of Pontida Square") for Bergamasque. It thus keeps the same rules of the *Ortografìa bresàna modèrna* ("Modern Brescian Orthography") and of the *Ortografìa del Dücàt Semplificàda* ("Simplified Duchy Orthography") with minor variations.

2.2.3. Local Orthographies

Ortografia Milanesa The *Ortografia Milanesa* (MILCLASS, "Milanese Orthography"),[13] also called *classega* ("Classical") is rooted in the tradition of the Milanese literature, starting from the XVII century with the works of Carlo Maria Maggi (1630-1699). This system appears as a compromise between the Italian orthography and the French or Provencal one. With the spread of Italian, the Milanese orthography started to be considered too complicated and became obsolete, thus being referred to as "*classega*". The *Circol Filògich Milanes* ("Milanese Philological Club") adapted the classical orthography for the modern use.

Ortograféa del Dücat The *Ortograféa del Dücat* (BERGDUC, "Orthography of the Duchy")[14] is the system used by the cultural association "*Dücàt de Piàsa Puntìda*", founded in 1924. It is based on older systems from the XIX century, with the vocabularies from Zappettini (Zappettini, 1859) and Tiraboschi (Tiraboschi, 1873). It was used also by the most important Bergamasque authors of the first half of the 1900.

Grafia LSI[15] First used in 1907 to compile the *Vocabolario dei dialetti della Svizzera italiana* ("Vocabulary of the Dialects of Italian Switzerland")[16], it is based on the Italian orthography with the addition of umlauts and elements from Classical Milanese. It is maintained and developed by the *Centro di dialettologia e di etnografia* ("Centre for Dialectology and Ethnography") of Canton Ticino, where it is primarily used (together with Grigioni) for street signs and local toponyms. This makes it the only Lombard orthography with an "official" status. It is a phonologic system created for the classification and study of the different variants of Lombard in Switzerland.

Ortografìa Bresàna Moderna The *Ortografìa Bresàna Moderna* (BREMOD, "Modern Brescian Orthography")[17] is based on the system used by Angelo Maria Canossi (the most important contemporary Brescian author), even if with some variations to rationalize and solve some ambiguities. It largely uses the same rules as Italian, but employs diacritics such as accents and umlauts to distinguish between different vowel sounds and mark the tonic accent.

Urtugrafìa Cremàsca The "Urtugrafìa Cremàsca" (CRES, "Cremasque Orthography")[18] is similar to other Eastern Lombard systems in that it is largely based on Italian rules with some added diacritics to mark accents and roundedness and openness.

Table 1 gives samples of text in each orthography. The Pan-Lombard orthographies (NOL and SL) are explicitly built to be used independently by all Lombard speakers, regardless of the variant. The vast majority of the data, however, is written either in macro-dialectal or local orthographies which are by nature intertwined with the Lombard variants for which they are constructed, and used by the speakers of those specific varieties. Thus, orthography and variant are closely connected in practice, leading not only to differences in orthographical rules,

[11] https://lmo.wikipedia.org/wiki/ Urtugrafia_insübrica_ünificada *Insubria* is an historical region that during the Classical antiquity was populated by the Insubres. For a time, it denoted the western part of Lombardy, plus Ticino and the province of Novara. Today, it can also be used to refer to Milan and the surrounding territories.

[12] https://lmo.wikipedia.org/wiki/ Ortograféa_orientàl_ünificàda

[13] https://lmo.wikipedia.org/wiki/ Ortografia_milanesa

[14] https://lmo.wikipedia.org/wiki/ Ortograféa_del_Dücat

[15] https://it.wikipedia.org/wiki/ Ortografia_ticinese. The tag LSI comes from "*LSI - Lessico dialettale della Svizzera italiana*" ("Dialectal Lexicon of Italian Switzerland"), a vocabulary of the Lombard variants in Switzerland.

[16] https://www4.ti.ch/decs/dcsu/cde/pubblicazioni/vocabolario-dei-dialetti-della-svizzera-italiana

[17] https://lmo.wikipedia.org/wiki/ Ortografìa_del_Bresà

[18] https://lmo.wikipedia.org/wiki/ Urtugrafìa_Cremàsca

[19] From *La Fuggitiva* ("The Fugitive") by Tommaso Grossi (1791-1853). Retrieved from https://lmo.wikipedia.org/wiki/Dialett_milanes. Own translation.

Orthography	Sample Text
MILCLASS	*Sera settada in terra col coo in man, e i gombed sui genœucc: me ziffolava el vent in di cavij: demaneman che vegneva on quaj bôff, el me portava comè ona vôs che vegneva de lontan;* [...] "I was seating on the ground with the head in the hands, and the elbows on the knees: as it came in whiffs, it brought me as a voice from afar; [...]"
LOCC	*Sera setada in tera cul coo in man, e i gumbed süi genögg: me zifulava el vent in di cavii: demaneman che vegneva un quai buf, el me purtava cumè una vus che vegneva de luntan;* [...]
SL	*S'era setada in terra col coo in man, e i gombeds sui jencægg: me cifolava el vent ind i cavei: demaneman qe vegneva un quai bof, al me portava comè una vox qe vegneva de lontan;* [...]
NOL	*S'era setada in terra col coo in man, e i gombet sui sgenoeugg: me scifolava el vent in di cavej: demaneman che vegneva un quaj bof, el me portava comè una vos che vegneva de lontan;* [...]
LSI	*Cula paròla Ticines a sa inteend i dialett che i è parlaa in Tesin (vün dai 26 cantón svizzer) e in Mesulcina e Calanca (dó vall dal Canton Grison).* [...] "With the word *Ticines*, we mean the dialects spoken in Ticino (one of the 26 Swiss cantons) and in Mesolcina and Calanca (two valleys of Grisons)"
LORUNIF	*L'ortografìa orientàl ünificàda l'è 'n tentatìf de dàga 'na rispòsta al bizògn de 'n sistéma de scritüra bù per töte le varietà piö parlàde del Lombàrt Orientàl* [...] "The unified oriental orthography is an attempt to answer the need for a writing system for all the most spoken varieties of Eastern Lombard."
BREMOD	*Le régole dopràde endèi artìcoi marcàcc come scriìcc segónt l'ortografìa modèrna le se rifà a la grafìa del Canossi che l'è sènsa döbe l'autùr dialetàl bresà contemporàneo piö 'mportànte.* "The rules used in the articles tagged as written following the modern orthography are inspired by the orthography of Canossi who without a doubt was the most important contemporary brescian dialectal author."
BERGDUC	*L'ortograféa del Dücàt l'è ol sistéma de régole de trascrissiù del dialèt bergamasch dovrade de l'associassiù del Dücàt de Piassa Püntida in di sò püblegassiù.* "The orthography of the Duchy is the system of transcription rules of the Bergamasque dialect used by the association of the Duchy of Piazza Pontida in its publications."
CRES	*Per capì la scritüra cremàsca sèrf adóma poche régule, simìi a chèle per scrìf an italià. [...] I acént i è quasi sèmper segnàt, anche 'ndù i è mia stretaménte necesàre.* "To understand Cremasque writing only few rules are needed, similar to those for writing in Italian. [...] The accents are almost always marked, even where they are not strictly needed."

Table 1: Summary of the Lombard orthographies in the dataset. The first column give their abbreviated name, while the second shows a text sample for each one. When available (MILCLASS, LOCC, SL, NOL) we used the same text.

19

but also in lexical choices and topics. So, while at a first glance many orthographies look similar from the rules point-of-view, it is still important to distinguish between them in order to grasp subtle differences and wider linguistic implications.

3. Related Work

To our knowledge, no prior published work was done for the automatic classification of Lombard orthographies. In the following section, we thus briefly introduce Language Identification (LI) for similar variants (Subsection 3.1) and some other NLP work and resources for Lombard (Subsection 3.2).

3.1. LI for Similar Variants

This work can be framed as a language identification task, that is the problem of determining the

natural language or variety that a document is written in. While one line of research aim at broadening the number of languages supported by one single system, another one is focused on groups of similar languages, covering idioms from various groups and families. Training a system to discern between similar languages, dialects, or variants is harder than having it distinguish completely different languages. For a broader survey of LI, we refer the reader to Jauhiainen et al. (2019).

The VarDial evaluation campaigns have long served as the primary benchmark for similar language and dialect identification (Zampieri et al., 2020). Recently, the campaign explicitly addressed the linguistic landscape of Italy through the Identification of Languages and Dialects of Italy shared task introduced in VarDial 2022 (Aepli et al., 2022). Participants were tasked with classifying text across several Italian language varieties, Lombard included. Findings from the shared task demonstrated that traditional machine learning models and simple character-level neural networks frequently outperformed massive pre-trained language models on this specific problem (Ceolin, 2022). This reflects our own findings for Lombard orthographies, where classical ML approaches prove highly competitive and more robust.

Character *n*-grams have been used effectively in text categorization and LI for decades (Jauhiainen et al., 2019), since the milestone method of Cavnar and Trenkle (1994) and its off-the-shelf implementation of van Noord (1997).

3.2. NLP for Lombard

Beyond pure language identification, there is a growing recognition of the need for speaker-centric, variety-aware NLP for Italy's endangered languages (Ramponi, 2024). Recent efforts have begun to map the diatopic variation of Italy in digital spaces, such as social media corpora (Ramponi and Casula, 2023). However, most existing computational approaches still implicitly treat local languages as unified, monolithic entities with standardized writing systems. By framing orthography classification as a foundational precursor to dialect identification, we address the reality that varieties like Lombard are a linguistic continuum lack a single codified standard, a practical dimension often overlooked in broad-coverage language identification systems.

While at least some research has been done on other language varieties of Italy, NLP work explicitly for Lombard is scarce (Ramponi, 2024).

Signoroni (2022) describes a human-evaluated, revised, and corrected Lombard-Italian parallel corpus destined to train machine translation systems. With the help of bilingual annotators, they audit an automatically aligned Wikipedia corpus from OPUS (Tiedemann, 2009).

Usually if Lombard is present, it is in a multilingual setting: it is part of benchmark datasets (FLORES+ and OLDI Seed) (Costa-jussà et al., 2024), supported by some multilingual MT models (NLLB-200) (NLLB Team et al., 2022) and LLMs (mBERT) (Devlin et al., 2019).

4. Methodology

4.1. The LombardoGraphia Corpus

Data Source and Collection We collect data and text from the Lombard Wikipedia.[20] The Lombard Wikipedia suggests writing the articles in one of the pan-Lombard orthographies, the SL and NOL, but accepts also other macro-dialectal and local systems. It is strongly suggested to mark which variant is used in an article by using the corresponding template. This feature was crucial to build and process the data.

Filtering We subject the data to extensive filtering.

We first process the raw Wikipedia XML using `wikiextractor`[21] to remove markup and extract plain text.

We scan each entry to find if there is an orthography tag; which we use to categorize the article. We assume these as gold data, since they are chosen by the authors of the relative article. The rest of the articles are assigned to the "no_tag" class. We extract 295,379 total lines from the Wikipedia dump, of which 200,859 contain orthographic tags and 94,520 are untagged.

We deduplicate the resulting lines, and then we manually check the output to further remove other noise: leftover boilerplate text, sentences in a language other than Lombard,[22] short lines[23], recurring lines of bot-generated articles,[24] and the like.

This ensures that the text in the corpus is primarily in Lombard, contains substantive linguistic content, exhibits orthographic features, and has clear orthographic tags assigned by human contributors.

Corpus Description Table 2 presents the corpus composition. LombardoGraphia contains 11,186

[20] `https://lmo.wikipedia.org/wiki/Pagina_principala` We retrieved the dumps from July 2, 2025.

[21] `https://github.com/attardi/wikiextractor`

[22] Mostly Italian, English, and some text in Cyrillic alphabet.

[23] Lines with 3 words or less were mostly foreign named entities, or dates.

[24] Such as pages about years, towns, and stations, etc., e.g. *El 1901 a l'è 'n ann del secol quell de vint.* ("1901 is a year of the XX century.")

Class	Type	Total Lines	Train		Valid		Test		Dataset Total		Removed	
			N	%	N	%	N	%	N	%	N	%
MILCLASS	Local	79,196	3,606	40.29	471	42.13	446	39.89	4,523	40.43	74,673	94.29
LOCC	Macro-dialectal	34,794	2,907	32.48	345	30.86	380	33.99	3,632	32.47	31,162	89.56
LORUNIF	Macro-dialectal	76,455	1,901	21.24	240	21.47	229	20.48	2,370	21.19	74,085	96.90
SL	Pan-Lombard	3,761	174	1.94	22	1.97	16	1.43	212	1.90	3,549	94.36
NOL	Pan-Lombard	2,349	109	1.22	9	0.81	8	0.72	126	1.13	2,223	94.64
CRES	Local	633	98	1.09	17	1.52	19	1.70	134	1.20	499	78.83
BREMOD	Local	2,675	94	1.05	9	0.81	13	1.16	116	1.04	2,559	95.66
BERGDUC	Local	990	59	0.66	5	0.45	6	0.54	70	0.63	920	92.93
LSI	Local	6	2	0.02	0	0.00	1	0.09	3	0.03	3	50.00
no_tag	-	94,520	-	-	-	-	-	-	-	-	94,520	100.00
Total		**295,379**	**8,950**	**100.0**	**1,118**	**100.0**	**1,118**	**100.0**	**11,186**	**100.0**	**284,193**	**96.21**

Table 2: Distribution of orthographic classes across train, validation, and test sets. **Total Lines** shows the number of lines in the Wikipedia corpus before filtering. **Dataset Total** reports the number of lines in the cleaned corpus. **Removed** shows lines filtered out due to quality criteria (N) and the percentage removed relative to **Total Lines** (%). The no_tag category contains lines without orthography tags that were excluded from the dataset.

samples distributed across 9 orthographic classes, with imbalance reflecting the orthographic preferences of Wikipedia contributors. Major classes represent 94.01% of the data (MILCLASS, 40.43%; LOCC, 32.47%; LORUNIF, 21.19%), while minor classes are 5.96% of the examples (SL, 1.90%; NOL, 1.13%; CRES, 1.20%; BREMOD, 1.04%; BERGDUC, 0.63%). LSI has only a minimal 0.03%, with only 3 samples.

The corpus is split into training (8,950 samples, 80%), validation (1,118 samples, 10%), and test (1,118 samples, 10%) sets. The class distribution is similar across all sets.

The vast majority of the training data is in Western Lombard (MILCLASS+LOCC, 72.77%), around a quarter of the lines is in Eastern Lombard (LORUNIF+CRES+BREMOD+BERGDUC, 24.04%), while just a tiny fraction is written in a Pan-Lombard orthography (SL+NOL, 3.16%), even if the use of SL and NOL is clearly suggested by Wikipedia.

After cleaning, 96.21% of lines were filtered out. LSI was left with only 2 viable examples in the train split, thus we decided to exclude it from the experiments.

LombardoGraphia is released in JSONL format:

```
{"text": "sample text in Lombard",
"tag": "ORTHOGRAPHY_CLASS"}
```

4.2. Model Training

We train both traditional and neural classifiers on the tagged Wikipedia corpus. We train traditional models using byte, char, or word 1- to 4-grams; and a combination of all three features. Class imbalance is addressed through balanced class weighting in all classifiers. We evaluate the models on the validation set during training and choose the one with the best accuracy score. Final evaluation is performed on the held-out test set. We train 24 models in total, combining four traditional classifiers (Logistic Regression, SVM, Naive Bayes, Random Forest) and four neural architectures (LSTM, CNN, Deep CNN, Transformer) with four feature types for traditional models (character n-grams, byte n-grams, word n-grams and combined features), and two encodings for neural ones (byte and characters).

4.3. Traditional Approaches

We employ four traditional machine learning classifiers using `scikit-learn` implementations. Models use TF-IDF weighted byte, character, or word n-grams (n=1-4) with a maximum of 10,000 features per vectorizer. The byte and character-level approach captures spelling patterns without requiring explicit tokenization, while words are trivially split as white-spaced strings. Multiple feature types (character, byte, and word n-grams) can be combined by concatenating their feature vectors.

Logistic Regression We use multinomial logistic regression with L-BFGS optimization (max_iter=1000) and balanced class weights. The multinomial formulation enables direct multi-class classification across all orthographic variants simultaneously.

Support Vector Machine We employ LinearSVC with balanced class weighting (max_iter=4000). The linear kernel provides efficient training while maintaining good performance on the high-dimensional sparse feature space.

Naive Bayes We apply MultinomialNB with default scikit-learn parameters. Despite its independence assumption, Naive Bayes often performs

well on text classification with sparse features.

Random Forest We train an ensemble of 100 decision trees with balanced class weights. The ensemble approach helps capture complex patterns in orthographic variation while being robust to overfitting.

4.4. Neural Approaches

We train neural classifiers with both character-level and byte-level encoding. Character encoding uses a learned vocabulary (with special tokens for padding and unknown characters), while byte encoding uses fixed representations (0-255 plus padding at 256). Models operate on sequences of maximum length 200, with embedding dimension 128, trained using the Adam optimizer (learning rate 0.001) and batch size 128 for 10 epochs. Best models are selected based on validation accuracy and tested on the held out test set. All neural models use cross-entropy loss and implement early stopping based on validation accuracy to prevent overfitting.

Long Short-Term Memory Our bidirectional LSTM uses 256 hidden units per direction across 2 layers with 0.5 dropout. The final classification uses the concatenated last hidden states from both directions, allowing the model to capture both left and right context for each character or byte.

Convolutional Neural Network We implement two CNN architectures. The *Wide CNN* uses 3 parallel convolutional layers with different kernel sizes (3, 4, and 5) and 256 filters each, followed by max pooling and concatenation. This architecture captures different n-gram patterns simultaneously. The *Deep CNN* employs 6 stacked convolutional layers (256 filters each, alternating kernel sizes of 7 and 3) with batch normalization and progressive max pooling (stride 2).

Transformer Our transformer uses 4 layers with 8 attention heads and learned positional encoding. Dropout is set to 0.3. We apply mean pooling over non-padding tokens before classification, allowing the model to attend to relevant patterns regardless of position in the sequence.

5. Results

We evaluate all models on a held-out test set of 1,118 samples across 9 orthographic classes. Table 3 shows overall results and detailed per-class accuracies for selected top-performing models. Figure 1 visualizes the performance variation across all models.

Overall Performance The SVM classifier with combined byte, character, and word n-gram features achieves the highest overall accuracy at 96.06%, correctly classifying 1,074 out of 1,118 test samples. However, when considering average per-class accuracy, Logistic Regression with byte features performs best at 85.78%, indicating more balanced performance across all orthographies. This is relevant for downstream application, given the dataset's severe class imbalance, where the three dominant classes (MILCLASS, LOCC, LORUNIF) represent 94.36% of the test set.

Combined features (byte+char+word) generally improve overall accuracy for traditional models, with SVM and Logistic Regression both achieving their best overall performance using this combination. However, the same cannot be said for average class accuracy, which is worse or unchanged when the combined features are used.

Naive Bayes is significantly worse than other traditional methods, due to its nearly complete inability to correctly classify the minority classes.

Among neural models, the CNN architectures achieve the highest overall accuracy (94.27% for byte-level, 94.18% for character-level), approaching the performance of traditional methods. However, all neural architectures show substantially lower average class accuracy compared to Logistic Regression and SVM, since they unsurprisingly struggle with minority classes. The best neural model (CNN with byte encoding) achieves only 60.21% average class accuracy, compared to 85.78% for Logistic Regression with the same features.

Per-Class Performance Performance varies across orthographic classes, as shown in Table 3. The three major classes (MILCLASS, LOCC, LORUNIF) achieve consistently high accuracy across nearly all models, with most configurations exceeding 90%. The best performance on MILCLASS reaches 98.88% (Naive Bayes with byte features), while LOCC peaks at 96.58% (SVM with combined features) and LORUNIF at 99.13% (both SVM variants with byte+char+word and byte features).

Minority classes show extreme performance variation and much lower overall accuracy. The accuracy range (difference between best and worst model) quantifies this variation: SL, NOL, CRES, BREMOD, and BERGDUC all show ranges of 0.69-1.0, meaning the best model achieves 69-100 percentage points higher accuracy than the worst. For these low-resource classes, Logistic Regression and SVM with byte or combined features consistently outperform neural methods.

Multiple models achieve 0% accuracy on minority classes: Naive Bayes fails completely on SL, NOL, CRES, BREMOD, and BERGDUC; all neural models fail on NOL and BREMOD; Deep CNN and Transformer achieve 0% on several classes. Neural models require substantially more training data to learn distinctive orthographic patterns than

Model	MILCLASS	LOCC	LORUNIF	SL	NOL	CRES	BREMOD	BERGDUC	Overall	Avg Class
Traditional Models - Logistic Regression										
Log. byte	97.31	92.89	89.08	**100.0**	**75.00**	94.74	53.85	**83.33**	93.38	**85.78**
Log. byte+char+word	97.31	94.74	94.76	**100.0**	**75.00**	**100.0**	46.15	66.67	95.08	84.33
Log. char	95.07	93.42	84.28	**100.0**	**75.00**	94.74	53.85	**83.33**	91.67	84.96
Log. word	94.62	94.21	88.65	93.75	37.50	94.74	**69.23**	66.67	92.39	79.92
Traditional Models - Support Vector Machine										
SVM byte	97.76	94.21	**99.13**	87.50	**75.00**	94.74	30.77	50.00	95.43	78.64
SVM byte+char+word	97.76	**96.58**	**99.13**	93.75	50.00	94.74	23.08	50.00	**96.06**	75.63
SVM char	98.21	94.74	98.25	93.75	50.00	94.74	30.77	50.00	95.52	76.31
SVM word	95.52	95.53	95.20	93.75	12.50	89.47	30.77	50.00	93.73	70.34
Traditional Models - Naive Bayes										
NB byte	**98.88**	91.05	93.45	0.00	0.00	0.00	0.00	0.00	89.62	35.42
NB byte+char+word	98.21	93.42	96.07	0.00	0.00	0.00	0.00	0.00	90.69	35.96
NB char	98.65	91.05	94.32	0.00	0.00	0.00	0.00	0.00	89.70	35.50
NB word	97.09	92.37	94.76	12.50	0.00	15.79	0.00	0.00	90.06	39.06
Traditional Models - Random Forest										
RF byte	98.43	92.37	96.94	68.75	0.00	68.42	0.00	0.00	92.75	53.11
RF byte+char+word	98.65	92.11	95.63	75.00	0.00	68.42	7.69	0.00	92.66	54.69
RF char	98.65	93.16	95.20	50.00	0.00	68.42	7.69	0.00	92.57	51.64
RF word	97.31	87.89	93.89	68.75	0.00	63.16	15.38	0.00	90.24	53.30
Neural Models										
CNN byte	97.98	95.53	97.38	75.00	12.50	78.95	7.69	16.67	94.27	60.21
CNN char	98.21	94.47	98.69	87.50	12.50	63.16	7.69	16.67	94.18	59.86
Deep CNN byte	83.86	94.74	92.14	75.00	12.50	73.68	15.38	0.00	87.20	55.91
Deep CNN char	97.09	92.37	96.94	0.00	12.50	57.89	7.69	0.00	91.23	45.56
LSTM byte	97.53	91.05	95.63	0.00	0.00	63.16	0.00	0.00	90.60	43.42
LSTM char	98.21	91.05	95.63	0.00	0.00	68.42	0.00	16.67	91.05	46.25
Transformer byte	96.64	85.00	94.76	68.75	0.00	5.26	0.00	0.00	88.00	43.80
Transformer char	96.41	88.95	93.89	0.00	0.00	31.58	0.00	0.00	88.54	38.85
Performance Statistics										
Best accuracy	**98.88**	**96.58**	**99.13**	**100.0**	**75.00**	**100.0**	**69.23**	**83.33**	**96.06**	**85.78**
Worst accuracy	83.86	85.00	84.28	0.00	0.00	0.00	0.00	0.00	87.20	35.42
Accuracy range	15.02	11.58	14.85	100.0	75.00	100.0	69.23	83.33	8.86	50.36

Table 3: Complete per-class accuracy (%) across all 24 models, with overall and average class accuracy. Bold indicates the best performance per metric. Classes are ordered by sample size. Overall accuracy measures performance on all test samples, while high average class accuracy shows consistent performance across classes.

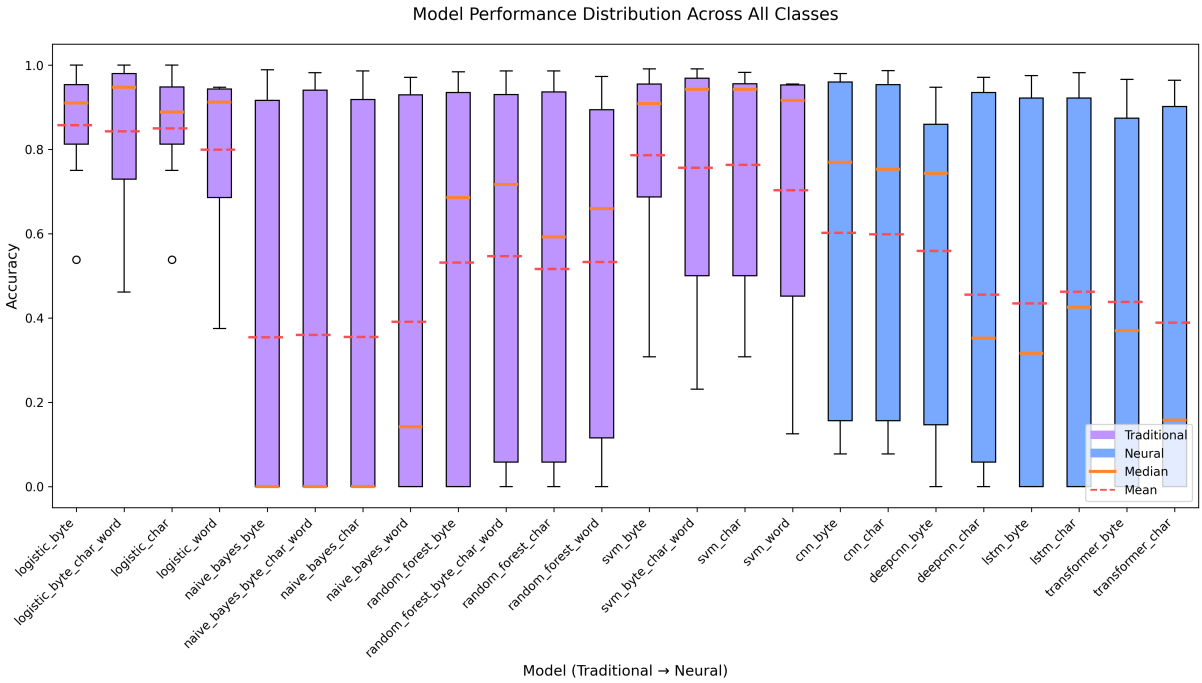

Figure 1: Boxplot of the results for all models. Traditional models are in purple, and neural models are in blue.

traditional approaches.

Feature Encoding Comparison For traditional models, byte-level features often perform competitively with or better than character-level features on balanced metrics: Logistic Regression with byte features achieves 85.78% average class accuracy versus 84.96% for character features and 84.33% for combined features. Word-level features, when used alone, underperform other encoding types, but contribute to improved performance when combined with character and byte features in traditional models. Word features with logistic regression lead to the best accuracy for BREMOD. Also worth noting, combining all features significantly improves performance on the same classifier for LORUNIF.

For neural models, the choice between byte and character encoding has minimal impact on overall performance (CNN: 94.27% vs 94.18%, LSTM: 90.60% vs 91.05%), though character-level encoding shows slight advantages for some classes.

Confusion Analysis We identify systematic confusion patterns common across all models. The most frequent error is misclassifying LOCC as MILCLASS (56 total errors across all models), which is unsurprising given both are Western Lombard orthographies with overlapping graphemes and features. The reverse confusion occurs less frequently (24 times).

Eastern Lombard orthographies also show expected mutual confusion. BREMOD is misclassified as LORUNIF (13 errors). The pan-Lombard orthographies (SL, NOL) are misclassified as MILCLASS when errors occur. This pattern may result from multiple factors: insufficient training examples to distinguish their distinctive features; lexical influence, as many SL and NOL articles are likely written by Western Lombard speakers; and the fact that both orthographies are fundamentally based on MILCLASS, with SL being more distinctive, while NOL remains closer to the classical milanese standard. This is also exacerbated by the fact that models are incentivized to classify dubious examples as MILCLASS, being it the most frequent class in the training data.

Looking at the output of the `Log.byte` model, qualitative analysis on classification errors regarding Eastern Lombard orthographies (the hardest ones for the model) shows that confusion arises due to the high degree of similarity between BERGDUC, BREMOD, and LORUNIF. In many cases of LORUNIF false positives, the text contained named entities, such as toponyms, or not enough orthographical information to allow for meaningful distinction. On the other hand, when LORUNIF is classified as either BERGDUC or BREMOD, the model may be influenced by lexical choices. This may be also the reason behind the almost perfect performance on CRES for logistic regression and SVM models. While on the surface CRES is very close to BERGDUC, BREMOD, and LORUNIF, lexical differences (e.g. *al* vs BERGDUC *ol* vs BREMOD *el*; masculine singular "the") may drive the models to tag it correctly.

We apply the same `Log.byte` model on the 94,520 untagged and unfiltered Wikipedia lines. Some of the patterns already discussed are observable also here. The resulting distribution is heavily skewed towards Western Lombard, with MILCLASS and LOCC being 81.0% of the data. The overall prediction confidence is relatively low, with an average score of 0.33. This is due to the brevity of the untagged samples: short sequences, such as foreign place names, lack the necessary orthographic and linguistic features to allow for high-confidence classification. This drives the model to assign them to the most frequent MILCLASS tag.

6. Conclusions

This paper presented LombardoGraphia, a curated corpus for the Lombard language, tagged by its main orthography variants. We used this dataset to experiment with and train traditional and neural models to automatically detect the variety of a given Lombard text. We make available both of these resources, to further NLP research for the Lombard language.

While our models achieve near perfect accuracy on majority classes, and common confusion patterns are still between very similar orthographies, improving the performance on minority classes is also desirable, if our models were to be used to build datasets for further research.

Thus, some work is still left for the future: a wider dataset to increase the coverage of minority classes and the models' accuracy for their classification; testing a document-level approach to classification; and apply these models to digitalized books and other sources, possibly with different orthographic variants. The classification itself can be improved by experimenting with other approaches, such as clustering and multi-label classification, and by adding an option for non-Lombard text.

These issues not withstanding, we believe that the dataset and the models in this work represent a critical stepping stone toward variety-aware NLP for Lombard.

Limitations

Our work has some limitations. First, our dataset is relatively small and imbalanced. While this is sufficient for most models to correctly classify the majority classes, minority ones are still very hard, especially for neural models. Second, we are limited to study readily available data from Wikipedia.

Even if these data are convenient and with reasonably good coverage, they exclude text that could be found in everyday life and published books, most often written in subjective orthographies, which can be both very similar or completely different from the ones used on Wikipedia.

Ethical Considerations

NLP, especially when underresourced and endangered languages are involved, should be primarily concerned to work for and with the speakers. With this work, we hope to provide Lombard speakers and language activists with a useful tool that can further the presence and vitality of the Lombard language.

We are aware that our models are not perfect, especially when dealing with underrepresented orthographic variants, and thus their predictions should be taken with reasonable caution.

Acknowledgements

We would like to acknowledge the volunteers contributing to and maintaining the Lombard Wikipedia without whom these data would not exist. We would like to thank the reviewers for their useful comments. This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062.

Bibliographical References

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

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.

Giovanni Bonfadini. 2010. lombardi, dialetti. In Treccani Eds., editor, *Enciclopedia dell'italiano*. Treccani, online at https://www.treccani.it/enciclopedia/dialetti-lombardi_(Enciclopedia-dell%27Italiano)/#Studi.

Lissander Brasca. 2011. *Scriver Lombard*, 1st adj. edition. Menaresta, Monza.

William Cavnar and John Trenkle. 1994. N-gram-based text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*.

Andrea Ceolin. 2022. Neural networks for cross-domain language identification. phlyers @Vardial 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 99–108, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: a survey. *J. Artif. Int. Res.*, 65(1):675–682.

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

Giovan-Battista Melchiori. 1817. *Vocabolario Bresciano-Italiano*. Franzoni e socio.

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

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.

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.

Antonio Tiraboschi. 1873. *Vocabolario dei Dialetti Bergamaschi Antichi e Moderni*. Editrice Fratelli Bolis.

Gertjan van Noord. 1997. Textcat.

Marcos Zampieri, Preslav Nakov, and Yves Scherrén. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595-612.

Stefano Zappettini. 1859. *Vocabolario Bergamasco-Italiano per ogni classe di persone e specialmente per la gioventù*. Pagnoncelli, Bergamo.

Language Resource References

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.

Alan Ramponi and Camilla Casula. 2023. Diatoplt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Edoardo Signoroni. 2022. Piötòst ché nièn, mèi piötòst-a manually revised lombard-italian parallel corpus. *RASLAN 2022 Recent Advances in Slavonic Natural Language Processing*, page 105.

Jörg Tiedemann. 2009. News from opus - a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248.