Low- vs High-level Lemmatization for Historical Languages. A Case study on Italian

Chiara Alzetta[1,*,†], Simonetta Montemagni[1,†]

[1] Istituto di Linguistica Computazionale "Antonio Zampolli", Consiglio Nazionale delle Ricerche, Pisa, Italy

Abstract

Lemmatization remains a foundational yet challenging task in the processing of historical Italian texts, due to the complex interplay of orthographic, morphological, and diatopic variation. A crucial, yet often overlooked, aspect is the degree of normalization applied during lemmatization. A conservative approach preserves attested historical forms, ensuring greater linguistic fidelity but increasing data sparsity. Conversely, an abstract normalization strategy aligns historical variants with standardized contemporary lemmas, improving generalization but potentially introducing inaccurate mappings. In this paper, we present a comparative evaluation of conservative and normalized lemmatization strategies for historical Italian. To our knowledge, this is the first study to explicitly assess the impact of lemmatization strategies in the context of historical languages, particularly those that are morphologically rich. Our results indicate that high-level normalization offers a promising trade-off between precision and generalization.

Keywords

Data-driven Lemmatization, Historical Italian, Universal Dependencies, Normalization

1. Introduction

Lemmatization is the task of identifying the canonical form, or *lemma*, of a given inflected wordform. While this mapping is often straightforward and based on well-established criteria, it can also involve a considerable degree of discretion, especially in the case of diachronic language data. In historical lexicography, lemma selection remains a well-known and unresolved challenge due to the high number of attested variant forms, many of which diverge significantly from the standard form. Choosing a specific lemma to serve as the headword — i.e. capable of effectively subsuming all its variants — is a widely debated issue. As Porter and Thompson [1] and Manolessou and Katsouda [2] have noted, it constitutes a genuine *dilemma*. In computational linguistics, by contrast, lemmatization criteria are rarely made explicit and are often taken for granted. While this may pose only minor issues in the lemmatization of contemporary language, it becomes a critical concern for historical language data. This paper investigates the role and impact of different lemma identification strategies in automatic lemmatization, with a focus on historical varieties.

Lemmatization is one of the fundamental tasks that facilitate downstream Natural Language Processing (NLP) applications and is particularly relevant for highly inflected languages. Traditionally, this task has been addressed using rule-based morphological analyzers and dictionary lookup. However, recent years have seen the rise of data-driven lemmatization approaches, where models learn to produce lemmas without relying on predefined linguistic rules and/or lexical resources. A key turning point in this methodological shift was the SIGMORPHON 2016 Shared Task, which reconceptualized lemmatization as a special case of morphological reinflection (Cotterell et al. [3]). This view paved the way for the current dominant approaches, based on neural models.

Within the data-driven paradigm, two main strategies have emerged. The generative character-level approach relies on encoder-decoder architectures that generate the lemma character by character, conditioned on the input form and its context (Qi et al. [4], Bergmanis and Goldwater [5]). In contrast, pattern-based models treat lemmatization as a supervised classification task (Straka [6]), where each class - derived from training data - corresponds to the edit operations that transform a specific wordform into its lemma. A comparative study on Estonian by Dorkin and Sirts [7] found that generative encoder-decoder models trained from scratch outperform both rule-based systems and pattern-based models fine-tuned from large pre-trained language models.

Among the most debated issues in lemmatization, particularly in data-driven models, there is the role of context and morphological information. Contextual information has been shown to be crucial for handling unseen and ambiguous words: see, among others, Bergmanis and Goldwater [5, 8] and McCarthy et al. [9]. The actual role of morphological information in performing contextual lemmatization was investigated by Toporkov and Agerri [10], who showed that fine-grained morphological information does not help to substantially improve

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author: Chiara Alzetta

[†] These authors contributed equally.

chiara.alzetta@cnr.it (C. Alzetta); simonetta.montemagni@cnr.it (S. Montemagni)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-11-30

lemmatization (not even for highly inflected languages) and that using basic part-of-speech tags (UPOS) seems to be enough for comparable performance across languages.

Although much progress has been made on lemmatization for standard, resource-rich languages, the task remains challenging in the case of historical varieties, especially for morphologically complex languages like Italian. Historical Italian presents both orthographic and morphological variation, not only over time but often in the same period and even within the same text. These challenges include, among others: alternations between etymological and phonetic spellings (e.g., *haveva* vs. *aveva* ’(it) had’, *chupola* vs. *cupola* ’dome’); phonetic variation (e.g. *pulito* vs. *polito* ’clean’, *eguale* vs. *uguale* ’equal’); morphologically distinct variants (e.g. *avria* vs. *avrebbe* ’(it) would have’); cliticized finite verbal forms (*aveagli* ’(it) had-to-him’, *avevalo* ’(it) had-it’). Additional challenges, also relevant to contemporary Italian, include the treatment of past participles (verbal vs. adjectival use) and derivative forms (the open issue is whether they represent an independent lemma or should be associated with the corresponding base form, e.g. the diminutive *angioletto* ’little angel’ is an independent lemma or should be lemmatized as *angelo* ’angel’).

A crucial but often neglected aspect of lemmatizing historical texts concerns the granularity and scope of the lemma list, as well as the criteria guiding lemma identification: in other words, the degree of normalization applied. This choice carries both theoretical and practical implications, influencing how linguistic variation is represented, how lexical continuity over time is interpreted, and how effectively the data can be searched, analyzed, or aligned across sources. Table 1 contrasts a conservative lemmatization approach - which preserves the graphical, phonological, and morpho-syntactic features of attested historical variants - with a more abstract normalization strategy that aligns such variants to a standardized contemporary (meta-)lemma. While the former offers greater linguistic precision and interpretability, it may lead to increased data sparsity. The latter, by contrast, reduces sparsity and facilitates generalization, though at the risk of introducing incorrect form–lemma associations.

The choice between these strategies is shaped by several practical factors, including the target application and the specific language involved. Linguistic analyses, for instance, may benefit from a conservative approach, whereas information retrieval systems and downstream NLP applications may perform better with normalized lemmas. Language-specific features also play a key role. As Manjavacas et al. [11] note, the highly heterogeneous nature of historical languages — marked by overlapping diachronic and diatopic variation and the absence of a stable standardized norm — makes it particularly challenging to carry out lemmatization and normalization simultaneously. In the case of diachronic Italian, a low-level lemmatization strategy was adopted by Favaro et al. [12, 13], deferring normalization to a later stage operating on lemma variants.

wordform	**Conservative Lemma**	**Normalized Lemma**	**POS**
brieve	BRIEVE	BREVE	ADJ
sanctissimo	SANCTO	SANTO	ADJ
chotesto	COTESTO	CODESTO	DET
alma	ALMA	ANIMA	NOUN
imperadori	IMPERADORE	IMPERATORE	NOUN
palagio	PALAGIO	PALAZZO	NOUN
utilitati	UTILITATE	UTILITÀ	NOUN
admettesse	ADMETTERE	AMMETTERE	VERB
diliberarono	DILIBERARE	DELIBERARE	VERB
guarentir	GUARENTIRE	GARANTIRE	VERB
surse	SURGERE	SORGERE	VERB

Table 1
Examples of conservative vs normalized lemmatization for historical Italian

In this paper, we present a comparative evaluation of these two lemmatization strategies for historical Italian, combining quantitative metrics with qualitative analysis. To our knowledge, this issue has not yet been explicitly addressed in the computational linguistics literature, where lemmatization choices are typically assumed rather than critically examined. We argue that this decision is especially relevant for morphologically rich languages, where different lemmatization strategies can have a substantial impact on both the performance and interpretability of downstream tasks.

The rest of the paper is organized as follows. In Section 2, the historical corpora selected as the basis of this study are described. Section 3 illustrates the strategy adopted for generating a version of these corpora with high-level normalized lemmatization. Section 4 describes the approach employed to train two models for lemmatizing Italian historical texts. Section 5 discusses the results obtained by the lemmatization models, focusing both on the results obtained in five-fold cross-validation experiments and against an external test set. Finally, Section 6 concludes the paper and presents some future prospects.

2. Data

For this study, we selected three corpora covering a wide timespan, going from the 14th to the 20th century, listed below:

- **UD-Italian Old** [14]: Italian-Old is a treebank containing Dante Alighieri’s Comedy, based on the 1994 Petrocchi edition and sourced from the DanteSearch corpus [15]. The treebank includes lemmatization, morpho-syntactic, and syntactic

Corpus	**Sentences**	**Tokens**
UD-Italian Old	3,419	122,038
GDLI-QC - GDLI Quotation Corpus	1,500	36,624
VGG - Voci della Grande Guerra	4,945	108,208
Total	9,864	266,870

Table 2
Size of the used corpora of historical Italian

annotation. A partial manual revision was carried out to align morpho-syntactic annotation and lemmatization with the Universal Dependencies (UD) guidelines, with particular attention to proper nouns and fixed multiword expressions. For our experiments, we used version 2.15 of the treebank, released in November 2024;

- **VGG - Voci della Grande Guerra** [16]: VVG is a corpus of texts that were written in Italian in the period of World War I or shortly afterwards (most of them date back to the years 1915-1919). The corpus includes different textual genres, namely: discourses, reports, and diaries of politicians and military chiefs; letters written by men and women, soldiers and civilians; literary works of intellectuals, poets, and philosophers; writings of journalists and lawyers. The corpus is annotated at the morpho-syntactic level and lemmatized. Annotation was carried out with UDPipe [17] trained on IUDT [18]v2.0; a subset was then manually revised [19]. For this study, we used the gold portion of the corpus;
- **GDLI-QC - GDLI Quotation Corpus** [12]: GDLI-QC is a corpus derived from an authoritative historical Italian dictionary, namely the *Grande dizionario della lingua italiana* (GDLI) edited by Salvatore Battaglia. GDLI presents a huge collection of quotations covering the entire history of the Italian language, from which a subset has been extracted, representative of the most cited authors and covering a wide chronological span (from the 14th to the 20th century). GDLI-QC has been morpho-syntactically tagged and lemmatized with Stanza [4]: annotation was carried out automatically, with full manual revision.

All of these corpora follow a conservative lemmatization strategy. In terms of annotation, they are all natively annotated according to the Universal Dependencies (UD) scheme[1] (De Marneffe et al. [20]), which has become the *de facto* standard nowadays. Lemmatization has been manually revised for each corpus — albeit only partially for UD-Italian Old — to ensure linguistic accuracy and internal consistency. As such, these corpora can be considered gold-standard resources. Table 2 provides details on their size in terms of sentences and tokens.

For the comparative study of the two lemmatization strategies, a normalized counterpart of each corpus, featuring high-level linguistic annotation, was required. To generate the normalized versions of the three corpora, we identified two historical Italian lexicons adopting this lemmatization approach.

One such resource is the MIDIA lexicon, which was built starting from the balanced diachronic corpus of written Italian texts called MIDIA (D'Achille and Grossmann [21]), fully annotated with lemma and part-of-speech (POS) information. Covering the period from the early 13th to the first half of the 20th century, the corpus is organized into five chronological periods and seven textual genres, comprising approximately 7.5 million tokens drawn from about 800 texts. In MIDIA, lemmatization and POS tagging were automatically performed using a version of TreeTagger (Schmid [22]) adapted for historical Italian (Iacobini et al. [23]). To handle the linguistic variation typical of earlier stages of the language, the contemporary Italian lexicon embedded in TreeTagger was enriched with approximately 230,000 word forms, primarily dating from the 14th to the 16th centuries. This substantially expanded the original MIDIA lexicon. The version we used contains 70,083 unique lemmata, 571,779 distinct wordform–lemma pairs, and 584,041 unique wordform–lemma–POS triples. Notably, there is a high degree of overlap between the wordform–lemma pairs from the corpora under study and those in the MIDIA lexicon: 89.91% for UD-Italian Old, 86.65% for GDLI-QC, and 81.66% for VGG.

Another key reference resource identified for these purposes is the *Tesoro della Lingua Italiana delle Origini* (TLIO) (Beltrami [24]), a historical dictionary of old Italian based on all extant documentation from the earliest texts recognizable as Italian up to the end of the 14th century, which includes manual lemmatization.

To fully understand the type of lemmatization performed in these two resources, we report below the set of wordforms sharing the nominal lemma AMMINISTRAZIONE 'administration' in the MIDIA and TLIO lexicons:

> **MIDIA**: *administratione, administrationi, administrazione, aministratione, amministratione, amministrationi, amministrazione, amministrazioni, nistrazione, strazione*
>
> **TLIO**: *adminestragione, administracion, administracione, administraciuni, administragione, administratione, administrationi, administrazione, aministracione, aministraciuni, aministragione, aministrascione, aministratione, amministracione, amministragione, amministragioni, amministratione, amministrazione, amministrazioni*

[1] https://universaldependencies.org/

MIDIA POS	UD POS	LEGEND
ART,POSS, DEMO,INDEF	DET	Determiner
PRE	ADP	Adposition
NPR	PROPN	Proper noun
ADV,NEG	ADV	Adverb
ARTPRE	ADP+DET	Articulated Prep.
VER	VERB	Verb
AUX	AUX	Auxiliary
CON	CCONJ,SCONJ	Conjunction
DEMO, INDEF, PRO, CLI	PRON	Pronoun
ADJ	ADJ	Adjective
NOUN	NOUN	Noun
PUN, SENT	PUNCT	Punctuation
CHE	PRON,SCONJ	
NUM	NUM	Numeral
WH	PRON,ADV,SCONJ	Interrogative

Table 3
Mapping between MIDIA and UD part of speech tags

3. Lemma Normalization

To carry out lemma normalization, the first step consisted of converting the part of speech tags of the MIDIA lexicon to the UD annotation scheme. Table 3 details the correspondences between the two tagsets. The conversion was carried out automatically, and the ambiguous underspecified cases (e.g. CHE and WH tags) were then revised manually.

The normalization process of the selected corpora was carried out in three successive phases, relying on lexicon-based validation and correction. The objective was to verify and, where appropriate, normalize wordform-lemma (WL) pairs extracted from the selected historical corpora using the MIDIA and TLIO historical lexicons.

In the first phase, each WL pair was checked against the MIDIA lexicon. If the WL pair was found in MIDIA, the case was marked as `f1-match-found` and left unchanged. If the wordform was present in the MIDIA lexicon but was associated with a different lemma, or with both a different lemma and POS, the unmatching information was modified with the values appearing in MIDIA (case marked as `f1-modified-lemma` or `f1-modified-lemma+pos`). If the *wordform* was not found in MIDIA, the case was labeled `f1-form-missing` and passed as input to the second phase.

In the second normalization phase, the wordforms labelled as missing (i.e. `f1-form-missing`) in MIDIA during Phase 1 were re-analyzed. For these cases, we checked whether MIDIA contained the lemma matching any other form. If the POS in the corpus and MIDIA lexicon coincided, then we marked the case as correct using the label `f2-validated-lemma`. If the lemma was present in MIDIA with a different POS, the original POS from the corpus was preserved, and the case was labeled `f2-different-pos`. If no matching form or lemma was found in MIDIA, the case was labeled `f2-missing`.

The final phase addressed the remaining unresolved cases from Phase 2 — those labeled `f2-missing` and `f2-different-pos` — by consulting the TLIO lexicon. As a first step, we checked whether the triple (wordform, lemma, POS) was present in the lexicon. If so, we marked the case as validated (`f3-valid-lemma-F`), or modified the lemma to match the triple in TLIO (`f3-modified-lemma-F`). If the *lemma* appeared as a *wordform* in TLIO with the same POS, the lemma was changed to match the lemma reported in TLIO (`f3-modified-lemma-L`) or validated against the lexicon (`f3-valid-lemma-L`). If the *form* was present but associated with a different POS, the case was labeled `f3-different-lemma-pos`. If none of the above conditions applied, the case remained unresolved and was labeled `f3-missing`.

Table 4 exemplifies the cases treated in the different normalization steps, reporting the corpus annotation and how it was revised based on the evidence of the MIDIA / TLIO lexicons.

For each step described above, Table 5 reports the distribution of cases in the three normalization steps. For the three historical corpora, the number of matching WL pairs is very high: the lemmatization in the corpus and the lexicon coincided in more than 96% of the cases (with minor differences across the corpora). Cases normalized during one of the three phases amount to 3.56% in the UD-Italian Old, 3.02% in VGG, and 2.97% in GDLI-QC. A neglectable number of cases were not normalized, ranging from 0.09% in the UD-Italian Old, to 0.85% and 0.73% in VGG and GDLI-QC respectively.

4. Model Training

For the analysis of historical Italian texts, we trained the Stanza natural language processing neural pipeline [4], developed by the Stanford NLP Group. Stanza, following a generative character-level approach, offers a modular architecture with state-of-the-art models for tokenization, lemmatization, part-of-speech tagging, morphological analysis, dependency parsing, and named entity recognition. Built on a Python interface, it supports over 70 human languages and is trained on UD treebanks. In addition to its pre-trained models, Stanza allows users to train custom models from scratch using UD-formatted data. In this study, we specifically focused on the lemmatization component.

The lemmatization model was trained using the normalized versions of the selected historical corpora — UD-Italian Old, VGG, and GDLI-QC — as input data. To these, we added the contemporary Italian corpus ISDT (Italian

Label	Corpus (wordform, lemma, POS)	Lexicon (wordform, lemma, POS)	Change Description
Phase 1, Lexicon: MIDIA			
f1-match-found	(proposta, proposta, NOUN)	(proposta, proposta, NOUN)	No changes are made; the triple matches the lexicon.
f1-modified-lemma	(altipiano, altopiano, NOUN)	(altipiano, altipiano, NOUN)	The lemma in the corpus is corrected to match the lexicon.
f1-modified-lemma+pos	(esuberanti, esuberare, VERB)	(esuberanti, esuberante, ADJ)	Both lemma and POS are corrected to align with the lexicon.
f1-form-missing	(prevvede, prevedere, VERB)	–	The form is missing from the lexicon and flagged for review.
Phase 2, Lexicon: MIDIA			
f2-validated-lemma	(com’, come, ADV)	(come, come, ADV)	The corpus triple is validated despite form variation; lemma and POS match the lexicon.
f2-different-pos	(rassicurantissime, rassicurante, ADJ)	(rassicurante, rassicurare, VERB)	The same form appears in the lexicon with a different lemma and POS; the corpus POS is retained for further analysis.
f2-missing	(fidenti, fidente, ADJ)	–	The form and lemma are absent from the lexicon and marked as missing.
Phase 3, Lexicon: TLIO			
f3-valid-lemma-F	(accecamento, accecamento, NOUN)	(accecamento, accecamento, NOUN)	The triple is validated; it matches the lexicon entry.
f3-modified-lemma-F	(disolate, disolato, ADJ)	(disolate, desolato, ADJ)	The lemma is corrected to align with the TLIO lexicon.
f3-modified-lemma-L	(adirizar, adirizare, VERB)	(adirizare, addirizzare, VERB)	The triple is normalized using the lemma assigned to the variant in the lexicon.
f3-valid-lemma-L	(succian, succiare, VERB)	(succiare, succiare, VERB)	The triple is validated; the lemma is found in the lexicon with matching POS.
f3-different-lemma-pos	(ubbriachi, ubbriaco, ADJ)	(ubbriaco, ubriaco, NOUN)	Lemma and POS differ from the lexicon; no change is applied.
f3-missing	(addobbamenti, addobbamento, NOUN)	–	Both the form and lemma are missing from the lexicon; no change is made.

Table 4
Normalization examples for each phase.

Label	UD-Italian Old	GDLI-QC	VGG
f1-match-found	117,586 (96.35%)	35,270 (96.3%)	104,071 (96.13%)
f1-modified-lemma	1,888 (1.55%)	515 (1.41%)	156 (0.14%)
f1-modified-lemma+pos	196 (0.16%)	92 (0.25%)	85 (0.08%)
f2-validated-lemma	579 (0.47%)	177 (0.48%)	2,325 (2.15%)
f2-different-pos	43 (0.04%)	53 (0.14%)	66 (0.06%)
f3-modified-lemma-L	3 (0%)	5 (0.01%)	29 (0.03%)
f3-valid-lemma-L	102 (0.08%)	23 (0.06%)	105 (0.1%)
f3-modified-lemma-F	563 (0.46%)	96 (0.26%)	35 (0.03%)
f3-valid-lemma-F	560 (0.46%)	59 (0.16%)	57 (0.05%)
f3-different-lemma-pos	410 (0.34%)	68 (0.19%)	408 (0.38%)
f3-missing	2062 (0.69%)	266 (0.73%)	924 (0.85%)

Table 5
Distribution of cases across the three normalization steps for each source.

Stanford Dependency Treebank) (Bosco et al. [18]). For comparison purposes, we also trained a model using the original, non-normalized versions of the historical corpora. In the remainder of this paper, we refer to the model trained on normalized data as NORM_Lem, and to the one trained on unnormalized original data as ORIG_Lem.

To evaluate the performance of the NORM_Lem and ORIG_Lem models, we conducted two sets of experiments, each with a distinct objective. The first set was designed to assess the impact of low-level versus high-level normalization on lemmatization accuracy (Section 5.1). For this purpose, we performed 5-fold cross-validation: in each fold, the dataset was divided into a training set (containing 14,419 sentences, corresponding to the 80% of the full dataset), a validation set (4,806 sentences, 10%), and a test set (4,806 sentences, 10%). As detailed in Table 6, the internal composition of the validation and test sets was representative of the four different corpora used for training in similar proportions.

The second set of experiments aimed to evaluate the accuracy and robustness of the normalized lemmatization model on an external historical corpus (Section 5.2). In this case, the model was trained on the entire dataset and tested on a selection of sentences from the MIDIA corpus, which had been semi-automatically converted into the UD format. This evaluation allowed us to test

		ISDT		Italian-Old		GDLI		VGG	
Fold	Set	Sents	Toks	Sents	Toks	Sents	Toks	Sents	Toks
1	dev	61.55	53.58	14.58	20.84	2.11	6.72	21.76	18.86
	test	60.84	52.19	14.74	21.67	2.08	6.25	22.34	19.90
	train	62.66	52.79	15.20	21.96	0.73	6.55	21.41	18.71
2	dev	62.26	53.15	14.86	21.48	2.14	7.20	20.75	18.17
	test	61.55	53.58	14.58	20.84	2.11	6.72	21.76	18.86
	train	62.17	52.48	15.16	22.05	0.73	6.20	21.94	19.27
3	dev	61.94	52.66	15.15	21.82	2.06	6.43	20.84	19.08
	test	62.26	53.15	14.86	21.48	2.14	7.20	20.75	18.17
	train	62.05	52.79	14.96	21.75	0.73	6.25	22.25	19.21
4	dev	61.18	52.35	14.95	22.14	2.14	5.81	21.73	19.70
	test	61.94	52.66	15.15	21.82	2.06	6.43	20.84	19.08
	train	62.41	53.14	14.93	21.49	0.74	6.75	21.92	18.61
5	dev	60.84	52.19	14.74	21.67	2.08	6.25	22.34	19.90
	test	61.18	52.35	14.95	22.14	2.14	5.81	21.73	19.70
	train	62.78	53.15	15.07	21.67	0.74	6.77	21.41	18.41

Table 6
Composition of folds (percentage of sentences and tokens).

the generalizability of the NORM_Lem model beyond the data it was trained on.

5. Lemmatization Results

5.1. Low- vs High-level Normalization Results

The first set of experiments was conducted using 5-fold cross-validation. The NORM_Lem and the ORIG_Lem models were tested on the normalized and original versions of the treebanks respectively. Table 7 presents the accuracy scores for each fold, as well as for the entire DEV and TEST sets. In all cases, the NORM_Lem model consistently outperforms the ORIG_Lem model, both across individual folds and on average. A reduction in the number of incorrectly lemmatized tokens is observed for source corpora, with the most notable improvement in the UD-Italian Old corpus, where NORM_Lem yields a 0.38% decrease in lemmatization errors on both the DEV and TEST sets. An exception to this trend is GDLI-QC, for which both models show a slight drop in accuracy (−0.18 on both DEV and TEST). The VGG corpus is less affected by normalization, showing a reduction in lemmatization errors of 0.11%.

We also analysed the results by part-of-speech (POS). Table 8 reports the error rates in the TEST set. Aside from NUM (numerals), which is the worst-performing category with an increase of errors with the NORM_Lem model, the POS with the highest error rates (above 3%) are ADJ, VERB, and PROPN, followed by NOUN and PRON, with error rates of 2.37% and 1.87% respectively. All other POS categories show error rates below 1%. Errors involving ADJs and VERBs are mainly ascribable to the ambiguous use of past participles, which often alternate between verbal and adjectival function, a frequent source of lemmatization errors. As for NOUNs, the observed errors may also be linked to the treatment of derived forms, whose lemmatization may not always be consistent across treebank sources. Regarding NUM, the category with the highest error rate, we noted that most errors involve Roman numerals, often misinterpreted as PROPN.

ORIG_Lem model		
Fold	Lemma Acc. (DEV)	Lemma Acc. (TEST)
Fold 1	0.9827	0.9830
Fold 2	0.9817	0.9829
Fold 3	0.9824	0.9821
Fold 4	0.9830	0.9825
Fold 5	0.9828	0.9826
Average	**0.9825**	**0.9826**
NORM_Lem model		
Fold	**Lemma Acc. (DEV)**	**Lemma Acc. (TEST)**
Fold 1	0.9851	0.9841
Fold 2	0.9841	0.9847
Fold 3	0.9852	0.9835
Fold 4	0.9852	0.9841
Fold 5	0.9847	0.9851
Average	**0.9848**	**0.9843**

Table 7
Lemma accuracy obtained with the ORIG_Lem and the NORM_Lem models over 5-fold cross-validation on DEV and TEST portions.

POS	ORIG_Lem	NORM_Lem	Note
ADJ	4.42	3.95	<
ADP	0.29	0.30	=
ADV	2.01	0.38	<
AUX	0.12	0.12	=
CCONJ	0.18	0.18	=
DET	0.76	0.75	<
NOUN	2.63	2.37	<
NUM	**0.43**	**0.48**	**>**
PRON	2.19	1.87	<
PROPN	3.66	3.61	<
PUNCT	0.18	0.18	=
SCONJ	0.23	0.22	<
VERB	3.88	3.63	<

Table 8
Percentage of erroneously lemmatized tokens by POS, obtained by the ORIG_Lem and the NORM_Lem models on the TEST sets.

5.2. Testing NORM_Lem with an External Historical Corpus

In the second set of experiments, we focused on the NORM_Lem model with the aim of evaluating its accuracy and robustness on an external historical corpus. The test set comprises a selection of sentences from the MIDIA corpus, for a total of 5,116 tokens. The sentences are acquired from ten different texts to ensure diversity in terms of genre and period of composition. In fact, the texts span a broad chronological range, from the early 14th century to the mid-19th century, thus offering a representative sample of linguistic variation across different evolution stages of the Italian language. In terms of genre distribution, the dataset includes three subsets of expository essays, three of scholarly or scientific texts, two of literary prose texts, and two of personal correspondence. This selection, which includes textual genres not represented in the training corpus, aims to evaluate the robustness of the NORM_Lem model in the face of stylistic, genre, and diachronic variation.

The overall lemmatization accuracy achieved by the NORM_Lem model on the external test set is 96.59%. While this score is slightly lower than the average accuracy obtained in the 5-fold cross-validation experiment described above, such a difference is expected given that the test set comprises previously unseen texts that partially differ both in genre and chronological coverage from the training data. The slight performance drop reflects the increased difficulty posed by domain shift, particularly with respect to historical variation (in this MIDIA sample there are periods which are not covered in the training corpus) and text type.

A closer analysis of the accuracy of lemmatization over time, shown in Figure 1, reveals that the performance remains relatively stable over the centuries, with significantly high values, ranging from 93.58% to 97.44%. The lowest accuracy is observed for the text dated 1505 by Leonardo Da Vinci (93.58%). However, this drop seems more related to the complexity and idiosyncrasies of the text's genre (i.e., technical and fragmentary scientific notes) rather than to its chronological distance. Excluding this outlier, lemmatization accuracy across the remaining texts shows limited variance, with most scores clustering around 96–97%, indicating the robustness of the model to diachronic variation.

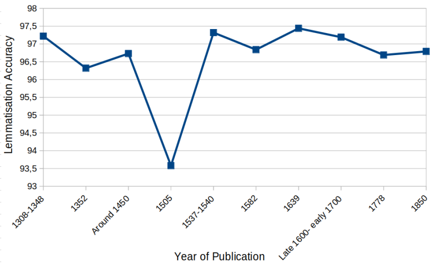

Figure 1: Lemmatization accuracy for different periods in the MIDIA test.

The genre-based evaluation further confirms this trend. The model performs best on personal correspondence and expository texts, achieving in both cases an accuracy of 96.94%, closely followed by literary prose (96.87%). Slightly lower accuracy is recorded for scientific texts (95.88%), very likely due to genre-specific linguistic characteristics, such as technical terminology, irregular syntax, and less standardized spelling. However, the performance remains consistently high across all genres, confirming the generalizability of the NORM_Lem model to different types of historical texts.

An analysis of lemmatization errors by part-of-speech (POS) on the external test set (Table 10) reveals patterns that are largely consistent with those observed in the five-fold evaluation, while also highlighting genre- and domain-specific challenges. As in the internal evaluation, ADJ, VERB, and PROPN remain among the POS with the highest error rates, recording values of 9.59%, 6.71%, and 6.80%, respectively, in the full test set. These results confirm the persistent difficulty posed by adjectives and verbs, often due to the ambiguous status of past participles that can function both as verbal and adjectival forms. Errors in the PROPN category remain notably high, particularly in scientific texts (21.43%). However, this result should be interpreted with caution, as it is influenced by the low frequency of proper nouns in these texts. Although the proportion of incorrectly lemmatized proper nouns appears substantial, the scientific subcorpus contains only 14 PROPN tokens in total. This small sample size limits their overall impact on the test set and may inflate the observed error rate due to sampling effects. ADV,

POS	Expos.	Letters	Lit. Prose	Science	All Test
ADJ	5.83	6.67	6.49	16.26	9.59
ADP	0.48	0	0.63	0	0.29
ADV	1.96	3.17	2.7	0.8	1.83
AUX	0	0	0	0	0
CCONJ	0	0	0	2.35	0.68
DET	0	1.05	0	2.27	1
INTJ	0	0	0.65	0	0
NOUN	6.83	6.82	5.73	6.3	6.41
NUM	10	0	0	0	2.70
PRON	4.92	5.56	6.67	3.7	4.88
PROPN	7.69	5.56	3.95	21.43	6.80
PUNCT	0	0	0	0	0
SCONJ	2.7	3.85	0	0	1.50
VERB	6.67	4.46	7.08	7.85	6.71
Global	3.06	3.06	3.13	4.12	3.41

Table 9
Percentage of erroneously lemmatized tokens by POS and by genre obtained by the NORM_Lem model against the MIDIA test set.

Genre	Wrong	Correct
Letters	0.25	0.75
Lit.Prose	0.30	0.70
Science	0.35	0.65
Expositive	0.35	0.65
All	0.32	0.68

Table 10
Percentage of wrong and correct lemma predictions by genre in Out-of-vocabular words.

SCONJ, and DET also show minor fluctuations in accuracy, but their overall contribution to the global error rate remains limited. Errors in NOUN lemmatization reveal a range of recurrent challenges, including both lexical variation and morphological ambiguity. Several errors involve orthographic variants or archaic spellings that are typical of historical texts, such as *uppinione* lemmatized as UPPINIONE (instead of OPINIONE), or phonological or dialectal interference, e.g. *ariento* lemmatized as such instead of ARGENTO. Other errors highlight semantic or derivational mismatches, where the model fails to associate the inflected form with the appropriate lemma. For example, the wordform *diletti* is incorrectly lemmatized as DILETTARE (VERB) rather than DILETTO (NOUN). Finally, some errors involve mislemmatization due to homography or syntactic ambiguity, as seen, e.g., with *mostra* lemmatized as MOSTRARE, where the model incorrectly assumes a verbal or adjectival interpretation. Such cases may be tied to the POS-lemmatization interaction, where contextually ambiguous forms are resolved incorrectly, possibly due to inconsistent POS-tag/lemma alignments in training data.

Interestingly, NUM errors are less prominent in the external test set compared to the five-fold validation, likely due to the lower frequency of Roman numerals or a more predictable usage context. Other categories such as ADP, CCONJ, and AUX remain highly stable, with error rates below 1%, suggesting that closed-class words are generally well handled by the model, even in previously unseen texts.

Overall, the distribution of errors confirms the robustness of the NORM_Lem model across POS categories, while also emphasizing the influence of genre-specific lexical and morphological variation, particularly in scientific and early modern texts.

Last but not least, we analyzed how the NORM_Lem model handles the challenge of Out-Of-Vocabulary (OOV) words — i.e., words not included in the pre-trained vocabulary — which typically lead to degraded model performance. The results reported in Table 10 are consistent with our previous observations: the highest percentage of incorrect predictions is found in Science and Expository texts (35%). This percentage decreases to 30% in Literary Prose and to 25% in Letters. We further examined the incorrect predictions by part of speech (POS), revealing that the most problematic categories are still NOUNs (30%), VERBs (27%), ADJECTIVEs (22%), and PROPER NOUNs (5%), which together account for 84% of the errors in OOV words. A closer inspection of individual cases suggests that there is still room for improvement: several errors are due to case mismatches, while others involve derivative formations.

6. Conclusion and Future Work

This paper has addressed the role and impact of different lemma definition strategies in automatic lemmatization, with a particular focus on historical language varieties. Specifically, we presented a comparative study of two lemmatization strategies for historical Italian: a conservative approach and a normalized one. The model trained on normalized data (NORM_Lem) was compared to a counterpart trained on unnormalized corpora, i.e. following a conservative lemmatization approach (ORIG_Lem). Both models were evaluated intrinsically via five-fold cross-validation. Results consistently favored the NORM_Lem model, which outperformed ORIG_Lem across all folds, achieving higher accuracy and reducing the number of incorrectly lemmatized tokens.

To further evaluate the effectiveness and generalization capacity of the NORM_Lem model, we tested it on an external dataset including textual genres and historical periods not represented in the training data. Although overall accuracy on this out-of-domain test set was slightly lower — due to domain and temporal variation — the model maintained strong generalization capabilities, with stable lemmatization accuracy across different historical periods. From a genre-specific perspective, lower accuracy was observed in scientific texts, where challenges such as domain-specific terminology and Latinized proper names were more prominent. A detailed POS-based error analysis confirmed that adjectives, verbs, and proper nouns remain problematic, often due e.g. to morphological ambiguity or derivational complexity. These findings align with previous observations on the limitations of character-based neural models in capturing morpho-syntactic regularities in low-frequency or irregular data, especially in historical language varieties.

Overall, our results provide empirical evidence that high-level normalized lemmatization improves the performance of data-driven models applied to morphologically rich and orthographically variable languages like historical Italian. In particular, high-level normalization emerges as a valuable preprocessing step for lemmatization tasks involving historical corpora. However, the trade-off between normalization and linguistic fidelity should be carefully considered, especially in philological or interpretative contexts where access to attested variants is essential.

Future work will explore hybrid approaches that combine normalization with variant-aware lemmatization strategies, potentially through multitask learning or post-lemmatization clustering techniques. Another promising direction involves assessing the impact of different lemmatization strategies on downstream tasks — such as information retrieval, syntactic parsing, or historical named entity recognition — in order to evaluate their broader utility within practical NLP pipelines.

Acknowledgments

We gratefully acknowledge the support of the project *CHANGES – Cultural Heritage Innovation for Next-Gen Sustainable Society* (PE00000020), funded under the NRRP program of the Italian Ministry of University and Research (MUR) and financed by the European Union through NextGenerationEU. Furthermore, we express our sincere gratitude to the team who designed, developed, and currently maintains the MIDIA corpus, and, in particular, to Claudio Iacobini for his great support. Last but not least, we thank Felice Dell'Orletta and Alessio Miaschi for their precious suggestions in designing the experiments, and Elisa Guadagnini for her helpful comments on lemmatization criteria of historical Italian.

References

[1] N. A. Porter, P. A. Thompson, Lemmas and dilemmas: Problems in old english lexicography (dictionary of old english), International Journal of Lexicography 2 (1989) 135–146.

[2] I. Manolessou, G. Katsouda, On Lemmas and Dilemmas again: Problems in Historical Dialectal Lexicography, Brill, 2024, pp. 298–326.

[3] R. Cotterell, C. Kirov, J. Sylak-Glassman, D. Yarowsky, J. Eisner, M. Hulden, The SIGMORPHON 2016 shared Task—Morphological reinflection, in: M. Elsner, S. Kuebler (Eds.), Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 10–22.

[4] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2020.

[5] T. Bergmanis, S. Goldwater, Context sensitive neural lemmatization with Lematus, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1391–1400.

[6] M. Straka, UDPipe 2.0 prototype at CoNLL 2018 UD shared task, in: D. Zeman, J. Hajič (Eds.), Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 197–207.

[7] A. Dorkin, K. Sirts, Comparison of current approaches to lemmatization: A case study in Estonian, in: T. Alumäe, M. Fishel (Eds.), Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), University of Tartu Library, Tórshavn, Faroe Islands, 2023, pp. 280–285.

[8] T. Bergmanis, S. Goldwater, Data augmentation for context-sensitive neural lemmatization using inflection tables and raw text, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4119–4128.

[9] A. D. McCarthy, E. Vylomova, S. Wu, C. Malaviya, L. Wolf-Sonkin, G. Nicolai, C. Kirov, M. Silfverberg, S. J. Mielke, J. Heinz, R. Cotterell, M. Hulden, The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection, in: G. Nicolai, R. Cotterell (Eds.), Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, Association for Computational Linguistics, Flo-

rence, Italy, 2019, pp. 229–244.

[10] O. Toporkov, R. Agerri, On the role of morphological information for contextual lemmatization, Computational Linguistics 50 (2024) 157–191.

[11] E. Manjavacas, Á. Kádár, M. Kestemont, Improving lemmatization of non-standard languages with joint learning, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1493–1503.

[12] M. Favaro, E. Guadagnini, E. Sassolini, M. Biffi, S. Montemagni, Towards the creation of a diachronic corpus for italian: A case study on the gdli quotations, in: Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, 2022, pp. 94–100.

[13] M. Favaro, M. Biffi, S. Montemagni, Pos tagging and lemmatization of historical varieties of languages. the challenge of old italian, Italian Journal of Computational Linguistics (IJCoL) 9 (2023) 99–120.

[14] C. Corbetta, M. C. Passarotti, F. M. Cecchini, G. Moretti, Highway to hell. towards a universal dependencies treebank for dante alighieri's comedy, in: Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30—Dec 02, 2023, Venice, Italy, CEUR-WS, 2023, pp. 1–8.

[15] M. Tavoni, Dantesearch: il corpus delle opere volgari e latine di dante lemmatizzate con marcatura grammaticale e sintattica, in: Lectura Dantis 2002-2009. Omaggio a Vincenzo Placella per i suoi settanta anni, volume 2, Università degli Studi di Napoli" L'Orientale", Il Torcoliere-Officine …, 2012, pp. 583–608.

[16] F. Boschetti, I. De Felice, S. Dei Rossi, F. Dell'Orletta, M. Di Giorgio, M. Miliani, L. C. Passaro, A. Puddu, G. Venturi, N. Labanca, A. Lenci, S. Montemagni, "voices of the great war": A richly annotated corpus of italian texts on the first world war, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association (ELRA), 2020, pp. 911—-918.

[17] M. Straka, J. Hajič, J. Straková, UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), 2016, pp. 4290–4297.

[18] C. Bosco, S. Montemagni, M. Simi, Converting Italian treebanks: Towards an Italian Stanford dependency treebank, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 61–69. URL: https://aclanthology.org/W13-2308.

[19] I. De Felice, F. Dell'Orletta, G. Venturi, A. Lenci, S. Montemagni, Italian in the trenches: linguistic annotation and analysis of texts of the great war, in: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Accademia University Press, 2018, pp. 160–164.

[20] M.-C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational linguistics 47 (2021) 255–308.

[21] P. D'Achille, M. Grossmann, Per la storia della formazione delle parole in italiano: un nuovo corpus in rete (MIDIA) e nuove prospetive di studio, Franco Cesati Editore., 2017.

[22] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of International Conference on New Methods in Language Processing, 1994, pp. 1–9.

[23] C. Iacobini, A. De Rosa, G. Schirato, Part-of-speech tagging strategy for midia: a diachronic corpus of the italian language, in: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014, Pisa University Press, 2014, pp. 213–218.

[24] P. G. Beltrami, Il tesoro della lingua italiana delle origini (tlio), in: Italia linguistica anno Mille, Italia linguistica anno Duemila: atti del XXXIV Congresso internazionale di studi della Società di linguistica italiana (SLI), Firenze 19-21 ottobre 2000.-(Pubblicazioni della Società linguistica italiana; 45), Bulzoni, 2003, pp. 1000–1004.

Declaration on Generative AI

During the preparation of this work, the author(s) did not use any generative AI tools or services.