Annotating Manzoni: Challenges in the Annotation of Lemmas, POS and Features in "I Promessi Sposi"

Rachele Sprugnoli[1,*,†], Arianna Redaelli[2]

[1]Università Cattolica del Sacro Cuore, Largo Gemelli, 1, 20123 Milano, Italy
[2]Università di Parma, Via D'Azeglio, 85, 43125 Parma, Italy

Abstract

In this paper we introduce a dataset of *I Promessi Sposi* annotated with lemmas, UPOS tags, and features aligned with Universal Dependencies (UD). Three representative chapters from Manzoni's 1840 edition (791 sentences, almost 26 K tokens) were automatically tagged with UDPipe and fully manually corrected. Tailored guidelines extended standard UD practice with: (i) a double lemmatization approach, one that maintains archaic spellings and altered forms and one that normalizes lemmas, (ii) novel features that capture specific important characteristics of the novel, such as the use of apocopated and altered forms. Using the resulting dataset, we retrained the Stanza pipeline to obtain an in-domain model. Augmenting training data with ISDT sentences yielded further, although smaller, gains. Finally, a CRF sequence tagger was developed to identify apocopated forms.

Keywords

annotation, Italian literature, computational literary studies, Alessandro Manzoni

1. Introduction

In recent years, there has been a growing interest in the application of Natural Language Processing (NLP) to texts within the humanities, particularly in the literature domain. This trend is evidenced by the papers published in the proceedings of numerous conferences and workshops specifically dedicated to this area of research, such as those organized by the Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities of the Association for Computational Linguistics (LaTeCH-CLfL)[1], the Digital Literary Studies Special Interest Group of the Alliance of Digital Humanities Organizations (SIG-DLS)[2], or the Computational Humanities Research (CHR) community[3]. Other key venues include the International Conference on Natural Language Processing for Digital Humanities (NLP4DH)[4] and the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)[5].

The European consortium CLARIN has compiled a list of 45 literary corpora, each representative of a single author or a specific period.[6] However, this list does not include any Italian corpora. Nevertheless, literary texts are present within Italian diachronic corpora such as DiaCORIS [1], CODIT [2], and MIDIA [3]. The latter two include some works by Alessandro Manzoni, though not the complete texts but only selected portions. In contrast, the full text of *I Promessi Sposi* is accessible and searchable through platforms such as Intertext,[7] the LIZ (*Letteratura Italiana Zanichelli*) database,[8] and the CBook website [4]. However, there are currently no publicly available linguistic annotations nor any models that have been developed or tested specifically on the novel.

This paper aims to begin addressing this existing gap by offering the following contributions:

- A manually annotated dataset comprising three chapters of the novel, totaling 791 sentences and approximately 26,000 tokens. The annotations include lemmas, UPOS tags, and morphological features following the Universal Dependencies (UD) framework. Particular attention was given to (i) using features described in the Italian UD guidelines that are not yet widely adopted across existing treebanks, (ii) applying a dual lemmatization strategy (normalizing and conservative), (iii) defining additional features that capture stylistic and linguistic peculiarities of the novel.

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

*Corresponding author.

† This paper is the result of the collaboration between the two authors. For the specific concerns of the Italian academic attribution system, Rachele Sprugnoli is responsible for sections 2, 3.2, 4, 5; Arianna Redaelli is responsible for sections 1 and 3.1. Section 6 was collaboratively written by the two authors.

rachele.sprugnoli@unicatt.it (R. Sprugnoli); arianna.redaelli@unipr.it (A. Redaelli)
0000-0001-6861-5595 (R. Sprugnoli); 0000-0001-6374-9033 (A. Redaelli)

[1]https://sighum.wordpress.com/events/
[2]https://dls.hypotheses.org/
[3]https://computational-humanities-research.org/
[4]https://www.nlp4dh.com/home
[5]https://circse.github.io/LT4HALA/
[6]https://www.clarin.eu/resource-families/literary-corpora
[7]https://www.intratext.com/Catalogo/Autori/Aut246.HTM
[8]https://www.zanichelli.it/ricerca/prodotti/liz-4-0-letteratura-italiana-zanichelli

- An in-domain model trained on the aforementioned annotated dataset.
- A joint model trained on the combined data from *I Promessi Sposi* and the ISDT treebank.
- A dedicated model for the recognition of apocopated forms, which are characteristic of the novel's language.

All datasets and models are publicly available in a dedicated GitHub repository: https://github.com/RacheleSprugnoli/CoNLL-U_Manzoni.

2. Related Work

The application of NLP tools to Italian literary texts has been approached through targeted experiments since the early 2000s. Basili et al. [5] employed machine-learning techniques to semantically classify narrative fragments from Alberto Moravia's novel *Gli indifferenti*, whereas Pennacchiotti and Zanzotto [6] evaluated the accuracy of a morphological analyzer and a POS tagger on a range of prose and poetry texts dating from the thirteenth century to the late nineteenth century, revealing a drop in performance compared with results obtained on contemporary Italian. More recently, within the TrAVaSI project (*Trattamento Automatico di Varietà Storiche di Italiano*), texts of various genres, including literary works dated from 1861 onwards, have been annotated according to the UD framework, but using the same annotation layers we adopt for Manzoni, i.e., excluding dependency parsing. As in our study, these annotated data have been exploited to train automatic models [7]. Particular attention has been devoted to lemmatization, adopting a conservative approach that preserves the original token's graphical, phonological, and morphological characteristics [8]. By contrast, dependency parsing is included in the annotation of Dante Alighieri's *Divina Commedia*, which has in turn enabled the release of the Italian-Old treebank[9] and the development of models specifically tailored to this text [9]. In this annotation, lemmatization follows the criteria established in the *DanteSearch* project, from which the data were drawn [10] before applying the UD framework. In this case as well, a conservative strategy is adopted, whereby *pecorelle* ("little sheep") is lemmatized as *pecorella*. The same methodology has also been employed in the *Edizione dell'Opera Omnia di Luigi Pirandello* [11] and in the *Archivio Lessicografico della Poesia Italiana dell'Otto-Novecento* (ALPION) [12], although in these projects the data are accessible only through concordances.[10] Different lemmatization choices have been made in the compilation of other linguistic resources for Italian. For instance, in MIDIA, altered forms are linked to their corresponding base lemmas, but other word forms have not been normalized, resulting in distinct lemmas for each variation: for example, the archaic spelling *imaginando* ("imagining") is lemmatized as *imaginare*, while the modern form *immaginando* corresponds to *immaginare*. In COLFIS (*Corpus e Lessico di Frequenza dell'Italiano Scritto*), altered nouns and adjectives were initially lemmatized as independent lemmas and then a reference to the corresponding base form was added.[11] Finally, in LIPSI (*Lessico di frequenza dell'italiano parlato nella Svizzera italiana*), altered forms are mapped to a base lemma when weakly lexicalized: e.g., *chiesina* ("little church") is lemmatized with *chiesa* ("church"). On the contrary, independent entries are created when there is a significant semantic divergence between the derived form and the base: e.g., *lampadina* ("light bulb") is treated as a separate lemma with respect to *lampada* ("lamp") [13, 14]. This same strategy is also adopted in the compilation of the *Nuovo De Mauro* dictionary[12] and in our work, as explained in detail in Section 3.

3. Annotation

Chapters 1, 8, and 23 [13] of the final edition of *I Promessi Sposi* (1840) were automatically annotated with UDPipe 2 (ISDT model, version 2.15) [15] [16] and then manually corrected.[14] We adopted the CoNLL-U Plus format[15] to arrange specific annotation requirements designed for the novel, as explained in the following subsection (see Figure 1).

3.1. Guidelines

The annotation guidelines were developed collaboratively, discussed in multiple revision rounds, and refined to their current form. Their purpose was to guide the annotation process while remaining as consistent as possible with the official UD guidelines for Italian[16]. However, existing Italian treebanks do not always strictly follow UD's recommendations. Whenever discrepancies were

[9] https://github.com/UniversalDependencies/UD_Italian-Old

[10] https://vocabolari.pirandellonazionale.it/; https://alpion.unict.it/vocabolario/ricerca/

[11] https://linguistica.sns.it/CoLFIS/Home.htm

[12] https://dizionario.internazionale.it/avvertenze/2

[13] These chapters were selected for their stylistic and structural variety. Chapter 1 introduces the setting of the novel and includes a long descriptive passage, some dialogic sections, and even pseudo-documentary parts marked by archaic lexical choices; chapter 8 plays a central role in the narrative, featuring multiple scenes, thematic shifts, and dialogic exchanges, as well as a semi-lyrical closing section; chapter 23 is characterized by its predominantly dialogic structure and includes a lengthy final soliloquy.

[14] As we report in Table 2, the performance of this model is not optimal.

[15] https://universaldependencies.org/ext-format.html

[16] https://github.com/UniversalDependencies/docs/tree/pages-source/_it

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC ARCH-ALT:LEMMA
# sent_id = 1
# text = Il cardinal Federigo, intanto che aspettava l'ora d'andar in chiesa a celebrar gli ufizi divini, stava studiando,
com'era solito di fare in tutti i ritagli di tempo; quando entrò il cappellano crocifero, con un viso alterato.
1   Il          il          DET     RD  Definite=Def|Gender=Masc|Number=Sing|PronType=Art   _   _   _   _       *
2   cardinal    cardinale   NOUN    S   Gender=Masc|Number=Sing                             _   _   _   Variant=Apoc    *
3   Federigo    Federigo    PROPN   SP                                                      _   _   _   _   SpaceAfter=No   *
```

Figure 1: Snippet of CoNLL-U Plus format (beginning of Chapter 23).

encountered between the UD guidelines and currently available treebanks, our guidelines prioritized the official UD specifications. This involved both substitutions and additions.

Among the substitutions, we systematically replaced the use of `VerbForm=Ger`, commonly found in current treebanks for the traditional Italian gerund (e.g., *dicendo*, "saying"), with the correct label `VerbForm=Conv`. Similarly, for superlative adjective forms (e.g., *pessimo*, "very bad"), we replaced `Degree=Sup` with `Degree=Abs`.

Among the additions, we decided to use the feature `Reflex=Yes` for reflexive forms (e.g., *sé*, *si*, *proprio*, "him/her/itself", "themselves"): although this feature is listed among the ones to be used in Italian,[17] it is still rarely applied in most currently available treebanks.[18] We also annotated indefinite pronouns functioning as total quantifiers (e.g., *ogni*, "each", "every", *tutto*, "all", "everything" and *ciascuno*, "everyone", "each one") with the feature `PronType=Tot`, in line with the UD guidelines, despite its inconsistent use across current resources.

Beyond these additions, we introduced a set of features not prescribed by the UD Italian guidelines, but intended to account for morphosyntactic phenomena of particular historical or stylistic relevance in *I Promessi Sposi*. All such features were annotated in the MISC field.

Firstly, we used the feature `Variant=Apoc` to annotate apocopated forms, only excluding indefinite articles (e.g., *un*, "a"), which are fully grammaticalized in contemporary Italian and therefore not stylistically significant. As observed by Bianchi [17], Manzoni drew both on postconsonantal and postvocalic apocopes (e.g., respectively, *fecer* instead of *fecero*, "they did", and *cagion* instead of *cagione*, "cause") to evoke the rhythms and informality of spoken language, at times even extending beyond Florentine usage, which was his main language model. Unlike elisions, which involve the omission of a final vowel before an initial vowel and are graphically marked with an apostrophe, apocopes generally drop final phonemes regardless of the phonological context and are not marked. However, some apocopated forms in the novel, such as *que'* instead of *quei* ("those"), do include an apostrophe. In such cases, the apostrophe reflects a graphic convention rather than a genuine elision, and our annotation still treats these forms as apocopated.

Furthermore, we extended the set of possible values for the feature `Degree` to include morphological alterations, which are also frequently attested in the novel:

- `Degree=Dim` for diminutives (e.g., *casetta*, "little house");
- `Degree=Aug` for augmentatives (e.g., *spadone*, "big sword");
- `Degree=Pej` for pejoratives (e.g., *occhiacci*, "nasty eyes");
- `Degree=End` for endearments (e.g., *poverina*, "poor little girl").

Rather than relying exclusively on morphological structure, the annotation of this feature was guided by contextual interpretation, focusing on the expressive or affective nuance that the altered form conveys in each occurrence. As Perotti [18] noted, many of these altered forms were introduced by Manzoni only in later revisions of *I Promessi Sposi*, reflecting his pursuit of greater precision and expressive depth. The extended feature set was thus designed to capture and document this stylistic evolution through a consistent, context-sensitive, and fine-grained annotation approach. Altered forms were lemmatized in the third field with their standard, non-altered base forms; the altered lemma, instead, was reported in the eleventh field (e.g., *occhiacci*, "nasty eyes"; third field: *occhio*, "eye"; eleventh field: *occhiaccio* "nasty eye"). By lemmatizing altered forms under their standard base lemma, the annotation facilitates lexical querying and quantitative analysis, avoiding the dispersion of occurrences across multiple lemmas while preserving the expressive variation. For the same reason, namely to ensure consistency and semantic clarity in lexical analysis, fully lexicalized altered forms whose meaning significantly diverges from that of the base lemma were instead treated as independent lemmas (e.g., *cavallone*, "large water wave", was lemmatized separately from *cavallo*, "horse").

In a nineteenth-century corpus like *I Promessi Sposi*, lemmatization also required additional care to account for archaisms and diachronic variation. In all cases, we prioritized the modern form of the lemma as the primary entry, placing it in the third field, regardless of the degree of obsolescence or morphological variation. This criterion was adopted to support both practical usability and

[17] https://github.com/UniversalDependencies/docs/blob/pages-source/_it/feat/Reflex.md

[18] `Reflex=Yes` is currently present in the following treebanks: PUD (3 occurrences), ParTUT (14), OLD (2,346).

interpretive clarity: lemmatizing under a standard modern lemma ensures ease of information retrieval, even for users who may not be familiar with historical or literary Italian. However, such standardization was not pursued at the expense of losing linguistically significant traces of the novel's historical and stylistic identity. On the contrary, we aimed to preserve this richness by systematically annotating archaic and obsolete forms through a dedicated feature in the MISC field and/or an additional lemmatization in the eleventh field.

More specifically, in line with this approach, we distinguished two main cases for archaic forms:

- when the form was both obsolete and corresponded to an archaic lemma whose modern counterpart differed only in orthography or morphology (not in lexical identity or meaning), we annotated the feature `Style=Arch` in MISC field and reported the archaic lemma in the eleventh field (e.g., *annunzio*; field LEMMA: *annunciare*, "to announce"; MISC field: `Style=Arch`; eleventh field: *annunziare*);
- when the form was only the archaic spelling of a lemma that is still used today (i.e., the lemma itself was not obsolete), we only annotated `Style=Arch` in the MISC field without adding any lemma in the eleventh field (e.g., *varjo*; field LEMMA: *vario*, "various"; MISC field: `Style=Arch`). The same criterion was also applied to inflected forms that appear archaic but whose corresponding lemma is still current and unaltered (e.g., *chieggio*, which is the first person singular of *chiedere*, "to ask").

In case of uncertainty, we referred to *Nuovo De Mauro* [19], which provides mappings between obsolete or literary forms and their modern equivalents.

Finally, consistent with the principles outlined above, we applied a contextual approach to UPOS tagging and morphological features assignment, following the conventions of current Italian treebanks: for example, infinitives and participles were annotated as NOUN or ADJ when used as nouns or adjectives, respectively. In the case of infinitives used as nouns, no morphological features were assigned, as these forms are not inflected for gender or number. For participles, instead, the annotation also had consequences on lemmatization: when used as adjectives, they were lemmatized with the corresponding masculine singular form, in line with standard adjectives; when retaining a verbal function, they were lemmatized with the infinitive of the corresponding verb[19]. To help distinguish between participles and adjectives, we referred to Guasti [20], indicating three diagnostic tests, also adopted in the annotation of CoLFIS:

- participles cannot be modified with the suffix *-issimo* or intensifying adverbs (e.g., *molto*, "very"), while adjectives can;
- past participles can host clitic pronouns, while adjectives cannot;
- participles can co-occur with both *essere*, "to be", and *venire*, "to come", while adjectives can't.

3.2. Inter-Annotator Agreement

The IAA was calculated on the first 100 sentences of Chapter 38, the last one of the novel. This chapter is not part of the current dataset and the completion is in progress at the time of writing this paper. The annotators involved are two students of the Master's degree in "Linguistic Computing" at Università Cattolica del Sacro Cuore; they are Italian native speakers who have studied UD during a couple of courses of the degree but have not participated in the writing and discussion of the guidelines and are at their first experience of extensive annotation. Before beginning their work on Chapter 38, the students read the guidelines and analyzed the annotations already made for Chapters 1, 8, and 23.

The Cohen's kappa recorded for the different annotation levels was as follows:

- Lemmatization: 0.80;
- UPOS tagging: 0.97;
- Morphological features identification: 0.84;
- Other features: `Degree`, 0.80; `Style`, 0.86; `Variant`, 0.99.

Table 1
Cohen's kappa on the first 100 sentences of Chapter 38.

UPOS		Morphological Features	
X	1	Polarity	0.89
NUM	1	Definite	0.82
INTJ	1	Gender	0.81
PROPN	1	Foreign	0.8
PUNCT	0.99	NumType	0.8
NOUN	0.99	Number	0.8
CCONJ	0.99	Person	0.8
ADP	0.98	Clitic	0.78
VERB	0.98	VerbForm	0.78
PRON	0.96	Poss	0.77
AUX	0.96	PronType	0.77
DET	0.95	Tense	0.76
ADV	0.94	Mood	0.76
ADJ	0.92	Degree	0.45
SCONJ	0.89	Reflex	0.39

[19] As for present participles, their usage is almost exclusively limited to either a nominal or, more rarely, a verbal function. The nominal use is generally easy to identify, as present participles functioning as nouns are typically preceded by a determiner (e.g., an article).

Table 1 provides details on the Cohen's kappa achieved for each UPOS tag and morphological feature. Overall, the results for the various annotation levels are good, often above 0.80 (indicating substantial or almost perfect agreement), with a few exceptions only for some features.

As for lemmatization, there are 27 discordant lemmas that fall into 4 categories. Some cases are clear errors due to superficial annotation: e.g., in *si sana ogni piaga* ("every wound is healed"), *sana* is lemmatized as *sano* ("healthy") instead of *sanare* ("to heal"). A recurring issue concerns the lemmatization of unstressed personal pronouns. Sometimes, the lemma matches the token itself; other times, it corresponds to the masculine form: e.g., in *l'era stata compagna* ("she had been her companion"), *l'* is lemmatized with *le* (feminine) or with *lo* (masculine). Another disagreement concerns the lemmatization of words in an archaic form, which also has repercussions on the feature `Style=Arch`. For example, *pronunziar* ("to pronounce") is lemmatized alternatively as *pronunciare*, in this case by adding the feature `Style=Arch`, or as *pronunziare*, without the feature.

Regarding the annotation of UPOS tags, the lowest agreement is recorded on subordinate conjunctions, confused with adpositions (2 times), adverbs (4 times) and pronouns (7 times, always in the annotation of *che*, meaning "who", "which" or "that").

The results concerning the annotation of morphological features show greater variability. Notably, the features `Degree`, which is employed for marking comparative and superlative forms of adjectives and adverbs, and `Reflex`, which is used for reflexive pronouns, have relatively low kappa scores (0.45 and 0.39 respectively), indicating moderate and fair IAA. As mentioned in subsection 3.1, these features were subject to modifications that appear to have been insufficiently assimilated by the annotators. For instance, one annotator consistently employed the `Sup` value of `Degree` rather than `Abs` for absolute superlatives, and frequently omitted the `Reflex=Yes` feature.

By contrast, the level of agreement is high for the newly introduced features in the MISC column. An interesting example of annotation divergence concerns the token *figliuoli* ("children"): one annotator interprets it as an archaic form of the lemma *figlio* ("child") with an endearing suffix, whereas the other annotator assigns the lemma *figliuolo*, without marking it with either the `Degree=End` or `Style=Arch` features.

4. Retraining Stanza

The dataset was split into training, development, and test sets using an 80/10/10 ratio, with the division based on the number of syntactic words as units, in accordance with the guidelines of the UD framework. The number of syntactic words was taken proportionally equally from the three chapters. Following this approach, the partitions are the following:

- training set: 615 sentences, 20,806 tokens;
- development set: 101 sentences, 2,670 tokens;
- test set: 75 sentences, 2,457 tokens.

Using this partition, a new Stanza [21] model for Manzoni's novel has been developed.

Table 2 presents the performance of the retrained model on the test set, in comparison with results obtained on the same file from other models, namely the ISDT [15] and OLD [9] 2.15 models of UDPipe 2, as well as the spaCy `it_core_news_lg`[20] and the Stanza combined models. The retrained model outperforms the other evaluated ones across all tasks. Obviously this is also due to the different annotation choices, especially those related to the features (see Section 3).

All models are nearly equivalent and highly reliable in token segmentation. The biggest divergence occurs for sentence splitting: as previously shown by Redaelli and Sprugnoli [22], this task is challenging due to the distinctive punctuation of the novel, particularly the use of guillemets and long dashes as closing quotation marks, thus the development of a dedicated model is especially necessary. Syntactic word segmentation has high scores (> 90) across all models but spaCy proved to be the least reliable.

With regard to UPOS tagging, the retrained Stanza model achieves an improvement of 2.44 F1 points compared to the Stanza combined model. The tag with the lowest F1 score under the retrained setting is INTJ (F1=0.79, P=1, R=0.65). For example, the only occurrence of *ohimè* (a roughly equivalent interjection to "alas") is misclassified as a NOUN, while *addio*, "farewell", is classified three times as an INTJ and three times as a NOUN. All other tags have values above 0.80 but we can notice some recurring errors in the case of the SCONJ tag. Indeed, subordinating conjunctions (F1=0.85, P=0.84, R=0.85) are confused with prepositions (ADP, especially for *dopo*, "after"), pronouns (PRON, as in the case of *che*, "who/that"), or adverbs (ADV, as in the case of *dove*, "where").

As for Universal features (UFeats), the 3.71 point improvement over the Stanza combined model is likely due to differences in the handling of specific features such as `Reflex=Yes` and `VerbForm=Conv`. The features with the lowest F1 scores are `PronType=Int` (F1=0.50, P=0.50, R=0.50), which marks interrogative pronouns and determiners, and `PronType=Exc` (F1=0.44, P=0.67, R=0.33), which is applied to exclamative pronouns and determiners. These categories are sparsely represented in the test set, with only 8 and 6 instances respectively. However, there is evidence of confusion between the two: for example, in the sentence *"Come stava allora il povero don*

[20] https://spacy.io/models/it#it_core_news_lg

Table 2
F1 score of different models.

	UDPipe-ISDT	UDPipe-OLD	spaCy-large	Stanza-combined	Stanza-retrained
Token	99.87	99.94	99.81	99.81	100
Sentences	22.66	66.99	21.62	61.08	100
Words	98.32	95.11	92.05	98.03	99.63
UPOS	93.94	87.85	85.08	93.59	96.03
UFeats	93.94	75.57	86.01	93.10	96.81
Lemmas	95.29	88.55	85.50	94.29	97.13

Table 3
Examples of lemmatization errors involving altered (on the left) and archaic (on the right) forms.

FORM	LEMMA-GOLD	LEMMA-PRED
bravacci	*bravo*	*brave*
campicello	*campo*	*campice*
paesello	*paese*	*paesello*

FORM	LEMMA-GOLD	LEMMA-PRED
maraviglia	*meraviglia*	*meravigliare*
leggiero	*leggero*	*leggiere*
edifizi	*edificio*	*edifizio*

Abbondio!" ("How was poor Don Abbondio feeling at that moment!") the word *come*, "how", is annotated as `PronType=Exc` in the gold data, but the model incorrectly predicts `PronType=Int`. The feature `Mood=Cnd`, indicating verbs in the conditional mood, also yields a relatively low F1 score (F1=0.73, P=1, R=0.57). Although this class includes only a small number of instances (7), misclassifications occurred, including one case where it was confused with the indicative mood (*fiaterebbe*, "he would breathe") and another with the subjunctive mood (*leverebbe*, "he would take away").

For lemmatization, the improvement is of 2.84 points with respect to the Stanza combined model, with a total of 82 incorrect lemma predictions. Notably, lemmatization choices involving altered forms and archaic variants do not appear to be major sources of inaccuracy: indeed, only 12% of errors involve altered forms, and 4% involve archaic ones. Table 3 provides examples of these types of errors. The remaining instances mostly concern the prediction of non-existent lemmas (e.g., *riunendo* (gerund of "reunite") → *riunere* instead of *riunire*; *mangi* (present subjunctive of "eat") → *manire* instead of *mangiare*); and of feminine forms instead of the correct masculine ones (e.g., *scure* ("dark") → *scura* instead of *scuro*; *forestiera* ("female foreigner") → *forestiera* instead of *forestiero*). It is interesting to note that the UDPipe model trained on the *Divina Commedia* (UDPipe-OLD) exhibits low performance on lemmatization, despite the fact that the target domain is literary, as is the case for Manzoni. This discrepancy can likely be attributed to the considerable temporal and stylistic differences between the two sources: the *Divina Commedia* is dated back to the 14th century and is composed in poetic form, whereas Manzoni's work dates to the 19th century and is written in prose. Indeed, the lexical overlap between the lemmas in the training set of the OLD treebank and those in our corpus amounts to only 50%, compared to a higher overlap of 69% with the lemmas in the training set of the ISDT treebank.

4.1. One Novel, Three Versions

Alessandro Manzoni revised *I Promessi Sposi* multiple times, resulting in three versions. The earliest, a handwritten draft composed in 1823 and known as *Fermo e Lucia*, differs in both content and style from later editions. The language used, for example, is an original combination of Italian, Lombard, French and Latin calques, also rich in author's neologisms. In 1827, Manzoni published a revised version, commonly called the *Ventisettana*, which introduced substantial linguistic refinements aimed at improving clarity and accessibility for Italian readers. The definitive version, released starting from 1840 and known as the *Quarantana*, incorporated further stylistic and linguistic changes based on the Florentine language, reflecting Manzoni's efforts to promote a unified Italian language.

Given the linguistic differences among these versions, it is of particular interest to assess the extent to which the model trained on the *Quarantana* generalizes to earlier texts. Table 4 presents the F1 scores obtained in the first chapter of *Fermo e Lucia* (5,760 tokens) and the *Ventisettana* (7,407 tokens). Notably, performance on the *Ventisettana* is even higher in terms of morphological features and lemmatization, although there is a slight decrease in UPOS tagging. Morphological features identification is still good on the 1823 version but UPOS tagging and lemmatization show a more evident drop.

4.2. A Joint Model

An additional experiment involved the creation of a combined model trained on the merged training and development sets of ISDT and the training set of *I Promessi Sposi*. ISDT was selected because its corresponding model

Table 4
F1 score on the first chapter of the two previous versions of *I Promessi Sposi*. Bold scores indicate an improvement over the results obtained on the *Quarantana* test set.

	Fermo e Lucia	*Ventisettana*
UPOS	95.78	95.87
UFeats	**97.12**	**97.56**
Lemmas	95.15	**97.55**

achieved better results than the other off-the-shelf models, although it still underperformed compared to the in-domain retrained model. The resulting combined training set consisted of 14,300 sentences.

Table 5 reports the performance of this combined model on the first chapters of *Fermo e Lucia* and the *Ventisettana*, as well as on the test set from the *Quarantana*. The increased training data, despite being from a different domain and not always consistent with our annotation guidelines, led to a modest overall improvement in performance, particularly on the 1840 test set. These generally positive results align with findings from previous experiments conducted on the *Voci della Grande Guerra* [23] and *VoDIM* [7] corpora. In contrast, joint models developed for syntactic parsing of the *Divina Commedia* have shown lower performance compared to in-domain models [24].

Table 5
F1 score of the joined (ISDT+Manzoni) model on the test set of *I Promessi Sposi* (*Quarantana*) and on the first chapter of the previous novel's versions. In bold the score that are improved with respect to the ones obtained with the in-domain mode.

	Fermo e Lucia	*Ventisettana*	*Quarantana*
UPOS	**95.95**	94.35	**96.24**
UFeats	96.86	96.11	**97.14**
Lemmas	**96.58**	**97.61**	**97.50**

5. Modeling Apocopes

We implemented a supervised sequence labeling pipeline for identifying apocopated forms using Conditional Random Fields (CRFs) and the same train, development and test sets used for the retraining of Stanza. For the time being, we have focused on apocopated forms only, as among the three specific features we added to the annotation, `Variant=Apoc` is the most frequent, whereas the others are too sparsely represented.[21] Although more frequent than the other features, the number of instances was still insufficient to support the use of neural methods, which require larger amounts of training data to perform effectively and generalize well. Therefore, we adopted a CRF-based approach instead.

The model is trained using the `sklearn-crfsuite` library and hyperparameters (c1 and c2 regularization coefficients) are optimized via randomized search with 5-fold cross-validation. The feature set includes orthographic (e.g., lowercase form, word suffixes and prefixes), morphological (e.g., UPOS and FEATS) and lexical (lemma) features from the preceding and following tokens. The results of the model's binary classification on the test set are reported in Table 6.

Table 6
Results of the CRF model on the `Variant=Apoc` feature.

	P	R	F1
Apoc	1	0.85	0.91
None	1	1	1
Avg.	1	0.92	0.95

The test set contains 59 apocopated forms corresponding to 41 tokens and 33 lemmas; 12 of these forms do not appear in the training set, which includes 611 apocopated instances corresponding to 220 tokens and 169 distinct lemmas. Among the model's 9 false negatives, 4 are apocopated forms that were not seen during training: i.e., *timor* ("fear"), *almen* ("at least"), *passan* ("they pass by"), *ondeggiar* ("to ripple"). As for the remaining cases, the model fails to correctly classify *par* ("it seems", seen 7 times in the training set), *fra* ("friar", 3 times), *star* ("to stay", 2 times), and *siam* and *cagion* ("we are" and "cause", each seen once in the training data).

6. Conclusion

In this paper, we have introduced several new resources: (i) a manually annotated dataset of 3 chapters of *I Promessi Sposi*, comprising 791 sentences and approximately 26,000 tokens, enriched with lemmas, UPOS tags, Universal Dependencies morphological features and ad-hoc features designed for capturing specific stylistic characteristics of Manzoni's novel; (ii) an in-domain NLP model trained specifically on this dataset; (iii) a joint model combining data from the novel and the ISDT treebank; (iv) a specialized model for recognizing apocopated forms, which are a distinctive feature of Manzoni's text.

All data and models developed in this study are made publicly available in a dedicated GitHub repository, hopefully laying the groundwork for future research on Italian literary texts through computational approaches.

As for future work, a key priority is to extend the annotation to additional chapters. Thanks to the new models developed in this study and their relatively low error rates, the manual correction process is expected to be significantly accelerated. The expansion

[21] The whole dataset, at the moment of writing, contains 735 apocopated forms, 109 altered forms and 106 archaic forms.

of the dataset will also enable the development of models targeting the other two specific features introduced in our annotation scheme, namely `Style=Arch` and `Degree=Aug/Dim/End/Pej`. Another future step will involve syntactic annotation, with the ultimate goal of incorporating Italy's most important novel among the UD treebanks. This will continue the broader effort to integrate Italian literary texts into syntactically annotated resources, following the precedent set by the annotation of the *Divina Commedia* [9].

Acknowledgments

The authors thank Flavio Massimiliano Cecchini for annotating chapters 1, 8, 23 of *Quarantana* and *Ventisettana*, Alessia Leo and Michael Mostacchi for annotating chapter 38 of *Quarantana*, Chiara Febbraro for the annotation of chapter 1 of *Fermo e Lucia* and Giovanni Moretti for technical assistance.

References

[1] C. Onelli, D. Proietti, C. Seidenari, F. Tamburini, The DiaCORIS project: a diachronic corpus of written Italian, in: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, D. Tapias (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: https://aclanthology.org/L06-1371/.

[2] M. S. Micheli, Codit. a new resource for the study of italian from a diachronic perspective: Design and applications in the morphological field, Corpus (2022).

[3] P. D'Achille, C. Iacobini, Il corpus midia: concezione, realizzazione, impieghi, Corpora e Studi Linguistici (2022) 207.

[4] A. Bolioli, M. Casu, M. Lana, R. Roda, Exploring the betrothed lovers, in: 2013 Workshop on Computational Models of Narrative, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2013, pp. 30–35.

[5] R. Basili, A. Di Stefano, R. Gigliucci, A. Moschitti, M. Pennacchiotti, et al., Automatic analysis and annotation of literary texts, in: Wokshop on Cultural Heritage, 9th AIIA Conference, Milan, Italy, 2005.

[6] M. Pennacchiotti, F. M. Zanzotto, Natural language processing across time: an empirical investigation on italian, in: International Conference on Natural Language Processing, Springer, 2008, pp. 371–382.

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

[8] M. Favaro, M. Biffi, S. Montemagni, et al., Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione, in: Proceedings of the 16th international conference on statistical analysis of textual data, Edizioni Erranti di S. Pellegrino, 2022, pp. 392–399.

[9] C. Corbetta, M. Passarotti, F. M. Cecchini, G. Moretti, Highway to hell. towards a Universal Dependencies treebank for dante alighieri's comedy, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 154–161. URL: https://aclanthology.org/2023.clicit-1.20/.

[10] 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.

[11] A. Di Silvestro, A. Sichera, «pirandellonazionale». una scommessa filologica ed ermeneutica, Griseldaonline 20 (2021) 173–180.

[12] A. Di Silvestro, C. D'Agata, G. Palazzolo, P. Sichera, Conservazione e fruizione di banche dati letterarie: l'archivio della poesia italiana dell'otto/novecento di giuseppe savoca, Atti del Convegno AIUCD (2022) 98–104.

[13] E. M. Pandolfi, LIPSI: Lessico di frequenza dell'italiano parlato nella Svizzera italiana, Osservatorio linguistico della Svizzera italiana Bellinzona, 2009.

[14] M. Prada, Lipsi. il lessico di frequenza dell'italiano parlato in svizzera, Italiano LinguaDue 2 (2010) 182–182.

[15] C. Bosco, S. Montemagni, M. Simi, et al., Converting italian treebanks: Towards an italian stanford dependency treebank, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, The Association for Computational Linguistics, 2013, pp. 61–69.

[16] M. Straka, Udpipe 2.0 prototype at conll 2018 ud shared task, in: Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies, 2018, pp. 197–207.

[17] E. Bianchi, I promessi sposi e il parlar fiorentino, Annali Manzoniani III (1942) 281–312.

[18] P. A. Perotti, Alcuni alterati nei promessi sposi: studio lessicale-statistico, Rivista di Letteratura italiana XXXII (2014) 55–70.

[19] T. De Mauro, Il dizionario della lingua italiana, n.d. URL: https://dizionario.internazionale.it, accessed May 26, 2025.

[20] M. T. Guasti, Il sintagma aggettivale, in: L. Renzi, G. Salvi, A. Cardinaletti (Eds.), Grande grammatica italiana di consultazione, vol. II, libreriauniversitaria.it Edizioni, 2022, pp. 321–340. First published in 1991 by Il Mulino. Anastatic reprint.

[21] 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, 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[22] A. Redaelli, R. Sprugnoli, Is sentence splitting a solved task? experiments to the intersection between NLP and Italian linguistics, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 813–820. URL: https://aclanthology.org/2024.clicit-1.88/.

[23] I. De Felice, F. Dell'Orletta, G. Venturi, A. Lenci, S. Montemagni, et al., 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.

[24] C. Corbetta, G. Moretti, M. Passarotti, Join together? combining data to parse Italian texts, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 251–257. URL: https://aclanthology.org/2024.clicit-1.30/.