ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 9, Number 2
december 2023

aA Accademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255000945

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_9_2

IJCoL

Volume 9, Number 2
december 2023

CONTENTS

POS Tagging and Lemmatization of Historical Varieties of Languages. The Challenge of Old Italian *

Manuel Favaro**
ILC-CNR, Pisa

Marco Biffi†
Accademia della Crusca

Simonetta Montemagni‡
ILC-CNR, Pisa

The paper discusses the challenges of POS tagging and lemmatization of historical varieties of Italian, and reports for both tasks the results of experiments carried out in a classical supervised domain adaptation scenario using the diachronic and typologically differentiated corpus built for the "Vocabolario Dinamico dell'Italiano Moderno" (VoDIM). For what concerns POS tagging, the effectiveness of retrained models is illustrated and substantiated with quantitative data, with a specific view to linguistic annotation results obtained with respect to specific language evolution stages, domains and textual genres. For lemmatization, different customized models have been developed, including lexicon-assisted ones and models retrained with historical annotated texts. In both cases, a detailed error analysis is provided.

1. Introduction

The literature on Natural Language Processing (NLP) for Digital Humanities (DH) typically deals with historical texts. "Historical" is the keyword that qualifies texts, corpora, and language varieties within publications and events focused on NLP for DH (Piotrowski 2012). Historical texts encompass both texts in classical languages such as Latin, Greek, or Biblical Hebrew, i.e. autonomous languages, and texts testifying historical varieties of a given language, i.e. coherent sets of linguistic elements (forms, structures, features, etc.) that tend to appear in conjunction with specific extralinguistic conditions, defined by diachronic, but also diastratic, diatopic, diafasic, and diamesic variables. The scope of language varieties addressed in DH extends beyond the dimensions of linguistic variation listed above to also include different textual genres, linguistic registers and styles.

* **Author Contributions:** Conceptualization: Manuel Favaro, Marco Biffi, Simonetta Montemagni; methodology: Manuel Favaro, Marco Biffi, Simonetta Montemagni; writing - original draft preparation: Manuel Favaro (Section 3, 4, 5), Marco Biffi and Simonetta Montemagni (Section 1 and 6), Simonetta Montemagni (Section 2); writing - review and editing: Manuel Favaro, Marco Biffi, Simonetta Montemagni.
** Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa, Italy.
E-mail: manuel.favaro@ilc.cnr.it
† Accademia della Crusca, Italy. E-mail: marco.biffi@unifi.it
‡ Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa, Italy.
E-mail: simonetta.montemagni@ilc.cnr.it

From the NLP perspective, the processing challenges associated with the two classes of "historical" texts differ significantly. This paper focuses on the second class, representing the most common but also insidious case, which involves dealing with varieties of language for which automatic linguistic annotation tools already exist, but refer to a language variety differing from the one to be dealt with: typically, these tools have been trained on contemporary newswire language, which is thus taken to be the standard language, as opposed to so-called non-standard language varieties, which also include diachronic ones[1].

Due to the recent digitization of large volumes of historical texts, there is an increasing need for the automatic analysis of historical varieties of language use: over the last decade, there have been several attempts to efficiently deal with this kind of data in computer applications. The challenges to be tackled for developing language technologies for historical texts are twofold. On the one hand, historical varieties of language typically exhibit considerable variation at different levels, ranging from spelling to lexicon, morphology and syntax, which vary not only across time but also genres and authors of the same period, and even within the same text. On the other hand, diachronic language varieties are often under-resourced with regard to annotated data needed for training NLP tools.

Depending on the resources available (both tools and datasets) on the one hand, and the languages to be dealt with on the other hand, two main approaches can be distinguished in the literature on the automatic linguistic analysis of historical varieties of language, namely:

1. use of NLP tools developed for the modern language, with adaptation of the input data;
2. use of NLP tools specifically adapted to the task.

Under the first approach, the distance between standard training data and historical test data is reduced by normalizing the input text. Normalization is proposed as a solution to one of the key challenges of NLP on historical texts: spelling variation. Normalization is carried out as a pre-processing step, before the application of NLP tools, that automatically maps historical variant spellings to a single, contemporary normalized form.

The alternative approach takes a reversed perspective: rather than adapting historical texts to "fit" existing NLP tools, the problem is tackled by adapting the tools to fit the language testified in the text. In machine learning, test and training data are typically assumed to share the same underlying distribution. However, in practice, this assumption often does not hold, resulting in a decline in performance when a model trained on a source language variety (typically contemporary) is tested against a different but related (e.g. historical) target variety. This gives rise to the problem generically referred to in the literature as "domain adaptation", which deals with adapting the NLP model from a training distribution to a different distribution attested in the test corpus. Research on historical language processing has primarily focused on supervised domain adaptation; in this classical setup, there is a small amount of labeled target data available along with a larger amount of labeled source domain data. In recent years, neural approaches

1 As Plank (2016) claims, there are no reasons for considering newswire texts as more standard or more canonical than other text types: simply, it seems that what is considered canonical is mostly due to a historical coincidence and motivated largely by the availability of resources.

to unsupervised domain adaptation in natural language processing (NLP) have been increasingly explored, even extending to historical languages, yielding interesting results. In such cases, the necessity for labeled target domain data is obviated, as learning is solely derived from unlabeled target data. This type of data is generally accessible for both source and target domains.

In this paper, we focus on the linguistic annotation of different varieties of Old Italian. In particular, we address key research questions related with POS tagging and lemmatization of historical language varieties of Italian. The study covers the period from Italy's unification in 1860 until contemporary language, a time span relatively short, but the peculiar history of the Italian language renders it sufficient to bring to light the issues and challenges associated with processing historical language varieties.

The creation of customized models for POS tagging and lemmatization of selected old Italian texts has been carried out in a classical supervised domain adaptation scenario. The contribution of this paper can be summarised as follows. For what concerns POS-tagging, the effectiveness of the customized models is illustrated and substantiated with quantitative data, with a specific view to the impact and role on linguistic annotation results of the specific language evolution stage, domain and textual genre. For lemmatization, different customized models have been developed, including lexicon-assisted ones (using gold and/or automatically constructed morphological lexicons) and models retrained with historical annotated texts. In both cases, a detailed error analysis is provided, highlighting the peculiarities of the results obtained with the different models with specific attention to the classification of unknown words.

The paper is organized as follows. Section 2 surveys related work, with a specific view to the processing of historical varieties of Italian. Section 3 describes the corpus composition and organization, as well as the method followed for the annotation. Sections 4 and 5 report and discuss the results of the annotation experiments carried out on the test corpora that were selected for the evaluation of the models built, respectively for POS tagging and lemmatization. Finally, Section 6 draws conclusions about NLP of historical varieties of Italian, by also indicating ongoing and future developments.

2. Historical Language Processing

2.1 Challenges

Before introducing the different strategies proposed in the literature for dealing with historical varieties of language, let's briefly survey the challenges to be tackled, ranging across various levels, with a specific view to Italian.

One of the most evident and fundamental differences between modern and historical texts or even between different synchronic varieties (especially in old times) concerns spelling. Piotrowski (2012) distinguishes between diachronic and synchronic spelling variation, with the former resulting from linguistic change over time, and the latter referring to different spellings co-occurring in the same period, even within the same text. Synchronic variation typically occurs in periods when spelling is still fluctuating: orthography only became standardized in many languages fairly recently. Consider, as an example, the alternation between etymological and phonetic spellings in old Italian (e.g., *haveva* vs. *aveva* for '(s)he had', or *chupola* vs. *cupola* for 'dome').

At the lexico-semantic level, new words constantly enter a language's vocabulary (e.g., through derivation or borrowing), typically to name new concepts or technological innovations, while other words fall out of use: e.g. neologisms obtained through derivation like *buonismo* 'do-goodery' from *buono* 'good', or through borrowing like

chattare 'to chat' from the English *chat*, or words that have fallen out of use like *donzello* for 'young man of noble family'. Another area of word-level variation is represented by processes of semantic change, such as semantic extension (e.g., metaphorical or metonymic extensions), or semantic specialization.

In morphology, change typically involves the appearance or disappearance of morphological categories or distinctions, or the rules underlying the use of morphemes (e.g. by analogy). In all cases, we encounter new forms or obsolete forms that are not part of the morphological repertoires used by NLP tools (e.g., *enno* for *sono* 'they are', formed by analogy to *hanno* 'they have'). In particular, the higher the diachronic distance between variants, the higher the probability of identifying alternations and polymorphism, especially in verbal morphology (Mengaldo 1987; Antonelli 2003). Among the morphological variants attested in the texts of post-unification Italian, i.e. the VoDIM period, it is worth mentioning here thematic alternations such as *chieggo/chiedo* 'I ask', or inflectional variation, e.g. in past participles such as *concesso/conceduto* 'granted'.

Finally, at the syntactic level, the most relevant phenomena concern the order of constituents (e.g., subject or object relative to the verb, or adjective relative to the modified noun, etc.), but also the way the sentence is constructed, leading to sentences e.g. of varying length, characterized by more or less frequent use of subordination. Some examples from the corpus *Voci della Grande Guerra* (Lenci et al. 2020) testifying some of the main features of the Italian language of the early XX^{th} century follow: the length of sentences is, on average, significantly longer than contemporary Italian (25.08 vs 21.04 tokens per sentence), with text types such as memoirs and essays showing a much higher average value, respectively 35.41 and 31.65 tokens per sentence; subordinative constructions, often recursively embedded, are widely used (in letters and diaries they represent more than 40% of the clauses), whereas in contemporary Italian the recourse to subordination is more limited, especially for what concerns embedded structures; the average distance between the head and the dependent, calculated in terms of tokens, is longer with respect to contemporary Italian (i.e. 3.36 vs 2.67); the relative order of subject and verb varies, and the corpus records greater variability in terms of constituent ordering, as testified by numerous post-verbal subjects (as in *Dichiarino* ***essi*** *di accettare il terreno di discussione* 'They declare to accept the ground for discussion').

The presence of words not included in the reference lexical repertoires of automatic analysis tools, regardless of their nature as spelling or morphological variants, or lexical variants like archaisms, geosynonyms, neologisms, etc., constitutes a distinctive feature of historical texts. These words often introduce complexity, contributing to the failure to recognize the correct grammatical category of the form in the specific context, with unavoidable consequences at the level of the lemma identification. It is also often the case that, despite the correct identification of the grammatical category, the appropriate lemma cannot be reconstructed. In addition to the challenges related to the recognition and classification of individual words, the consideration of syntactic variability, more difficult to trace and quantify, is also crucial.

Pennacchiotti and Zanzotto (2008) report the results of an exploratory study on the difficulties arising from the automatic processing of historical varieties of the Italian language. To this end, they constructed a corpus gathering Italian texts from the XIIIth to the end of the XIXth century, reporting dictionary coverage in terms of the number and percentage of words attested in contemporary Italian dictionaries used as a reference. The analysis of the percentages of recurring forms in various texts covered by the reference dictionary shows low lexical coverage, ranging from 19.9% for G. Battista Basile's texts to 55.8% for Giuseppe Parini's. The average dictionary coverage for historical texts is 44%, 19% lower than the value recorded for contemporary journalistic

Italian. Similar results emerge in relation to the accuracy of morphological and morpho-syntactic annotation: for historical texts, the average accuracy is lower by 22% and 24%, respectively, compared to journalistic Italian.

In light of these data, the question that arises concerns the typology of words not covered by the reference dictionary. Let's compare, as an example, the different forms attested for the verb *avere* 'to have' in historical texts from the XVIIth and XIXth century. In the *Galileian Textual Corpus* collecting the correspondence of 1633, *avere* is realized by 91 different forms, including spelling variants (e.g. *abbia* vs. *habbia*; *haver* vs. *aver*), morphological variants (e.g. *havrà / avrà* vs. *harà*; *havrebbero / avrebbero* vs. *havrebbono*), and a variety of forms with clitics. On the other hand, in the collection of *Periodici Milanesi* from the first half of the XIX^{th} century (De Stefanis Ciccone, Bonomi, and Masini 1984), *avere* is realized through a significantly larger number of different forms (120), still characterized by a wide range of variation but with significant differences: while spelling variants have almost disappeared, morphological variants abound (e.g., *avrebber / avrebbero / avrebbono*) and, in particular, verbal forms (including finite forms) with clitics (e.g., *aveagli*, *avevalo*), which represent more than a third of the type forms. The nearly two centuries that separate these texts justify the different distribution of different types of variation, but leave unchanged the fact that the attested forms of the verb *avere* go well beyond the repertoire of standard forms.

The examples of variation reported above do not exhaust the typology of words not covered by reference dictionaries. These also include lexical variants corresponding to archaisms, neologisms, as well as dialectal forms or terminology of a specific domain. We report below, by way of example, some cases recorded in the *Voci della Grande Guerra* corpus, which collects texts of different genres and linguistic registers from the period of the First World War (De Felice et al. 2018): obsolete forms rarely used in contemporary Italian (e.g., *costì*, *tardanza*); literary forms, such as *pelago* and *nocumento*; variants of current forms and/or lemmas, such as *comperare* for *comprare*, *spedale* for *ospedale*; diatopically marked forms, typical of a regional variety of Italian like *cocuzza* or *mencio*, or dialectal forms like *batajun* or *preive*. In addition to these, there are graphical variants of contemporary forms (such as *pei* for *per i*, *pur troppo* for *purtroppo*) that also have an impact on sentence segmentation.

2.2 Solutions

Given the challenges connected with the processing of historical varieties of language sketched above with a specific view to the Italian language, let us now turn to the responses offered by the language technology research community for different languages, including Italian. The possible solutions range from creating an annotation component "from scratch" by manually or semi-automatically annotating a corpus of the target language variety to be used for training, to extending its coverage to deal with the target historical variety. In this paper we focus on the latter, which can be achieved in different ways: by "modernizing" the spelling of the historical texts to more closely match the modern spelling and then use a tagger trained for the modern language; by expanding the lexicon with historical forms; or by extending the training corpus of a modern tagger with an annotated sample representative of the target historical variety. The applicability of the approaches listed above clearly depends on the availability of resources: a major challenge in developing language technology for historical text is that diachronic language varieties are typically under-resourced with regard to needed available data. Also the properties of the historical target language (e.g. its linguistic distance from the modern language) play a key role in identifying the most appropriate

solution. Given the complementarity of the different approaches, it is often the case that a combination of them is proposed with promising results.

Spelling normalization of historical texts represents the dominant solution for dealing with historical varieties of languages: besides improving text search, it is used as a pre-processing step for improving linguistic annotation results. Spelling normalization involves mapping historical spellings to their canonical forms in modern languages, thus bridging the gap between contemporary training corpora and target historical texts. A recent literature review on converting historical spelling to present-day spelling (Bollmann 2019) proposes five categories in which modern normalization approaches can be subdivided: substitution lists like VARD (Rayson, Archer, and Smith 2005) and Norma (Bollmann 2012), rule-based methods (Baron and Rayson 2008; Porta, Sancho, and Gómez 2013), edit distance based approaches (Hauser and Schulz 2007; Amoia and Martinez 2013), statistical methods and - most recently - neural methods (Partanen, Hämäläinen, and Alnajjar 2019; Duong, Hämäläinen, and Hengchen 2020). For what concerns languages, historical text normalization has been applied to different languages, from different language families. They include, among others, English, Finnish, German, Hungarian, Icelandic, Spanish, Portuguese, Slovene, and Swedish. To our knowledge, spelling normalization has never been tested on historical varieties of Italian. Normalization is unproblematic for historical varieties that are closely related to a standardized modern language. It becomes less effective when spelling is not the most striking difference and co-occurs with morphological, lexical, and structural variation, as is the case with the Italian language (see above). For Italian, to our knowledge normalization has only been applied to contemporary social media language, see Weber and Zhekova (2016) and Van der Goot et al. (2020), which does not pose the challenges specific to old Italian. As pointed out by Manjavacas, Kádár, and Kestemont (2019), while for modern languages normalization is feasible, for historical languages like Italian this is not possible, because one is in front of an amalgam of language variants (diachronic, but also geographic, stylistic, etc.) lacking any sort of super-variant functioning as target.

Historical texts, whether normalized or not, then need to be linguistically annotated. In this paper, we focus on part-of-speech (POS) tagging and lemmatization, which represent critical pre-processing steps for many natural language processing tasks such as information retrieval, knowledge extraction, or semantic analysis.

POS tagging of historical texts is typically carried out against previously normalized texts: normalization has turned out to significantly increase tagging accuracy e.g. on historical English (Rayson et al. 2007), or early modern German (Scheible et al. 2011). Domain adaptation offers an alternative approach to the problem, which is more general (e.g. it can be applied to any corpus without requiring the design of a set of normalization rules). As Yang and Eisenstein (2016) demonstrate for historical English, domain adaptation methods significantly improve the POS tagger performance. They also show that normalization and domain adaptation combine to yield even better performance than that obtained by either approach alone.

Together with spelling normalization, lemmatization of historical language varieties represents the mostly investigated topic over the last years. Lemmatization is a crucial task: lemmas serve as gateways to lexical entries in dictionaries as well as to single occurrences of lexical items in textual corpora. It represents a particularly complex task for morphologically rich languages: Italian is among them. As in the case of POS tagging, lemmatization is successfully carried out against previously normalized texts (Petterson, Megyesi, and Tiedemann 2013; Hämäläinen, Partanen, and Alnajjar 2021). In other approaches, lemmatization is paired with PoS tagging: since inflected

forms can be ambiguous as to their lemma, PoS tags can be used to disambiguate among different lemmas. This information can be exploited as part of joint multi-task learning (Kondratyuk et al. 2018; Manjavacas, Kádár, and Kestemont 2019; Van der Goot et al. 2020) or, more traditionally, in a sequential approach, in which the models for two tasks are learned separately, but the lemmatizer relies on POS information during training and prediction, e.g. Stanza (Qi et al. 2020). Methods used in recent research on historical language lemmatization also include lexicon-assisted tagging, especially for classical languages, as in Eger, vor der Brück, and Mehler (2015), and Burns (2020). Following larger trends in NLP research, neural networks and deep learning approaches, using either word or character-level embeddings, often combined with PoS tagging, define the state-of-the-art performance for many languages; see, e.g., Kestemont et al. (2017), Bergmanis and Goldwater (2018), Manjavacas, Kádár, and Kestemont (2019).

For what concerns Italian, one of the first attempts to automatically annotate historical varieties is reported in Iacobini, De Rosa, and Schirato (2014), showing POS tagging results on the MIDIA corpus, containing texts from the XIIth to the XXth century, where specific patterns from old texts were added to the TreeTagger parameter set. More recently, POS tagging and lemmatization adaptation experiments have been carried out by using (relatively small) manually revised historical corpora to retrain the tools trained on contemporary language, with significantly improved results. This is the case of De Felice et al. (2018) for the *Voci della Grande Guerra* Corpus, of Favaro, Biffi, and Montemagni (2021, 2022a) for a subset of the VoDIM corpus (see below), and of Favaro et al. (2022) for the the quotations in the *Grande dizionario della lingua italiana* ('Great Dictionary of Italian Language', in short GDLI). Last but not least, Palmero Aprosio, Menini, and Tonelli (2022) introduce BERToldo, one of the BERT-like models, trained from scratch on historical data. The different transformer models built achieve high accuracy rates: POS tagging evaluation on D(h)ante corpus (Basile and Sangati 2016) shows an accuracy ranging between 93% and 96%, depending on the different versions.

3. The Corpus

In this study, we took as a reference resource the diachronic corpus of the *Vocabolario Dinamico dell'Italiano Moderno* 'Dynamic Vocabulary of Modern Italian', in short VoDIM (Marazzini and Maconi 2018), that collects texts testifying post-unitarian Italian. VoDIM texts are both oral and written, belonging to different textual genres and domains: art, economy, gastronomy, law, music, newspapers, poetry, politics, science. The VoDIM corpus, whose size is currently about 20 million tokens[2], is balanced, thanks to a subcorpora division that allows dynamic balancing of topics and chronology (Biffi and Ferrari 2020). VoDIM is available online[3], into Crusca's *Scaffali Digitali* 'Digital Bookshelves' (Biffi 2020).

The VoDIM corpus is an excellent starting point for constructing diachronically representative language resources, as its texts cover a significant time span (1861-today) varying also along the diamesic and typological axes. For the specific concerns of historical language processing, we built a representative VoDIM subcorpus, with texts from seven prose genres, namely art, gastronomy, law, newspapers, literature, bestsellers, and science. Each of these sections maintains a diachronic balance, with a

2 In the future, this corpus will be extended to become a large web corpus, whose size is expected to increase to about 2 billion words (Biffi 2020; Biffi and Ferrari 2020)

3 `http://www.stazionelessicografica.it`

size of approximately 3,000 tokens. Consequently, the overall subcorpus size is around 21,000 tokens (corresponding to about 22,000 morphological words). Table 1 reports the subcorpus composition by post-unitarian Italian time periods, which are aligned with DiaCORIS (Onelli et al. 2006) and LIS[4] settings. An additional period corresponding to contemporary Italian is also included. Table 2 illustrates the distribution of the same texts across the different textual genres / domains.

Table 1
VoDIM subcorpus composition by time periods

Years	*Texts*	*Tokens*	*Words*	*Sentences*
1 1861-1900	6	3828	4111	193
2 1901-1922	5	3161	3342	173
3 1923-1945	5	2982	3174	136
4 1946-1967	5	3810	4073	141
5 1968-2001	6	3340	3546	144
6 2002-today	20	3503	3826	108
Tot.	47	20624	22072	895

Table 2
VoDIM subcorpus composition by textual genres / domains

Genre	*Years*	*Texts*	*Tokens*	*Words*	*Sentences*
art	(2-6) 1902-2009	5	3225	3436	110
gastronomy	(1-4) 1871-1947	5	3071	3275	157
law	(5-6) 2000-2016	18	2552	2812	64
newspaper	(1-5) 1867-1996	4	2572	2791	103
literature	(1-5) 1881-1982	5	3246	3391	209
bestsellers	(1-4) 1892-1954	5	3085	3252	142
science	(1-6) 1864-2015	5	2873	3115	110
Tot.		47	20624	22072	895

The VoDIM subcorpus was automatically annotated with Stanza (Qi et al. 2020), a state-of-art fully neural pipeline for multilingual NLP trained on Universal Dependencies (UD) treebanks (De Marneffe et al. 2021). UD today represents a *de facto* standard for morpho-syntactic and syntactic dependency annotation of texts, including historical varieties of some of the covered languages. Among the two main UD-compliant annotation pipelines, namely UDPipe (Straka and Straková 2017) and Stanza, we opted for the latter mainly because of its lemmatization strategy, which permits to customize the lemmatizer by providing a key-value dictionary: this represents a key feature when dealing with historical varieties of language, especially in the case of morphologically rich ones (see Section 5). Annotation concerned tokenization, POS tagging and lemmatization. As a baseline, the 'combined' model for Italian, trained on the combination of the different Italian UD corpora (namely, ISDT, VIT, PoSTWITA, and TWITTIRO)

4 For additional information about LIS (*Lessico Italiano Scritto* 'Written Italian Lexicon'), cfr. Biffi (2016)

was used[5]. Automatic annotation was then manually revised and, whenever needed, corrected to create a gold standard corpus to be used for both training and testing purposes (Favaro, Biffi, and Montemagni 2022b).

In the revision, we took advantage of the experience gained in the project *Voci della Grande Guerra* 'Voices of the Great War' (VGG) (De Felice et al. 2018; Lenci et al. 2020), especially for what concerns sentence splitting, tokenization and lemmatization problems derived from the automatic processing of a non-standard historical language variety. With regard to tokenization, some of the annotation problems observed in VGG are the same as those experienced in the annotation of the VoDIM subcorpus. This is the case, for instance, of the segmentation of pronominal clitics occurring with finite verbs (i.e. *prendevasi*, *prendeva+si* 'you took', which was automatically analyzed as a unique form), which represents a feature typical of older stages of Italian. Punctuation usage is another peculiar feature connected with different genres (Favaro, Biffi, and Montemagni 2021): in novels and bestsellers, it is not uncommon to find punctuation marks aimed at reproducing the speech pattern of dialogues or a greater emphasis of the utterance, such as 'mixed' dots (*?!*, *!?*); an excessive number of suspension marks is another feature creating sentence splitting problems. Furthermore, scientific texts include a large number of acronyms and symbols, causing various hyposegmentation issues (De Felice et al. 2018).

For what concerns lemmatization, as reported in Favaro et al. (2022), we opted for a 'low-level' (i.e. conservative) strategy: this entails that, at this level, we do not abstract away from graphical, phonological, morphological or lexical variants, e.g. *amministragione* and *amministrazione* 'administration' represent distinct lemmas rather than being seen as distinct variant forms of the same abstract lemma (Favaro et al. 2022). Miletić and Siewert (2023) discuss the pros and cons of such an approach. On the one hand, a lemmatization strategy respecting different levels of variation (lexical, morphological, orthographic) allows for the preservation of varietal differences, but limits the positive impact of lemmatization on data sparsity. On the other hand, choosing one language variety over the others for lemmatization purposes is more effective for what concerns data sparsity, but it combines together two distinct tasks, lemmatization and normalization, arguably making the process more difficult. In morphologically rich languages like Italian, lemma normalization is far from trivial, both computationally (Hämäläinen, Partanen, and Alnajjar 2021) and linguistically (Favaro et al. 2022), especially in relation to a diachronic corpus such as VoDIM, that includes both historical and contemporary texts. We thus decided to carry out lemmatization in two steps, the first one consisting in associating to a given inflected form the corresponding dictionary head-form (or lemma), and the second one in charge of the normalization of lemmas identified at the previous step. In our approach, normalization of lemma variants will thus be carried out as a post-processing step, in order to reduce data sparsity, thus making it possible — in perspective — to query the corpus at different abstraction levels. In this paper, we focus on the first step only.

5 https://stanfordnlp.github.io/stanza/combined_models.html

4. POS Tagging

4.1 POS Tagging Experiments

In this section, we describe the POS tagging experiments carried out on the VoDIM subcorpus. The creation of customized models for POS tagging has been carried out in a classical supervised domain adaptation scenario. The coverage of the Stanza POS tagger designed for modern Italian was extended to cover the historical varieties collected in the VoDIM corpus by retraining the tagger on the manually corrected corpus (see Section 3) which was used in combination with the ISDT corpus (Bosco, Montemagni, and Simi 2013), whose current size is 257,616 tokens (13,121 sentences). Once the data were collected, several testing experiments were carried out to analyze the impact of parameters on the training performance. Regarding POS tagging, the experiments were carried out by testing batch size. Considering that Stanza default batch size is 5000, we needed to set a lower value in order to avoid overfitting on training data. First, we tried to set mini-batches (32 and 64), that enabled fast training runs on the Tesla T4 GPU provided by Google Colaboratory[6], but did not lead to significant improvements. After other experiments with different sizes, the best results were achieved with a batch size of 512. In what follows, we report the results achieved with this setting.

Table 3
VoDIM subcorpus partitioning in terms of tokens and sentences, by time and textual genre

Fold	*Time Period*	*Genre/s*	*Test Tok.*	*Test Sent.*	*Train Tok.*	*Train Sent.*
1	2-6	art	2059	81	18565	814
2	1-4	art/gastronomy	2067	89	18557	806
3	1-4	gastronomy	2009	96	18615	799
4	5-6	law	2091	59	18533	836
5	1-6	law/newspapers	2053	95	18571	800
6	1-5	newspapers/literature	2096	97	18528	798
7	1-5	literature	2063	97	18561	798
8	2-5	literature/bestsellers	2048	88	18576	807
9	1-4	bestsellers/science	2053	98	18571	797
10	3-6	science	2085	95	18539	800
		Tot.	20624	895		

To estimate the POS tagging performance after retraining, the annotated VoDIM subcorpus was split into 10 equal parts, to prepare the data for a 10-fold cross-validation: 90% of the VoDIM subcorpus was used for retraining (corresponding, on average, to 18.500 tokens), and the remaining 10% (2.000 tokens on average) was used for testing. Instead of randomly splitting the VoDIM subcorpus, we opted for a corpus partitioning reflecting the chronological and domain classes described above, with the final aim of evaluating the potential influence of linguistic features related to the textual genre and/or the period, as described in the following sections. Table 3 shows in detail the partitioning into 10 folds of the VoDIM subcorpus, with specification - for each fold

6 https://colab.research.google.com/

- of the covered periods and genre / domain. As a baseline, the 'combined' model for Italian was used (see Section 3).

4.2 POS Tagging Evaluation

We run 10-fold cross-validation to assess the POS tagger performance on the VoDIM subcorpus: Table 4 details achieved results, both on average and in the individual folds. It turned out that the baseline POS tagging model is already effective for the different VoDIM text types, even in the case of older texts. The accuracy of retrained POS tagging models increases, on average by 0,3 for Universal POS tags (in short, UPOS)[7], 0,5 for language-specific part-of-speech tags (in short, XPOS)[8] and 0,9 for associated universal features (in short, UFeats)[9].

The distance in accuracy between the baseline POS tagging model and the retrained model is more pronounced in certain iterations, as evident in the case of the second fold (+2% for UPOS, XPOS, UFeats), covering art and gastronomy domains. On the other hand, the performance of the retrained model deteriorates compared to the baseline for all UPOS, XPOS and UFeats in the ninth fold (covering bestsellers and scientific texts). Since the texts in both second and ninth fold share the same time span (1-4, cfr. Table 3), the origin of the observed gap in the POS tagging performance should rather be looked for in other synchronic caracteristics. Our hypothesis is that genre differences have influenced the results in some way. As a matter of fact, if we broaden the perspective, the retrained model achieves better results in 6 training iterations (1, 2, 5-8), but the baseline outperforms in the remaining ones.

Table 4
10-fold cross validation results for UPOS, XPOS and UFeats

Fold	baseline	retrained	baseline	retrained	baseline	retrained
	UPOS		XPOS		UFeats	
1	0.975	0.980	0.952	0.962	0.964	0.977
2	0.963	**0.977**	0.953	**0.966**	0.961	**0.979**
3	0.980	0.974	0.973	0.966	0.971	0.973
4	0.974	0.973	0.963	0.961	0.963	0.969
5	0.964	0.976	0.948	0.963	0.946	0.969
6	0.974	0.978	0.971	0.974	0.968	0.982
7	0.974	0.983	0.967	0.976	0.967	0.984
8	0.968	0.976	0.957	0.965	0.959	0.969
9	**0.975**	0.965	**0.952**	0.948	**0.964**	0.949
10	0.987	0.980	0.964	0.960	0.976	0.975
average	0.973	0.976	0.960	0.964	0.964	0.973

Consider now, for the second and ninth folds, the performance results by POS. Tables 5 (fold 2) and 6 (fold 9) report, for each UPOS, the values of precision, recall

7 https://universaldependencies.org/u/pos/
8 http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf
9 https://universaldependencies.org/u/feat/index.html

Table 5
UPOS analysis of fold 2 (higher baseline values are marked in bold)

UPOS	Precision		Recall		F1-score		Freq
	baseline	retrained	baseline	retrained	baseline	retrained	
ADJ	0.968	0.968	0.893	0.899	0.929	0.933	169
ADP	0.992	0.995	0.981	0.995	0.987	0.995	378
ADV	0.913	0.958	0.966	0.966	0.939	0.962	119
AUX	**1.000**	0.943	**0.985**	0.985	**0.992**	0.964	67
CCONJ	0.969	0.985	0.955	1.000	0.962	0.992	66
DET	0.997	0.997	0.987	0.989	0.992	0.993	380
INTJ	0.000	0.000	0.000	0.000	0.000	0.000	1
NOUN	0.957	0.977	0.987	0.975	0.972	0.976	472
NUM	0.846	0.978	0.489	1.000	0.62	0.989	45
PRON	0.967	0.976	0.952	0.976	0.959	0.976	124
PROPN	**0.786**	0.550	1.000	1.000	**0.880**	0.710	11
PUNCT	0.911	1.000	1.000	1.000	0.953	1.000	204
SCONJ	0.886	0.917	0.912	0.971	0.899	0.943	34
SYM	0.333	1.000	0.333	1.000	0.333	1.000	3
VERB	**0.983**	0.967	**0.994**	0.978	**0.989**	0.972	179
X	0.000	0.000	0.000	0.000	0.000	0.000	3

Table 6
UPOS analysis of fold 9 (higher baseline values are marked in bold)

UPOS	Precision		Recall		F1-score		Freq
	baseline	retrained	baseline	retrained	baseline	retrained	
ADJ	**0.960**	0.950	**0.933**	0.923	**0.946**	0.937	104
ADP	0.980	0.985	1.000	1.000	0.990	0.992	195
ADV	0.991	1.000	**0.915**	0.872	**0.951**	0.932	117
AUX	**0.989**	0.967	**0.968**	0.957	**0.978**	0.962	93
CCONJ	0.983	1.000	0.983	1.000	0.983	1.000	58
DET	0.996	1.000	1.000	1.000	0.998	1.000	226
INTJ	**0.846**	0.818	**1.000**	0.818	**0.917**	0.818	11
NOUN	**0.965**	0.950	**0.976**	0.962	**0.971**	0.956	338
NUM	**1.000**	0.923	1.000	1.000	**1.000**	0.960	24
PRON	**0.987**	0.981	**0.981**	0.968	**0.984**	0.975	158
PROPN	**0.731**	0.721	1.000	1.000	**0.845**	0.838	49
PUNCT	1.000	1.000	**1.000**	0.995	**1.000**	0.998	423
SCONJ	**0.920**	0.727	0.920	0.960	**0.920**	0.828	25
SYM	0.000	0.000	0.000	0.000	0.000	0.000	0
VERB	**0.983**	0.963	**0.992**	0.975	**0.987**	0.969	238
X	**1.000**	0.882	**0.515**	0.455	**0.680**	0.600	33

and f1-score. The "freq" column, instead, registers the frequency of occurrence of each UPOS tag.

For both folds, problematic POS categories are represented by the residual class (X, covering ideophones, onomatopoeias, etc.) and interjections (INTJ): they both represent rare and at the same time highly variable elements, very hard to be correctly predicted; their frequency is very limited, especially in fold 2.

Consider now the case of POS tags whose recognition is more accurate in the baseline model for both folds. This is the case of proper nouns (PROPN), whose classification is problematic in both folds, although to a different extent. Whereas recall in both cases is 1 (i.e. all proper nouns are correctly retrieved), there is a tendency to overextend the tag to capitalized common nouns. The typical case is the "ideological" representation of words such as *Governo* 'Government', *Stato* 'State', etc., which constitute one possible variant in Italian historical varieties. Interestingly, the overextension of the tag is more accentuated in fold 2 rather than fold 9: we wonder whether this difference could be due to the involved textual genres. Words characteristic of the 'art' genre are capitalized in contexts where they are common nouns (e.g. *Basilica* 'basilica' or *Duomo* 'cathedral') and thus erroneously tagged as proper nouns by the retrained model, as opposed to the baseline (see precision values for PROPN in fold 2, 0.786 for the baseline vs 0.550 for the retrained model). Another tricky case concerns the classification of auxiliaries (AUX) and main verbs (VERB), which is problematic for both folds in the retrained models: it seems that the baseline model performs better on these word classes.

Besides the AUX, PROPN and VERB cases discussed above, in fold 9 the baseline performance is better also for adjectives (ADJ), nouns (NOUN), numerals (NUM), pronouns (PRON) and subordinative conjunctions (SCONJ). The most striking difference is observed in subordinative conjunctions, for which the baseline precision is 0.920 against 0.727 of the retrained model; for what concerns recall, the reverse is the case, i.e. the retrained model correctly retrieves a higher number of cases. We are currently investigating the reasons underlying this state of affairs.

5. Lemmatization

In Section 3, the adopted two-step strategy for lemmatizing historical varieties of Italian was introduced and motivated. In this section, we focus on the first step, aimed at associating to a given inflected form the corresponding lemma, without normalization of the lemma variants which will be treated as a post-processing step. In particular, we illustrate the different customized models that have been developed, including both lexicon-assisted models and a model retrained with the addition of historical annotated texts, and compare achieved results quantitatively and qualitatively.

5.1 Lemmatization Experiments

Regarding lemmatization, the goal was to compare accuracy improvements between a retrained model and a baseline lemmatiser augmented with different morphological lexicons, both gold and extracted from automatically annotated corpora.

For what concerns the retrained model, the VoDIM subcorpus was subdivided in training and test sets with the same ratio as for POS tagging (i.e. 90 + 10). Differently from the previous case, we randomly split the whole corpus, covering all time spans and textual genres / domains (cfr. Table 7). Unlike POS tagging, to generate the retrained model we maintained the Stanza default settings, because the performance of the retrained model was already high in the first retraining session (see below).

Table 7
VoDIM subcorpus partitioning for lemmatization retraining, in terms of tokens and sentences

Test Tok.	*Test Sent.*	*Train Tok.*	*Train Sent.*
2204	92	18420	803

Besides model retraining, Stanza provides an alternative method to improve the performance of the baseline lemmatization model, by adding a morphological lexicon encoded as a "key-value" Python dictionary. We thus created different morphological lexicons from sources related to different historical periods and different text types, with the final aim of improving the lemmatizer performance against historical varieties of language. These lexicons were built with different methods. One was the wide coverage general purpose lexicon, testifying contemporary usage and used for annotating the ISST-TANL corpus (Montemagni and Simi 2007). Two other lexicons were extracted from annotated corpora: the one labelled as VoDIM was extracted from the automatically annotated VoDIM corpus; the other was extracted from the *Stampa Periodica Milanese* corpus (in short SPM) (De Stefanis Ciccone, Bonomi, and Masini 1984), which was manually annotated and including Milanese periodical press of the early XIXth century. ISST-TANL lexicon is, compared to the two others, the only lexicon with complete morphology, manually revised and totally expanded, while the coverage of the morphological VoDIM and SPM lexicons is limited to the wordforms (and lemmas) occurring in the reference corpora. As reported in Table 9, each lemma in the ISST-TANL lexicon is associated with an average of 6.5 different forms, while each lemma in the corpora-derived lexicons is associated with an average of 2.5. We also created different combinations of these lexicons (see below) to expand the coverage of individual lexicons.

In order to be used by the Stanza baseline model for lemmatization, all lexicons were automatically converted from the proprietary formats to the the universal POS tagset (UPOS). The mapping is illustrated in Table 8, where it can be observed that for conjunctions and verbs finer-grained POS tags needed to be reconstructed, in particular the distinction between coordinating and subordinating conjunctions, auxiliary and main verbs. The ISST-TANL tagset, also used by the LinguA pipeline (Attardi and Dell'Orletta 2009; Attardi et al. 2009; Dell'Orletta 2009), has 14 coarse-grained POS tags and 37 fine-grained POS tags[10]. The tagset used for the VoDIM and SPM annotation is the Pi-Morfo tagset, used the morphological analyzer of Pi-System (Picchi 2003), which also has 14 coarse-grained POS tags, largely overlapping with ISST-TANL tags, with which 49 grammatical subcategories are associated.

Starting from these lexicons, two other combined lexicons were created to evaluate the impact of combining the different lexicons on the lemmatization performance: the combination of all lexicons (in short SIV, resulting from the combination of SPM, VoDIM and ISST-TANL), and the combination of lexicons extracted from gold annotations (labeled as SI, combining SPM and ISST-TANL). On the one hand, SIV has a broader lemma coverage and wider chronological coverage (from the early XIXth century to the 2000s); on the other hand, SI is smaller but is expected to be more accurate.

10 `http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf`

Table 8
POS tagsets mapping

Class	*ISST-TANL*	*Pi-Morfo*	*UPOS*
adjective	A	A	ADJ
adverb	B	B	ADV
conjuction	C	C	CCONJ | SCONG
determiner	D	D	DET
adposition	E	E	ADP
puctuation	F	F	PUNCT
interjection	I	I	INTJ
numeral	N	N	NUM
pronoun	P	P	PRON
article	R	R	DET
noun	S	S	NOUN
predeterminer	T		DET
verb	V	V	AUX | VERB
residual	X		X

Table 9
Statistics of used morphological lexicons

Lexicon	*Forms*	*Lemmas*	*Form/Lemma Ratio*
ISST-TANL	391377	60206	6.5:1
VoDIM	199091	78754	2.5:1
SPM	44913	17404	2.5:1
	Combined Lexicons		
SIV	478364	106356	4.5:1
SI	402500	62738	6:1

5.2 Evaluation of Lemmatization

Table 10 reports the results obtained by the different lemmatization models: baseline and retrained models, as well as the five models resulting from the combination between the baseline model and the different developed lexicons.

It can be noticed that all models are quite accurate as far as lemmatization is concerned. However, the highest accuracy is achieved by the retrained model, showing a noticeable gain with respect to the baseline (+0,8%). By contrast, the baseline model augmented with morphological lexicon lookup does not appear to produce any improvement to the lemmatization process: in particular, SI, resulting from the combination of gold standard lexicons (SPM and ISST-TANL), that we expected to ensure higher accuracy than the lexicons extracted from automatic annotations (VoDIM and, consequently and partially, SIV), shows the worst performance (together with SPM).

Table 11 reports the number of errors made by each model; the "type" column refers to the number of different error types, while the "token" column reports the total number of errors.

Table 10
Lemmatization accuracy with the different models

	baseline	0.984
	retrained	0.992
baseline+lexicon	ISST-TANL	0.982
	VoDIM	0.984
	SPM	0.981
	SIV	0.984
	SI	0.981

Table 11
Number of lemmatization errors, by type and token

		n. of errors	
		type	token
	baseline	33	37
	retrained	15	18
baseline+lexicon	ISST-TANL	36	41
	VoDIM	32	37
	SPM	39	44
	SIV	33	38
	SI	40	45

Cross-referencing the data reveals that only six error types are shared by all models. Upon analyzing these data, it becomes apparent that, except for the preposition *da* meaning 'from', all errors also involve POS tagging. Notably, four out of the six errors revolve around the ambiguity between adjectives and past participles. For instance, the word *organizzato* ('organized') can function both as an adjective and as a past participle of *organizzare* ('to organize') (see Table 12). Similarly, *condotte* was mistakenly interpreted as the past participle of the verb *condurre* ('to bring') rather than the plural form of the noun *condotta* ('conduct'). Additionally, the error concerning the literary and rare subordinating conjunction *onde* ('whence') arises from the ambiguity with the plural form of the noun *onda* ('wave').

Table 12
Lemmatization errors shared by all models

form	lemma_gold	lemma_pred
organizzato	*organizzato*	*organizzare*
condotte	*condotta*	*condurre*
essiccato	*essiccato*	*essiccare*
onde	*onde*	*onda*
D'	*da*	*di*
aperto	*aperto*	*aprire*

Table 13 shows instead the errors shared by the baseline and each of the developed models. It should be noted that with the SIV lexicon (resulting from the combination of all lexicons) the number of errors shared with the baseline model is the lowest, amounting to 15. These errors are idiosyncratic in nature because they originate from specific features of each individual lexicon. They include - for instance - lemmatization choices, as highlighted in Favaro, Biffi, and Montemagni (2022a) for what concerns the ISST-TANL lexicon, or standardization of historical variants, as in SPM where *polito* is lemmatized as *pulito* ('cleaned') and *quistione* as *questione* ('question').

Table 13
Errors shared by the baseline and the other models

	n. of shared errors
baseline/retrained	9
baseline/ISST-TANL	19
baseline/VoDIM	17
baseline/SPM	22
baseline/SIV	15
baseline/SI	20

5.2.1 Error Analysis

Table 14 reports the typology of errors made by the different models. In the first column, errors originating at the level of POS tagging are reported: this is the case of the ambiguity between adjective and past participle (e.g. *sciupato/sciupare* 'damaged'/'to damage'), or between adjective and noun (e.g. *attrattiva/attrattivo*, 'attraction'/'attractant'). Besides these typical and pervasive POS ambiguities, there are other unpredictable cases, such as the *onde* homography discussed above. The second column collects ambiguous lemmatization cases where the POS tag was properly recognized. Consider, as an example, the verb form *rimandi* lemmatized as *rimanere* 'to remain' instead of *rimandare* 'to resend', or the preposition *D'* lemmatized as *da* 'from' instead of *di* 'of'. The third column in the table collects those errors that generate inadmissible lemmas in Italian (e.g. *sfuggono* as a form of the verb **sfuggere* instead of *sfuggire* 'to escape',

Table 14
Error typology

		ambiguity (diff. POS)	ambiguity (same POS)	pred. errors	punct.	Tot.
	baseline	15	10	7	5	37
	retrained	10	2	2	4	18
baseline+lexicon	ISST-TANL	20	12	5	4	41
	VoDIM	13	15	5	4	37
	SPM	16	16	8	4	44
	SIV	15	14	5	4	38
	SI	18	18	5	4	45

or *misi* lemmatized as **misare* instead of *mettere* 'to put'). The last column concerns errors relating to punctuation marks, also including typographical symbols, such as apostrophes.

The retrained model exhibits over half of its errors falling under the first category, primarily stemming from POS errors. Lemmatization errors, documented in the second column, account for only 22% of the total errors in the case of the retrained model, in contrast to the 45-50% observed in all other models. Consequently, it can be claimed that the retrained model shows a considerable degree of reliability in predicting the correct lemma.

In order to assess the performance of a lemmatization model, we also evaluated its ability to properly deal with out-of-vocabulary (OOV) words (i.e. tokens that do not appear in the training data).

Table 15
Lemmatization accuracy on out-of-vocabulary words

	accuracy
baseline	0.927
retrained	0.994

To this specific end, we excluded from the 2.204 test tokens those occurring in the training corpus: 177 tokens were left. Table 15 reports the accuracy of the baseline and retrained models for the subset of OOV words. The retrained model confirms its higher ability to predict proper lemmas, with an accuracy (0.994) slightly higher than that observed for the full test set (0.992).

6. Conclusions

In this paper, we illustrated our approach towards POS tagging and lemmatization of historical varieties of Italian, and reported the results of experiments carried out in a classical supervised domain adaptation scenario. We focused on the period going from Italy's unification in 1860 until the contemporary language, a relatively short period, which – given the history of the Italian language – is already sufficient to bring to light the challenges associated with the processing of old Italian.

Among the features of the proposed approach, it is worth mentioning here that spelling normalization, very often resorted to for dealing with historical varieties of languages, does not appear to us as a viable solution due to the peculiar history of the Italian language. Normalization is unproblematic for historical varieties that are closely related to a standardized modern language, which is not the case for old Italian where spelling variation co-occurs with other variation types (e.g. morphological, lexical, structural, but also geographic, diastratic, etc.).

Another important characteristic of the proposed approach is concerned with the adoption of a low-level conservative lemmatization strategy, where lemma normalization is postponed to a later stage. By decoupling the mapping of a specific word form to its dictionary headword from lemma normalization, the mapping process (coinciding with step 1) becomes restricted to the reconstruction of the lemma, without abstracting away from possible graphical, phonological, or morphological variation: this approach makes the lemmatization results more accurate, even if more fragmented.

For what concerns POS tagging, although the baseline model is already effective for the different VoDIM text types, even for older texts, the accuracy of retrained models increases: interestingly, higher improvements are reported for finer-grained categories, i.e. language-specific POS tags and associated features. In the 10-fold cross-validation, the splitting of the training corpus was carried out in such a way as to control the internal composition of individual folds: each fold was representative of specific periods and textual genres. As a result, the distance between the baseline POS tagging model and the retrained one turned out to be more pronounced in certain folds, thus suggesting the influence of also the textual genre/domain on the result. This outcome is in line with POS tagging experiments carried out on a chronologically wider corpus, collecting quotations from the *Grande Dizionario della Lingua Italiana* (Favaro et al. 2022). Achieved results show the heavy influence of the author's style compared to the linguistic evolution stage: e.g. the annotation of Vittorio Alfieri's texts, who lived in the XVIIIth century, shows lower accuracy values compared to those of authors such as Matteo Maria Boiardo, who lived three centuries earlier, both for the baseline and the retrained models.

Lemmatization experiments have been aimed at comparing accuracy improvements between a retrained model and a lexicon-assisted baseline lemmatizer, using different morphological lexicons, both gold and extracted from automatically annotated corpora. Although all models turned out to be quite accurate, the best performance was achieved by the retrained model, with a gain of +0,8% with respect to the baseline. Interestingly, the retrained model showed an increased accuracy on the subset of OOV words compared with that observed for the overall test set. The baseline model augmented with morphological lexicon lookup does not appear to produce any appreciable improvement: in this case, we suppose that some of the errors of lexicon-assisted models could originate in interpretative problems at the level of lemmatization criteria (e.g. the use of normalized lemmas rather than low-level ones).

The results achieved so far are encouraging: we thus believe that time is ripe for linguistically annotating bigger historical corpora with a high degree of accuracy, thanks to the high performance of retrained Stanza neural models for POS tagging and lemmatization. Lines of research currently being explored include: the extension of the gold annnotated corpus, covering other periods and textual genres / domains; the identification of the most appropriate model for annotating (both POS tagging and lemmatizing) texts of a specific variety of language use (e.g. a specific period, textual genre, or a given author); the design and implementation of the second step of the incremental strategy for lemmatizing texts characterized by a high degree of variability, corresponding to lemma normalization.

Acknowledgments

The resources discussed in this paper were developed in the framework of the project *Trattamento Automatico di Varietà Storiche di Italiano* ('Automatic processing of historical varieties of Italian', TrAVaSI), a project involving the Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC) and Accademia della Crusca, funded by Regione Toscana (POR FSE 2014 - 2020). Experiments and results reported in the previous sections have been partially carried out within the PNRR project PE20 "Cultural Heritage Active Innovation for Sustainable Society (CHANGES)", among the research activities of Spoke 3. Special thanks are due to Felice Dell'Orletta for his support in setting up the experiments.

References

Amoia, Marilisa and Jose Manuel Martinez. 2013. Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In *Proceedings of the 7th*

workshop on language technology for cultural heritage, social sciences, and humanities, pages 84–89, Sofia, Bulgaria, August.
Antonelli, Giuseppe. 2003. *Tipologia linguistica del genere epistolare nel primo Ottocento. Sondaggi sulle lettere familiari di mittenti cólti*. Edizioni dell'Ateneo, Roma.
Attardi, Giuseppe and Felice Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 261–264, Boulder, Colorado, June. Association for Computational Linguistics.
Attardi, Giuseppe, Felice Dell'Orletta, Maria Simi, and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December.
Baron, Alistair and Paul Rayson. 2008. VARD2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate conference in corpus linguistics*, Birmingham, UK, May.
Basile, Angelo and Federico Sangati. 2016. D(h)ante: A new set of tools for XIII century italian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May.
Bergmanis, Toms and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana, June. Association for Computational Linguistics.
Biffi, Marco. 2016. Progettare il corpus per il vocabolario postunitario, in l'italiano elettronico. vocabolari, corpora. In C. Marazzini and L. Maconi, editors, *L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori*. Accademia della Crusca, Firenze, pages 259–80.
Biffi, Marco. 2020. La galassia lessicografica della crusca in rete. In L. Leonardi e P. Squillacioti, editor, *Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo digitale*. Edizioni dell'Orso, Alessandria, pages 219–232.
Biffi, Marco and Angela Ferrari. 2020. Progettare e ideare un corpus dell'italiano nella rete: il caso del coliweb. *Studi di Lessicografia Italiana*, 37:357–374.
Bollmann, Marcel. 2012. (Semi-)automatic normalization of historical texts using distance measures and the norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, Portugal, August.
Bollmann, Marcel. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, June. Association for Computational Linguistics.
Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.
Burns, Patrick J. 2020. Ensemble lemmatization with the classical language toolkit. *Studi e Saggi Linguistici*, 58:157–176.
De Felice, Irene, Felice Dell'Orletta, Giulia Venturi, Alessandro Lenci, and Simonetta Montemagni. 2018. Italian in the trenches: Linguistic annotation and analysis of text of the great war. In *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it)*, Torino, Italy, December.
De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
De Stefanis Ciccone, Stefania, Ilaria Bonomi, and Andrea Masini. 1984. *La stampa periodica milanese della prima metà dell'Ottocento: testi e concordanze*. Giardini, Pisa.
Dell'Orletta, Felice. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December.
Duong, Quan, Mika Hämäläinen, and Simon Hengchen. 2020. An unsupervised method for OCR post-correction and spelling normalisation for finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, Iceland (Online), May/June. Linköping University Electronic Press.
Eger, Steffen, Tim vor der Brück, and Alexander Mehler. 2015. Lexicon-assisted tagging and lemmatization in latin: A comparison of six taggers and two lemmatization methods. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social*

Sciences, and Humanities (LaTeCH 2015), pages 105–113, Beijing, China, July.
Favaro, Manuel, Marco Biffi, and Simonetta Montemagni. 2021. Risorse e strumenti per le varietà storiche dell'italiano: il progetto TrAVaSI. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 178–186, Bologna, Italy, March.
Favaro, Manuel, Marco Biffi, and Simonetta Montemagni. 2022a. Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione. In M. Misuraca, G. Scepi, and M. Spano, editors, *Proceedings of the 16th international conference on statistical analysis of textual data*, volume I. Vadistat press, Napoli, July, pages 392–399.
Favaro, Manuel, Marco Biffi, and Simonetta Montemagni. 2022b. TrAVaSI_VoDIM corpus. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.
Favaro, Manuel, Elisa Guadagnini, Eva Sassolini, Marco Biffi, and Simonetta Montemagni. 2022. 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 An-cient Languages (LT4HALA 2022)*, pages 94–100, Marseille, France, June. European Language Resources Association (ELRA).
Hauser, Andreas W. and Klaus U. Schulz. 2007. Unsupervised learning of edit distance weights for retrieving historical spelling variations. In *Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, pages 1–6, Borovets, Bulgaria, September.
Hämäläinen, Mika, Niko Partanen, and Khalid Alnajjar. 2021. Lemmatization of Historical Old Literary Finnish Texts in Modern Orthography. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 189–198, Lille, France, June.
Iacobini, Claudio, Aurelio De Rosa, and Giovanna Schirato. 2014. 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 and of the Fourth International Workshop EVALITA 2014*, pages 213–218, Pisa, Italy, December. Pisa University Press.
Kestemont, Mike, Guy de Pauw, Renske van Nie, and Walter Daelemans. 2017. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, 32(4):797–815.
Kondratyuk, Daniel, Tomáš Gavenčiak, Milan Straka, and Jan Hajič. 2018. LemmaTag: Jointly tagging and lemmatizing for morphologically rich languages with BRNNs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4921–4928, Brussels, Belgium, October-November. Association for Computational Linguistics.
Lenci, Alessandro, Simonetta Montemagni, Federico Boschetti, Irene De Felice, Stefano Dei Rossi, Felice Dell'Orletta, Michele Di Giorgio, Martina Miliani, Lucia C. Passaro, Angelica Puddu, Giulia Venturi, and Nicola Labanca. 2020. Voices of the great war: A richly annotated corpus of italian texts on the first world war. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 911–918, Marseille, France, May. European Language Resources Association.
Manjavacas, Enrique, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. 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 1493–1503, Minneapolis, Minnesota, June. Association for Computational Linguistics.
Marazzini, Claudio and Ludovica Maconi. 2018. Il Vocabolario dinamico dell'italiano moderno rispetto ai linguaggi settoriali. proposta di voce lessicografica per il redigendo VoDIM. *Italiano Digitale*, 7(4):101–120.
Mengaldo, Pier VIncenzo. 1987. *L'epistolario di Nievo. Un'analisi linguistica*. Il Mulino, Bologna.
Miletić, Aleksandra and Janine Siewert. 2023. Lemmatization experiments on two low-resourced languages: Low Saxon and Occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 163–173, Dubrovnik, Croatia, May. Association for Computational Linguistics.
Montemagni, Simonetta and Maria Simi. 2007. The Italian dependency annotated corpus developed for the CoNLL–2007 shared task. Technical report, CNR-ILC.
Onelli, Corinna, Domenico Proietti, Corrado Seidenari, and Fabio Tamburini. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1212–1215, Genoa, Italy, May. European Language Resources Association (ELRA).

Palmero Aprosio, Alessio, Stefano Menini, and Sara Tonelli. 2022. BERToldo, the historical BERT for Italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 68–72, Marseille, France, June. European Language Resources Association.

Partanen, Niko, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect Text Normalization to Normative Standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146, Hong Kong, China, November.

Pennacchiotti, Marco and Fabio M. Zanzotto. 2008. Natural language processing across time: An empirical investigation on Italian. In *Proceedings of GoTAL - 6th International Conference on Natural Language Processing*, pages 371–382, Gothenburg, Sweden, August.

Petterson, Eva, Beata Megyesi, and Jorg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*, page 54–69, Oslo, Normay, May.

Picchi, Eugenio. 2003. Pisystem: sistemi integrati per l'analisi testuale. In A. Zampolli, N. Calzorali, and L. Cignoni, editors, *Computational Linguistics in Pisa. Linguistica Computazionale*. IEPI, Pisa-Roma, pages 597–627.

Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

Plank, Barbara. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing*, Bochum, Germany, Semptember.

Porta, Jordi, José-Luis Sancho, and Javier Gómez. 2013. Edit transducers for spelling variation in Old Spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*, Oslo, Norway, May.

Qi, Peng, Zhang Yuhao, Zhang Yuhui, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for many human languages. In *ACL2020 System Demonstration*, Online, July. Association for Computational Linguistics.

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the [ba]ard: Evaluating the accuracy of a modern pos tagger on early modern English corpora. In *Corpus Linguistics Conference*, Birmingham, UK, July.

Rayson, Paul, Dawn Archer, and Nicolas Smith. 2005. VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora. In *Proceedings of the Corpus Linguistics conference*, Birmingham, UK, July.

Scheible, Silke, Richard J Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, page 19–23, Portland, USA, June.

Straka, Milan and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In Jan Hajič and Dan Zeman, editors, *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Van der Goot, Rob, Alan Ramponi, Tommaso Caselli, Michele Cafagna, and Lorenzo De Mattei. 2020. Norm it! lexical normalization for Italian and its downstream effects for dependency parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6272–6278, Marseille, France, May. European Language Resources Association.

Weber, Daniel and Desislava Zhekova. 2016. TweetNorm: Text normalization on Italian Twitter data. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Germany, September.

Yang, Yi and Jacob Eisenstein. 2016. Part-of-Speech tagging for historical English. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1328, San Diego, California, June. Association for Computational Linguistics.