Evaluating Sentence Splitting on 19th-Century Italian Novels: a Comparative Analysis Across Approaches

Arianna Redaelli*
Università di Parma

Rachele Sprugnoli**
Università Cattolica del Sacro Cuore

The automatic segmentation of raw text into individual sentences, known as sentence splitting or sentence segmentation, is a fundamental task in text processing. Although it is often considered to be solved in standard domains such as news articles and Wikipedia pages, the performance of the system can vary significantly between different textual genres. This study evaluates eight sentence splitting tools employing rule-based, supervised, semi-supervised, and unsupervised approaches, and additionally tests two Large Language Models in a zero-shot setting, on a corpus of 19th-century Italian novels, namely "I Promessi Sposi", "I Malavoglia", "Le avventure di Pinocchio", and "Cuore". In addition, we train new sentence splitting models using the Stanza pipeline, creating individual models for each novel as well as a combined model trained on all available data. This work aims to highlight that, although literary texts have received relatively little attention in sentence segmentation research, they offer a rich and promising intersection between Natural Language Processing, Italian linguistics, and the Digital Humanities.[1]

1. Introduction

Sentence splitting is a Natural Language Processing (NLP) task whose aim is segmenting a text into sentences. By "sentence" we refer to a coherent sequence of words, constructed in accordance with the general rules of the language, and conveying a complete thought that is interpretable in isolation (Bonomi et al. 2010). A sentence typically ends with a strong punctuation mark (e.g., full stop, question mark, or exclamation point) and is usually followed by a capital letter. The definition adopted here, while inevitably subject to theoretical and practical limitations, is tailored to the specific objectives of the present study, as will become evident in the sections that follow.

Sentence splitting involves identifying sentence boundaries, which, in many Western languages including Italian, generally correspond to specific punctuation marks (Palmer 2000). Consequently, for a number of languages, sentence splitting can be viewed as a problem of punctuation disambiguation, i.e., determining whether a punctuation mark actually signals the end of a sentence.

Although often regarded as a trivial or solved task, sentence splitting plays a crucial role in many downstream NLP applications. Indeed, its accuracy has a significant influ-

* Università di Parma, Via D'Azeglio, 85, 43125 Parma, Italy. E-mail: arianna.redaelli@unipr.it
** Università Cattolica del Sacro Cuore, Largo Gemelli, 1, 20123.
E-mail: rachele.sprugnoli@unicatt.it

1 This paper is the result of the collaboration between the two authors. For the specific concerns of the Italian academic attribution system: Arianna Redaelli is responsible for Sections 1, 3, 5; Rachele Sprugnoli is responsible for Sections 2, 4, 6, 7, 8.1. Sections 8 and 9 were collaboratively written by the two authors.

ence on the performance of other tasks such as syntactic analysis (Dridan and Oepen 2013), machine translation (Wicks and Post 2022), and automatic summarization (Liu and Xie 2008). Errors introduced at this early stage can propagate through the analysis, leading to a substantial decline in overall system performance.

Widely used pipeline-based models, such as those implemented in *Stanza* (Qi et al. 2020) and *spaCy*[2], are typically trained and evaluated on relatively formal texts, including journalistic prose and Wikipedia articles. As a result, publicly reported performance scores, often above 0.90 in terms of F1, tend to reflect success in highly structured and standardized genres. However, text genre has a clear effect on sentence splitting accuracy. For example, in the CoNLL 2018 shared task "Multilingual Parsing from Raw Text to Universal Dependencies" (Zeman et al. 2018), the top-performing system achieved an F1 score of 0.99 on the Italian ISDT treebank (Bosco, Montemagni, and Simi 2013), which is composed of formal texts. In contrast, the best F1 on the PoSTWITA treebank, which consists of tweets (Sanguinetti et al. 2018), was only 0.66.[3]

These genre-specific variations suggest that focusing on less formal and more stylistically complex texts could yield deeper insights into the limitations and capabilities of current sentence splitting systems. Among such genres are literary texts, which often exhibit idiosyncratic and creative stylistic features that diverge from normative grammar and punctuation conventions (Tonani 2010b). These features are influenced by both individual authorial choices and the broader cultural and historical context. In particular, punctuation conventions may vary significantly across historical periods: literary works may conform to dominant trends or deliberately oppose them, thereby contributing to new stylistic norms.

This dynamic is especially evident in the 19th century, a period when the Italian *usus punctandi* underwent a significant shift from a primarily syntactic or pause-based usage, as prescribed by grammar handbooks, to a more communicative and textual interpretation of punctuation marks (Ferrari 2018). This transition was likely shaped by both theoretical reflections and the practical applications of major literary figures, notably Alessandro Manzoni (Mortara Garavelli 2003). For this reason, the present study focuses primarily on his historical novel, "I Promessi Sposi".

Manzoni devoted meticulous attention to the punctuation of his novel, overseeing revisions down to the final printed proofs. His approach combined classical conventions with personal and sometimes innovative choices made in collaboration with the publisher (Manzoni, Ghisalberti, and Chiari 1954). For example, in direct speech, low quotation marks («,») are used to mark spoken dialogue, long dashes (–) indicate internal thoughts, while italics is used to quote material from other written sources. Although these conventions are not always applied consistently, they render the novel particularly rich and challenging from a sentence segmentation perspective.

Furthermore, "I Promessi Sposi" holds a central place in the evolution of the Italian written language. Many of Manzoni's punctuation choices were later adopted as prescriptive models in grammar manuals, although only a subset of these choices has become part of the contemporary standard. Because punctuation was still undergoing codification during this period, its use reflected not only the norms of the time but also factors such as authorial style, text type, and even typographic constraints.

To gain a more comprehensive understanding of 19th-century punctuation practices, this study also includes excerpts from other literary works contemporaneous

2 https://spacy.io
3 https://universaldependencies.org/conll18/results-sentences.html

with Manzoni's. Specifically, we analyze "I Malavoglia" (1881) by Giovanni Verga, "Le avventure di Pinocchio. Storia di un burattino" (1883) by Carlo Collodi, and "Cuore" (1886) by Edmondo de Amicis. These novels offer a range of stylistic and punctuation strategies that reflect the diversity of literary practices of the time.

1.1 Contributions

The main contributions of this paper are as follows:

1. We evaluate the performance of different sentence splitting tools, each employing distinct methodological approaches, and two Large Language Models (LLMs) in a zero-shot setting, on a genre that is both stylistically and historically complex: 19th-century Italian literary fiction.
2. We analyze the results from the perspective of humanities scholars, particularly Italian linguists who are the principal stakeholders in the domain. In doing so, we aim to foster meaningful cross-disciplinary collaboration between NLP and Digital Humanities.
3. We release manually segmented data for four canonical 19th-century Italian novels, along with an open-access Google Colab notebook enabling the replication and extension of our experiments.[4]
4. We train and evaluate new sentence splitting models using the `Stanza` pipeline (Qi et al. 2020), creating individual models for each novel as well as a joint model trained on the combination of all available data.

2. Related Work

Sentence segmentation systems can broadly be classified into three macro-categories based on the underlying methodological approach. First, rule-based systems, such as `Sentence Splitter`[5] and the `Sentencizer` component of spaCy, rely on handcrafted heuristics tailored to specific languages, along with curated lists of exceptions and abbreviations. Second, supervised systems require annotated corpora in which sentence boundaries are explicitly marked for model training. Notable examples include `UDPipe` (Straka 2018) and `Stanza`, which are trained on treebanks from the Universal Dependencies (UD) initiative (De Marneffe et al. 2021). Finally, unsupervised systems are trained on raw, unsegmented text using features such as word length, punctuation patterns, and collocational information. A well-known system in this category is `Punkt` (Kiss and Strunk 2006), included in the NLTK library.

In this study, we empirically evaluate eight representative sentence segmentation systems, spanning all three aforementioned categories, on a benchmark dataset composed of historical literary fiction in Italian. This evaluation aims to assess their robustness in a domain where non-standard usage of punctuation is common.

While several works investigate the influence of text genre on sentence segmentation, literary texts remain underexplored. For example, Liu et al. (2005) proposed a CRF-based model incorporating prosodic and lexical features for sentence splitting in conversational speech. Rudrapal et al. (2015) addressed the challenge of noisy user-generated text in social media using a hybrid combination of rules and statistical models. Griffis et al. (2016) highlighted the limitations of general-purpose systems in clinical narratives,

4 https://github.com/RacheleSprugnoli/Sentence_Splitting_Manzoni
5 https://github.com/mediacloud/sentence-splitter

emphasizing the need for domain-specific adaptation. More recently, Sheik et al. (2022) introduced a deep learning model tailored for legal texts, demonstrating the advantages of lightweight neural architectures in specialized domains. These studies collectively underscore the importance of adapting sentence splitting techniques to the linguistic characteristics of specific text types. An initiative that further highlights the importance of focusing on domain-specific corpora is the FinSBD series of shared tasks, organized as part of workshops dedicated to the application of NLP tools in the financial domain. Specifically, the three editions of FinSBD ("Financial Sentence Boundary Detection"), held between 2019 and 2021, addressed the processing of financial prospectuses, that is official PDF documents in which investment funds describe their characteristics and investment strategies in detail (Azzi, Bouamor, and Ferradans 2019; Au et al. 2020; Au, Ait-Azzi, and Kang 2021).

It is important to note that most existing studies center on English-language data, with limited focus on other languages, including Italian. Two notable exceptions are the work of Brugger et al. (2023) and the "1st Shared Task on Sentence End and Punctuation Prediction in Natural Language Generation (NLG) Text" (SEPP-NLG 2021) held at the SwissText conference in 2021. The former offers a comprehensive evaluation of multilingual legal texts, including a detailed analysis of Italian. As we will show in our experiments, also in their work baseline off-the-shelf systems performed poorly on legal data due to the linguistic features of these texts, such as nested clauses, enumerations, and ambiguous punctuation. Thus the authors trained monolingual models (CRF, BiLSTM-CRF, Transformer) for each language using a dataset of over 130K manually annotated sentences spanning legal judgments and laws. Performances on Italian are strong, with the transformer-based model outperforming the other approaches (mean F1 score: 0.99). A multilingual Transformer model further demonstrated robust cross-language transfer. As for the SEPP-NLG 2021 shared task, its focus was on predicting sentence boundaries and punctuation in outputs from NLG systems across four languages, that is German, French, Italian, and English. Participants predominantly employed transformer-based classification models achieving high F1 scores (greater than or equal to 0.90) for sentence-end detection and common punctuation prediction (periods, commas, question marks) on in-domain Europarl data. Performance notably declined on out-of-domain TED talk transcripts, highlighting domain transfer challenges. For Italian, top systems achieved F1 scores around 0.92 for sentence-end detection and 0.76 for punctuation prediction, with hyphens and colons proving most difficult.

In addition to building on the above-mentioned works, our work draws methodological inspiration from Read et al. (2012), who evaluated sentence splitting systems on English literary texts, including the Sherlock Holmes stories. However, we extend this line of inquiry by turning to the Italian literary canon and, in particular, to 19th-century novels. These texts pose unique challenges due to their historical variation in punctuation practices, syntactic flexibility, and the evolving norms of Italian orthography. By focusing on this underrepresented domain, we aim to highlight both the limitations of current sentence segmentation systems and the importance of genre-sensitive approaches in NLP.

3. 19th-century Italian Punctuation

The development of Italian modern punctuation is closely linked to the rise and spread of movable type printing, which promoted both experimentation with different punctuation systems and the search for standardization. However, ambiguity and variation

in the number, type, and use of punctuation marks persisted well into the 19th century (Mortara Garavelli 2008).

The number and type of punctuation marks varied greatly over the centuries, as documented by the main Italian grammars. In fact, until the late 18th century, only the full stop, comma, colon, and semicolon were stably considered punctuation marks, sometimes accompanied by the question mark and exclamation mark, which were considered expressive variants of the full stop. The other punctuation marks (i.e., parentheses, suspension points, quotation marks, hyphens, and dashes), gradually introduced in writing, were added by grammarians without any real consistency, sometimes being categorized as diacritical signs rather than punctuation marks; on the opposite side, a new line was sometimes counted among the punctuation marks. This means that the use of many punctuation marks remained largely under- or unregulated for a long time, until a more stable classification was gradually achieved between the first and the second half of the 19th century.

In terms of usage, instead, two prevailing approaches took shape. The first approach interpreted punctuation marks as regulatory signals for pauses in oral reading: schematically, from this perspective, the full stop indicated a long pause, the comma a shorter pause, and the colon and semicolon an intermediate pause. The second approach was syntactic, that is, based on sentence structure: the full stop closed a complete sentence, the comma separated short sentence elements, and the colon or semicolon (often used interchangeably) separated longer ones. Moreover, in line with the second approach, punctuation marks should be regulated by strict grammatical rules: for instance, the comma was always prescribed before relative pronouns. Neither approach fully dominated: on the contrary, they frequently coexisted within the same normative text. To this situation must be added the fact that punctuation was generally granted a high degree of arbitrariness, often being treated as a matter of personal style or rhetorical effect rather than being governed by shared linguistic rules. This helps explain why, even throughout the 19th century, punctuation remained inconsistent in actual usage.

Such instability can be seen in printed works, where both the personal choices of the writer and the influence of the publisher affected typographical tendencies, but even more markedly in texts not intended for publication, such as private letters (Antonelli 2008) or early manuscript drafts of literary works. Manzoni's "Fermo e Lucia" belongs to the latter category. In these kinds of texts, in addition to the inconsistency of standard punctuation marks, unconventional symbols occasionally appeared (i.e., punctuation marks that were never officially stabilized by grammars): a key case is the double dash (=), frequently found in handwritten texts from the 18th and 19th centuries generally covering the functions of the dash, and formally codified in Giovanni Gherardini's mid-nineteenth-century grammar (Gherardini 1843), but no longer recorded in subsequent normative works.

As anticipated in Section 1, another important development emerged in the Italian context starting from the second half of the 19th century: the growing tendency to approach punctuation marks from a textual perspective [6]. This shift meant considering punctuation not only as a tool for syntactic structuring or prosodic indication, but also as a meaningful element for organizing and interpreting texts: a significant and clear example of this approach is represented by the colon, which came to be described as a

6 The textual value of punctuation had already been partially theorized by some Enlightenment-era French grammarians, who anticipated the idea of punctuation as a tool for organizing discourse and guiding interpretation (Lorenceau 1980).

mark used to introduce direct speech or to signal the explicit elaboration of preceding content. Within this framework, the two previous approaches (syntactic and pause-based) were not abandoned, but rather integrated into a broader conception of punctuation as a textual device. Since this perspective, which would gradually become the dominant interpretive model in the study of Italian punctuation (Ferrari et al. 2018)[7], was especially observable in the second half of the 19th century, especially in grammars influenced by Manzoni's choices, "I Promessi Sposi" may have played a role, directly or indirectly, in shaping this evolving understanding of punctuation (Ferrari 2020).

4. Tools

Sentence splitting is a fundamental analysis in text processing, for which there are many tools available, also for Italian. For our evaluation we have selected eight systems developed with different approaches, and we additionally tested two commercial LLMs.[8] Some tools are modules integrated in larger pipelines, others are systems specifically created to perform only sentence splitting, while the LLMs were employed in a zero-shot setting via prompting. It is important to note that selected tools do not split in the presence of a colon or semicolon. Indeed, although recent studies in the punctuation field identify colons and semicolons as punctuation marks capable of indicating the boundary of a sentence (Ferrari et al. 2018), in this work we have decided to not consider them as separating marks because of the various forms literary texts can take. To clarify the issue, we can consider the example of direct speech. In "I Promessi Sposi", direct speech can be introduced by a *verbum dicendi* and a colon, continuing without any interruption. In such cases, splitting at the colons would be relatively easy. However, direct speech can also be embedded within a sentence that continues after the quotation closes, creating a non-autonomous text portion that, during sentence splitting, should be manually reconnected to the one preceding the quotation itself (e.g., *Lucia sospirò, e ripeté: «coraggio,» con una voce che smentiva la parola.* EN: *Lucia sighed, and repeated, «courage,» in a voice that belied the word.*). An equally troublesome problem arises when the diegetic frame follows the quotation instead of preceding it. When this happens, the colons are absent, and other punctuation marks like commas are found before the closing quotation marks or dash (e.g., *«È il mio caso,» disse Renzo.* EN: *«That's my case,» said Renzo.*). The system would not split the sentences at these punctuation marks,

7 While this overview focuses on the 19th Italian tradition, similar distinctions between pause-based, syntactic, and discourse-organizational interpretations of punctuation have been discussed in international scholarship. Writing primarily with reference to English, Nunberg (1990) argues that punctuation does not transcribe speech prosody but functions as a conventional system showing syntactic relations and textual organization. From this perspective, the gradual shift observed in 19th century Italian grammars toward a more textual understanding of punctuation can be seen as part of a broader reorientation in Western writing practices (Parkes 1992; Crystal 2015).These frameworks are not applied systematically here, since the present study is historically and philologically grounded in 19th Italian usage; however, they provide a useful comparative backdrop. Moreover, since punctuation marks may simultaneously signal pauses, syntactic structure, discourse organization, and processing cues, their functions often overlap. In historical texts in particular, this functional layering makes it difficult to recover a single underlying principle; consequently, segmentation behaviour cannot be straightforwardly classified according to discrete punctuation models, nor can divergences be attributed with certainty to a single interpretive principle.

8 We made additional experiments with smaller, open-source LLMs, such as `Minerva-7B-instruct` and `gpt-oss-20b`, which, however, did not yield satisfactory results. The main issues observed were a tendency to alter words in the output text and to segment only the first sentences of the input, likely due to limitations in context window size and reduced robustness in handling longer passages. Consequently, these models were not included in the final evaluation.

yet the diegetic frame following the direct speech has the same value and autonomy as the one preceding it. Consequently, considering colons and semicolons as sentence boundaries would make the segmentation much more complex and often inaccurate.

We selected the following tools:

- `CoreNLP`[9]: an NLP pipeline written in Java and developed by Stanford University (Manning et al. 2014). It contains various modules including `ssplit` that divides a text into sentences via a set of rules. The latest version of the pipeline (4.5.7) supports eight languages including Italian.
- `spaCy`: an open-source NLP library which supports dozens of languages, including Italian, and provides four alternatives for sentence splitting. Among these, statistical models for Italian have been trained to split on colons and semicolons. For this reason, we tested the performance only of `Sentencizer`, the rule-based pipeline component.
- `Sentence Splitter`[10]: a Python module based on scripts developed for processing the Europarl corpus (Koehn 2005). It supports several languages with ad-hoc rules.
- `UDPipe`[11]: an NLP pipeline based on the UD framework performing tokenization, sentence splitting, PoS tagging, lemmatization and syntactic analysis. UDPipe 2 is written in Python and uses the tokenizer of UDPipe 1; among the 131 most recent models (version 2.15), eight are for Italian. We evaluated the model trained on the VIT (Delmonte, Bristot, and Tonelli 2007) and Italian-Old (Corbetta et al. 2023) treebanks that does not (always) split at colons and semicolons. VIT includes news, bureaucratic, political, financial, scientific, literary texts plus spoken dialogues whereas Italian-Old contains the text of Dante Alighieri's "Divina Commedia".
- `Stanza`[12]: an NLP package written in Python and based on neural network components. Sentence splitting is jointly performed with tokenization by the `TokenizeProcessor` module. The default Italian model is a combination of multiple UD treebanks.[13]
- `Ersatz`[14]: a language-agnostic neural model based on a semi-supervised training paradigm. It combines the use of regular-expressions to detect candidate sentence boundaries with a Transformer-based binary classifier (Wicks and Post 2021).
- `Punkt`: an unsupervised system which uses collocational information to identify abbreviations, initials, and ordinal numbers. All punctuation not included in these elements is considered an end-of-sentence marker.
- `WtP`[15]: an unsupervised multilingual sentence segmentation system based on a self-supervised learning approach tested on 85 languages, including Italian. It does not rely on punctuation or sentence-segmented training data thus it is a punctuation-agnostic system (Minixhofer, Pfeiffer, and Vulić 2023). Among the various available models, we adopted the `wtp-canine-s-12l` which, according

9 https://stanfordnlp.github.io/CoreNLP/
10 https://github.com/mediacloud/sentence-splitter
11 https://ufal.mff.cuni.cz/udpipe
12 https://stanfordnlp.github.io/stanza/
13 Namely, ISDT, VIT, PoSTWITA, and TWITTIRO, https://stanfordnlp.github.io/stanza/combined_models.html.
14 https://github.com/rewicks/ersatz
15 https://github.com/segment-any-text/wtpsplit

to the official documentation of the tool, has the best results on languages other than English.

- `GPT-5.2`: a large language model developed by OpenAI, prompted in a zero-shot setting to perform sentence splitting as a sequence labeling task. Unlike traditional tools, it does not rely on explicit tokenization or rule-based boundary detection, but infers sentence boundaries from contextual and syntactic cues acquired during large-scale pretraining.
- `Claude Sonnet 4.6`: a large language model developed by Anthropic, employed the same zero-shot prompting setting as GPT 5.2. Sentence boundaries are identified by instructing the model to return the input text segmented into individual sentences, relying on the model's in-context reasoning capabilities rather than explicit supervision.

For the evaluation, the tools were used as they are, with their default configurations, without making any customization. For this reason, given the choices motivated above, we did not consider other systems, such as Tint (Palmero Aprosio and Moretti 2018), which by default split at colons and semicolons. As for the LLMs, the following zero-shot prompt was used: *Segmenta in frasi il testo in input. Restituisci in output l'intero testo originale con ogni frase su una riga diversa.*[16] The prompt was submitted twice to each model in order to assess the consistency of the output across runs.

5. Dataset

The data used to evaluate the aforementioned tools are taken from "I Promessi Sposi" in its final version published in 1840-1842[17]. 3,095 sentences, corresponding to 12 chapters of the novel, were manually split. This dataset was divided into training, development and test sets according to the proportions 80/10/10 and using the UD rules for which this proportion was calculated using syntactic words as units.[18] To obtain syntactic words and calculate this splitting, sentences were segmented and tokenized by hand; this gold standard was then processed with the combined Stanza model.[19] Following this division, the test set is made of 324 sentences.

Table 1
End-of-sentence markers in the "I Promessi Sposi" test set.

MARK	# TOTAL	# EOS
.	277	237 (86%)
»	90	53 (59%)
?	47	22 (47%)
!	31	6 (19%)
…	23	3 (13%)
–	10	3 (30%)

16 EN: *Segment the input text into sentences. Return the entire original text as output, with each sentence on a separate line.*

17 The text, fully digitized and available online, was collated with the reference edition (Manzoni and Colli 2024) prior to analysis, to ensure maximum fidelity to the author's punctuation choices.

18 `https://universaldependencies.org/release\_checklist.html\#data-split`

19 The output of this process was used to train a new Stanza model as reported in Section 7.

Table 1 shows the sentence-ending punctuation marks in the test set. Both the total number of occurrences (TOTAL) and the number of times a sign is an end-of-sentence marker (EOS) are reported together with the corresponding percentage. In addition to the full stop, sentence boundaries can be indicated by expressive punctuation marks (!, ?) when followed by a capital letter. If followed by a lowercase letter, instead, these marks only have an expressive function, modifying the sentence's internal intonation without determining its end. Low quotation marks (also known as guillemets, «») and long dashes (–), used for direct speech and thoughts respectively, typically determine a sentence boundary when they appear with another demarcative punctuation mark (e.g., a full stop). In Manzoni's novel, if a closing quotation mark (guillemets or long dashes) appears with another punctuation mark, the latter is usually placed before the former, which formally closes the sentence. Lastly, in the novel, suspension points (...) can indicate a sentence boundary when they suggest a suspensive allusion or when they mark the interruption of a character's line due to linguistic or extra-linguistic contingencies. In such cases, suspension points' demarcative function is shown either by the following capital letter or by an opening quotation mark which indicates the beginning of a different character's line.

6. Evaluation

Table 2 reports the results of our evaluation in terms of F1. The best performance overall (0.98, SD: 0.005) is achieved by `Sonnet 4.6`. `GPT 5.2` shows good results (0.78, SD: 0.11), ranking second but at a considerable distance from `Sonnet 4.6`. For both LLMs, we report the standard deviation (SD) computed over two independent runs. The very low SD observed for `Sonnet 4.6` indicates highly stable performance across runs, whereas `GPT 5.2` shows greater variability, as reflected by its higher SD. Among the non-LLM approaches, the best performance (0.94) is registered with `SentenceSplitter`, a rule-based system. All other traditional tools do not exceed 0.69, thus showing substantially lower performance. The lowest score (0.51) is obtained by the unsupervised `WtP wtp-canine-s-12l` system. While the rule-based `SentenceSplitter` confirms that carefully designed rules can be effective even without adaptation, the results of the LLMs, especially `Sonnet 4.6`, highlight a substantial performance gap in favor of LLMs on this test set, together with different levels of robustness as measured by the standard deviation across runs.

Analyzing the outputs of the various systems, it is possible to notice some recurring errors (few examples are reported in Table 3):

1. Misinterpretation of guillemets («,»). The closing sign of the low quotation marks is not recognized as a sentence boundary, so in the automatic segmentation it can appear at the beginning or in the middle of a sentence.
2. In the `VIT` treebank and in the data used to train the combined `Stanza` model, sentences are segmented inconsistently: sometimes semicolons and colons are strong punctuation, and sometimes not. This causes inconsistent treatment of such punctuation marks.
3. Suspension points are often considered strong punctuation marks and the sentence is split after them.
4. A sentence is often split after an expressive punctuation mark (?, !) even if it is followed by a lowercase letter.

Table 2
Results (in terms of F1) of eight systems developed with different approaches: rule-based (RB), supervised (S), semi-supervised (SS) and unsupervised learning (U).

TYPE	SYSTEM	F1
RB	`spaCy sentencizer`	0.61
	`CoreNLP 4.5.7 ssplit`	0.66
	`SentenceSplitter`	0.94
S	`UDPipe 2 VIT model`	0.66
	`UDPipe 2 Old-Italian model`	0.67
	`Stanza combined`	0.69
SS	`Ersatz`	0.60
U	`Punkt`	0.68
	`WtP wtp-canine-s-12l`	0.51
LLMs	`GPT 5.2`	0.78 (SD: 0.11)
	`Sonnet 4.6`	**0.98** (SD: 0.005)

Table 3
Examples of errors in two of the tested systems compared with the manually split sentences.

TEST GOLD	**UDPipe 2 -VIT model**
1) «Al sagrestano gli crede?» 2) «Perché?»	1) » «Al sagrestano gli crede?» «Perché?»
1) – È lei, di certo!– 2) Era proprio lei, con la buona vedova.	1) – È lei, di certo!– Era proprio lei, con la buona vedova.
1) Anche Agnese, veda; anche Agnese…» 2) «Uh! ha voglia di scherzare, lei,» disse questa.	1) Anche Agnese, veda; anche Agnese…» «Uh! ha voglia di scherzare, lei,» disse questa.

TEST GOLD	**Ersatz**
1) «Al sagrestano gli crede?» 2) «Perché?»	1) » «Al sagrestano gli crede? 2) » «Perché?
1) – È lei, di certo!– 2) Era proprio lei, con la buona vedova.	1) – È lei, di certo! 2) – Era proprio lei, con la buona vedova.
1) Anche Agnese, veda; anche Agnese…» 2) «Uh! ha voglia di scherzare, lei,» disse questa.	1) Anche Agnese, veda; anche Agnese… » «Uh! 2) ha voglia di scherzare, lei,» disse questa. «

5. The long dash is not recognized as a sentence-ending marker; consequently, either the sentence continues after the dash or the dash appears at the beginning of the following sentence.

Although the performance of `Sentence Splitter` is generally excellent (F1=0.94), two types of errors can still be observed. Unlike other systems, suspension points are not consistently treated as sentence boundaries, and sentences are not split even when they are followed by a capital letter. For example, in *del resto, vedete, fin che c'è fiato… Guardatemi me* (EN: *after all, you see, as long as there's breath… Look at me*), the clause *Guardatemi me* (EN: *Look at me*) should be considered as a separate sentence but instead the system does not perform segmentation. Additionally, long dashes also pose

difficulties; for instance, the following three distinct sentences are erroneously merged into a single one:

1. *Una sera, Agnese sente fermarsi un legno all'uscio.* (EN: *One evening, Agnese hears a carriage stop at the door.*)
2. *- È lei, di certo! -* (EN: *- It's her, for sure! -*)
3. *Era proprio lei, con la buona vedova.* (EN: It was her, with the good widow.)

The same types of errors are found in the output of `Sonnet 4.6`. More specifically, there are 2 errors due to the presence of a dash followed by an upper-case letter where no split is applied and 3 errors related to suspension points where the model tends to under-segment, failing to insert a boundary after the ellipsis even when followed by an upper-case letter.

Another interesting case is that of the `UDPipe Italian-Old` model, trained on another cornerstone of Italian literature: "Divina Commedia" by Dante Alighieri. Although the textual genre is similar, the work is a poetic text from several centuries earlier, and the model's performance is comparatively low. Notably, this model also exhibits errors in handling suspension points and guillemets. The former, for instance, are consistently treated as sentence-ending markers, resulting in incorrect segmentation. For example, the sentence *Per domenica ne ho già… uno… due… tre; senza contarvi voi altri: e ne può capitare ancora.* (EN: *For Sunday I already have… one… two… three; not counting you: and more could happen.*) is split as follows:

1. *Per domenica ne ho già…*
2. *uno…*
3. *due…*
4. *tre; senza contarvi voi altri: e ne può capitare ancora.*

Guillemets («») appear in the Italian-Old treebank because they are used in the edition from which the text was sourced (Alighieri 1994), but they never occur at the end of a sentence. This is because strong punctuation is placed after the quotation marks, unlike in Manzoni, where it precedes them. An example from the Divine Comedy illustrates this: *E tutto in dubbio dissi: «Ov' è Beatrice?».* (EN: *All lost in doubt I said: «Where is Beatrice?».*) This results in systematic segmentation errors, as shown in the following case, where the closing quotation mark is incorrectly put at the beginning of the second sentence because sentence segmentation occurs after the period:

1. *E anche se io stessi zitto, già non servirebbe a nulla, perché parlan tutti; e vox populi, vox Dei.* (EN: *And even if I were silent, it would be of no use, because everyone is talking; and vox populi, vox Dei.*)
2. *» Trovarono appunto le tre donne e Renzo.* (EN: *» They found the three women and Renzo.*)

7. A New Stanza Model: Training and Evaluation

With the rest of the manually split data, namely 2,447 sentences for the training set and 324 for the development set, a new `Stanza` model specific for Manzoni's text ("I Promessi Sposi" = PS) was trained. Two different amounts of sentences were used as training in order to control the effect of the dataset size on the performance. The results obtained with 1500 steps are the following:

- PS_300 (300 training sentences): 0.97 F1
- PS_2447 (2,447 training sentences): 0.99 F1

With just 300 sentences there is already a clear improvement over the default model, obtaining an even higher result than the one obtained with `Sentence Splitter`, the system that had proven to be the best on our test set (see Table 2).

An analysis of the outputs produced by the retrained models reveals two significant differences compared to the original `Stanza Combined` model. The latter tends to insert an end-of-sentence marker between periods, question marks and exclamation points, and the closing guillemet (»). For instance, the sentence *«Speriamo,» concluse, «che il Signore gli avrà usato misericordia.»* (EN: *«Let us hope,» he concluded, «that the Lord will have had mercy on him.»*) is split into two separate sentences: the first ending with the period, and the second consisting solely of the closing quotation mark:

1. *«Speriamo,» concluse, «che il Signore gli avrà usato misericordia.*
2. *»*

This type of segmentation error occurs 33 times out of 53 with the original `Stanza Combined` model but is reduced to 20 instances with `PS_300` and is entirely absent with `PS_2447`. A second recurring error (25 occurrences) with `Stanza Combined` is the insertion of a line break after periods, question marks and exclamation marks even when these are followed by a lowercase letter. For example, the sentence *Ma che? di qualunque cosa si parlasse, il colloquio gli riusciva sempre delizioso.* (EN: *But what? Whatever they talked about, the conversation was always delightful for him.*) is segmented as follows:

1. *Ma che?*
2. *di qualunque cosa si parlasse, il colloquio gli riusciva sempre delizioso.*

This error does not occur at all with either `PS_300` or `PS_2447`.

The results obtained by applying the `PS_2447` model to the two previous versions of the novel, "Fermo e Lucia" (1823) and the so-called "Ventisettana" (1827), specifically to the first chapter (147 and 193 sentences, respectively), reflect the punctuation conventions characteristic of these versions. While the decrease in performance for the "Ventisettana" is minimal (F1 = 0.98, -0.01), a much more significant drop is observed for "Fermo e Lucia" (F1 = 0.68, -0.31). In the latter case, the model makes two main errors. First, it incorrectly splits the sentence after *D.* (the abbreviation for *don*, as in *D. Abbondio* and *D. Rodrigo*) in 3 out of 7 occurrences, as in the following example:

1. *Ma il povero D.* (EN: *But poor D.*)
2. *Abbondio non avrebbe voluto esser conscio…* (EN: *Abbondio would not have wanted to be aware…*)

Second, the model interprets the dash solely as a sentence-final punctuation mark, whereas in this text it actually marks the beginning and not the end of direct speech. For example, the following exchange between two characters:

1. *- Zitto zitto, a che serve tutto questo?* (EN: *- Hush hush, what's the point of all this?*)
2. *- Ma come farà Signor padrone?* (EN: *- But how will you do it, Master?*)

is split as follows:

1. *Zitto zitto, a che serve tutto questo? -*
2. *Ma come farà Signor padrone? -*

The retrained model appears to generalize well to other types of texts. Indeed, when applied to the test set of the VIT treebank, it achieves an F1 score of 0.91. Although this is slightly lower than the scores obtained by the corresponding in-domain VIT models (0.96 with Stanza and 0.95 with UDPipe 2), the performance drop is minimal.

Remarkably, the model performs even better on the test set of the Italian-Old treebank, achieving an F1 score of 0.98, the same score reported for the in-domain model trained specifically on the "Divina Commedia".

8. Other 19th-Century Novels

Table 4 displays the performance of the same systems tested on "I Promessi Sposi" on the first approximately 90 sentences of three other important 19th-century novels: "I Malavoglia" (1881) by Giovanni Verga (Verga and Cecco 2014), "Le avventure di Pinocchio. Storia di un burattino" (1883) by Carlo Collodi (Collodi and Castellani Pollidori 1983), "Cuore" (1886) by Edmondo de Amicis (De Amicis and Tamburini 2018 (1° ed. 1972)).[20] As with "I Promessi Sposi", the analysis of these three novels also relies on the standard critical editions, which follow different editorial criteria. In brief, the editions of "I Malavoglia" and "Cuore" are largely conservative, remaining close to the author's first edition, while the edition of "Pinocchio" adopts a more interventionist approach, introducing changes based either on interpretive considerations or on the recovery of elements from the textual tradition.

Table 4
Results on about 90 sentences taken from other 19th-century novels. `Stanza-PS_2447` refers to the model retrained on Manzoni's novel, as described in Section 7. Standard Deviations over two runs on LLMs are reported in parentheses.

	Malavoglia	Pinocchio	Cuore
`spaCy sentencizer`	0.73	0.35	**0.84**
`CoreNLP 4.5.7 ssplit`	0.76	0.72	0.62
`SentenceSplitter`	0.77	0.45	0.68
`UDPipe 2 VIT model`	0.75	0.79	0.67
`UDPipe 2 Italian-Old model`	0.85	0.65	0.69
`Stanza combined`	0.71	0.70	0.61
`Stanza PS_2447`	**0.90**	0.89	0.69
`Ersatz`	0.72	0.75	0.66
`Punkt`	0.73	0.77	0.66
`WtP wtp-canine-s-12l`	0.53	0.78	0.39
`GPT 5.2`	0.82 (0.11)	0.86 (0.12)	0.79 (0.13)
`Sonnet 4.6`	0.87 (0.002)	**0.93** (0.01)	0.66 (0.001)

As in the previous experiment, the results are overall lower than those typically reported for contemporary Italian texts.

Among the traditional tools, the model retrained on "I Promessi Sposi" (`Stanza PS_2447`) achieves the best performance on "I Malavoglia" (0.90) and shows a very strong result on "Le avventure di Pinocchio" (0.89). Compared to the default `Stanza combined` model, this corresponds to an improvement of +19 points on "I Malavoglia" and +19 points on "Le avventure di Pinocchio". The improvement is more limited on "Cuore" (+8 points). The rule-based approach remains promising but with different sys-

20 86 sentences are taken from "I Malavoglia", corresponding to the first chapter of the novel; 93 sentences, that is the first two chapters, come from "Le avventure di Pinocchio"; 87 sentences are taken from "Cuore", corresponding to the first three chapters of the novel.

tems depending on the text: `spaCy sentencizer` achieves the best non-LLM result on "Cuore" (0.84), while `SentenceSplitter` performs well on "I Malavoglia" (0.77). In contrast, the supervised `UDPipe 2 VIT model` is the best traditional system on "Le avventure di Pinocchio" (0.79).

When considering LLMs, `Sonnet 4.6` achieves the best overall result on "Le avventure di Pinocchio" (0.93, SD: 0.01), outperforming all other systems, including `Stanza PS_2447`. It also performs strongly on "I Malavoglia" (0.87, SD: 0.002), though slightly below `Stanza PS_2447` (0.90). On "Cuore", however, its performance drops to 0.66 (SD: 0.001). `GPT 5.2` shows more balanced results across texts, but with higher variability across the two runs, as indicated by the larger standard deviations. In contrast, `Sonnet 4.6` exhibits very low standard deviation values, indicating highly stable behavior across runs.

Finally, some tools obtain extremely different results depending on the text they process. `spaCy` and `SentenceSplitter` record very low scores on "Le avventure di Pinocchio" (0.35 and 0.45 respectively), while `WtP` achieves only 0.39 on "Cuore", nearly half of its performance on "Le avventure di Pinocchio" (0.78). These discrepancies confirm that sentence segmentation performance is strongly influenced by the stylistic and structural characteristics of each novel.

Indeed, among the main peculiarities of the novel is the original and personal use of quotation marks. For example, guillemets («,») are frequently used to refer to popular sayings and proverbs as well as to short formulas (Bronzini 1982), which sometimes intersperse the diegesis, occasionally preceded by a dash and introduced by colons or not (e.g., *La casa del nespolo era piena di gente; e il proverbio dice: «triste quella casa dove ci è la visita pel marito!»*; EN: *The house of the medlar tree was full of people; as the saying goes: «unlucky is the house that gets visitors because of the husband!»*), and sometimes isolate a complete enunciative section (e.g., *«Chi fa credenza senza pegno, perde l'amico la roba e l'ingegno».*; EN: *«Trusting without security means losing your friend, your belongings, and your sense».)*. The long dash (–)[21], instead, has a number of different functions (Tonani 2010a): one of these is to signal direct speech, but often marking only its beginning and not its end. This leads, on one hand, to a variety of ways of handling parenthetical elements and, on the other hand, to a blurred boundary between the characters' speech, the characters' speech mediated by the narrator, and the narrator's own discourse.

"Pinocchio", a novel written for a young audience, is characterized by a strongly dialogic style (Pellerey 2005). For direct speech, including the simulated dialogue between the narrator and the reader, the long dash (–) is abundantly used, but as for "I Malavoglia", the opening dashes are not always accompanied by the closing ones. However, in the original layout of Collodi's novel, the distinction between one line and the next is clear, marked by a new line, and the closing dash is used, in most cases, after the final sentence in a dialogic sequence, just before the diegesis resumes. Additionally, Collodi frequently uses punctuation clusters, specifically the exclamation mark followed by suspension points (!...), at the end of sentences (Castellani Pollidori 1983), a possibility mostly not contemplated by late 19th-century grammars. Following such clusters, both uppercase (e.g., *No, è meglio cuocerlo nel piatto!... O non sarebbe più saporito se lo friggessi in padella?*; EN: *No, it's better to cook it in the pot!... Or wouldn't it be tastier if I fried it in a pan?*) and lowercase letters (e.g., – *Pinocchio!... rendimi subito la mia*

21 In this study, the standard Italian dash (–) was used to ensure consistency with the contemporary Italian conventions. However, in the original 19th-century texts included in the corpus, the typographically prevalent form was the so-called English dash, which is longer. This form was instead preserved in all the critical editions consulted.

parrucca! –; EN: – *Pinocchio!... give me my wig back right now!* –) may appear. Guillemets («,») are rarely used, in order to indicate a reported speech within another reported speech (e.g., – *Non lo so, babbo, ma credetelo che è stata una nottata d'inferno e me ne ricorderò fin che campo. Tonava, balenava e io avevo una gran fame, e allora il Grillo-parlante mi disse: «Ti sta bene: sei stato cattivo, e te lo meriti»*; EN: – *I don't know, father, but believe me, it was a night from hell, and I'll remember it as long as I live. There was thunder and lightning, and I was terribly hungry, and then the Talking Cricket said to me: «It serves you right: you've been bad, and you deserve it»)*.

Lastly, Edmondo de Amicis's novel "Cuore" tells the story of a child's school experience from his point of view, adopting a diary-like structure. Each chapter is introduced by a title referring to the theme that will be addressed and a chronological indication. "Cuore" is generally regarded as a novel with a rather simple linguistic form: the prose is plain, and the sentences are mostly short. However, at the punctuation level, the text presents substantial complexities. First of all, guillemets are not used at all; direct speech is indicated by long dashes (–), but successive lines of dialogue are arranged consecutively on the page, and in such cases, the closing dash of the previous line also serves as the opening dash of the next line. Secondly, this system is fairly regular but not entirely consistent: some lines of dialogue may be opened but not closed by a dash even when isolated, while others appear without any punctuation marking at all. The latter is the case with reported speech embedded within other reported speech, which, though lacking any actual introductory marks, is typically preceded by a colon and begins with a capital letter (e.g., – *È quest'operaio,* – *rispose il maestro,* – *che è venuto a lagnarsi perchè il suo figliuolo Carlo disse al suo ragazzo: Tuo padre è uno straccione.*; EN: – *It's this worker* – *replied the teacher* – *who came to complain because his son Carlo said to his boy: Your father is a beggar.*). Since the lines of dialogue are integrated into the narrative structure, they can end with various punctuation marks, from commas to semicolons to full stops: when the punctuation mark is not strong, after the preliminary conclusion of the line, the text continues with the narrator's discourse. The dash is also frequently used to introduce or interrupt the narrator's discourse, functioning as a parenthetical marker. These insertions vary in length – from single words to full clauses – and often combine with other punctuation marks (e.g., *È anche un tipo curioso il mio vicino di sinistra,* – *Stardi,* – *piccolo e tozzo, senza collo, un grugnone che non parla con nessuno, e pare che capisca poco, ma sta attento al maestro senza batter palpebra, con la fronte corrugata e coi denti stretti: e se lo interrogano quando il maestro parla, la prima e la seconda volta non risponde, la terza volta tira un calcio.*; EN: *My neighbor on the left is a curious type too,* – *Stardi,* – *short and stocky, with no neck, a grumbler who doesn't talk to anyone and seems not to understand much, but he watches the teacher without blinking, with a furrowed brow and clenched teeth: and if he gets called on while the teacher is talking, he doesn't answer the first or second time, the third time he kicks.*); in such cases, the second dash may also be absorbed by a following stronger punctuation mark, such as a full stop. Finally, in some narrative transitions, the dash appears together with a full stop, serving as an additional boundary marker.

Beyond the specific differences listed schematically above, there are also some common typographical and punctuation features among the considered novels. For example, when a closing quotation mark appears with another punctuation mark, the latter in general occurs before the former (as found in "I Promessi Sposi"), with the only exception of the guillemets in "I Malavoglia".

Figure 1 presents the percentage of times punctuation marks are used as sentence-final markers in the test sets of the novels considered in our study. These percentages are calculated relative to the total number of occurrences of each punctuation mark in each text. The analysis reveals both consistent patterns and evident divergences across the

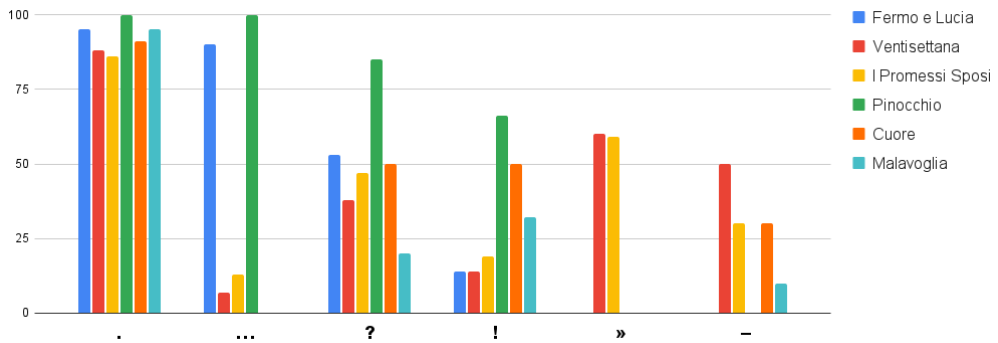

Figure 1
Percentage of times each punctuation mark is used as a sentence-end marker in the test sets with respect to their total number of occurrence.

texts. The period (.) consistently emerges as the dominant sentence-ending punctuation mark. Its usage ranges from 86% in "I Promessi Sposi" to 100% in "Pinocchio". More variation emerges with suspension points (...), which functions as a sentence ender predominantly in "Pinocchio" (100%) and "Fermo e Lucia" (90%), and to a much lesser extent in "Ventisettana" (7%) and "I Promessi Sposi" (13%). In the other two texts, "Cuore" and "Malavoglia", this mark is not used at all. The question mark (?) appears across all novels, but its frequency of sentence-final use varies widely. It is most prominent in "Pinocchio" (85%), while "Malavoglia" exhibits the lowest rate of use (20%). The percentage across the three versions of Manzoni's novel is between 38% (in the "Ventisettana") and 53% in "Fermo e Lucia". Similarly, the exclamation mark (!) shows significant differences in usage. It is most frequently used to end sentences in "Pinocchio" (66%) and "Cuore" (50%), while its use in "Fermo e Lucia", "Ventisettana", and "I Promessi Sposi" is markedly lower (less than 20%). A distinct pattern emerges with the closing guillemets (»), which appears as sentence boundaries exclusively in "Ventisettana" (60%) and "I Promessi Sposi" (59%). Finally, the dash (-) appears variably across the texts, with significant usage in "Ventisettana" (50%), "I Promessi Sposi" (30%), and "Cuore" (30%), while it is minimally present in "Malavoglia" (10%) and it is never used as a sentence boundary in "Fermo e Lucia" and "Pinocchio".

8.1 Experiments

Given the good results obtained with only 300 sentences in the case of "I Promessi Sposi" (see Section 7), we repeated the retraining experiment with the other novels. Specifically, we created training sets of 300 sentences and development sets of 100 sentences for each novel and retrained the `Stanza`'s sentence splitter using 1500 steps.

Table 5 provides a comparative analysis of F1 scores achieved by the models on the different test sets. Each row corresponds to a model trained on a specific novel, while the columns represent the test texts. `Stanza-Novels` is instead trained on the combination of all training sets, including that of "I Promessi Sposi". Performance varies between texts, but results are generally better than those obtained with other tools (see Table 4). Models tend to perform best on their corresponding training texts (e.g., `Stanza-Pinocchio` achieves an F1 of 0.98 on "Pinocchio"; `Stanza-PS_2447`

Table 5
F1 score obtained on different models across the various novels.

	Malavoglia	Pinocchio	Cuore	Promessi Sposi
`Stanza-Cuore`	0.86	0.85	**0.86**	0.98
`Stanza-Malavoglia`	0.87	0.96	0.75	0.93
`Stanza-Pinocchio`	0.72	**0.98**	0.78	0.93
`Stanza-PS_2447`	**0.90**	0.89	0.69	**0.99**
`Stanza-Novels`	0.89	0.95	0.75	**0.99**

reaches 0.99 on "I Promessi Sposi"), confirming that in-domain training leads to stronger sentence boundary detection. Conversely, cross-domain generalization is not always reliable: `Stanza-Pinocchio` drops to 0.72 on "Malavoglia", while `Stanza-PS_2447` scores only 0.69 on "Cuore". Interestingly, `Stanza-Malavoglia` is surpassed on "Malavoglia" itself by `Stanza-PS_2447` (0.87 vs. 0.90). This suggests that some sentence segmentation patterns learned from "I Promessi Sposi" may transfer effectively to other 19th-century novels, possibly due to the comparatively more regular punctuation and syntactic patterns of Manzoni's text. The `Stanza-Novels` model achieves top-tier results on "I Promessi Sposi" (0.99) and "Pinocchio" (0.95) and a good F1 on "I Malavoglia". On the contrary, "Cuore" poses challenges even for the combined model, confirming its status as a problematic text for sentence segmentation. As previously noted, this novel exhibits low regularity in the use of punctuation marks. In particular, there are inconsistencies in the function assigned to the dash, which is associated with the majority of splitting errors. For instance, as illustrated by the two examples below, the dash may appear either at both the beginning and end of a direct speech segment, or solely at its beginning.

1. *Egli aperse gli occhi, e disse:* – *La mia cartella!* – (EN: *He opened his eyes and said:* – *My schoolbag!* –)
2. *Mio padre domandò a un maestro:* – *Cos'è stato?* (EN: *My father asked a teacher:* – *What happened?* –)

9. Conclusions

This study has examined the effectiveness of diverse sentence splitting tools, including rule-based, supervised, semi-supervised, unsupervised approaches, and two LLMs, on a dataset of four canonical 19th-century Italian novels: "I Promessi Sposi", "I Malavoglia", "Le avventure di Pinocchio", and "Cuore". While these texts were all composed within a relatively narrow historical window, they differ significantly in narrative structure, authorial style, and punctuation practices, reflecting, alongside the personal choices of each author, the transitional state of Italian punctuation conventions during the late 19th century.

Our findings confirm that sentence splitting, often considered a solved task in NLP for standardized domains, remains a non-trivial challenge in historical and literary contexts. Among traditional tools, performance varies substantially depending on both the methodological approach and the specific novel. Rule-based systems can achieve very high accuracy in some settings, while supervised models retrained on in-domain data show substantial improvements over default models, particularly on "I Malavoglia"

and "Le avventure di Pinocchio". However, models trained on modern, well-structured data generally perform poorly on these novels, underscoring the limitations of existing tools when applied to text types with non-standard punctuation or genre-specific conventions. Notably, "Cuore" emerged as the most problematic of the four novels, due to its irregular and often ambiguous use of punctuation, especially the dash, frequently employed in ways that complicate sentence boundary identification.

The inclusion of LLMs provides further insight. `Sonnet 4.6` achieves the best overall result in the global evaluation, clearly outperforming all traditional systems. At the per-novel level, it obtains the highest score on "Le avventure di Pinocchio" and competitive results on "I Malavoglia", though it does not consistently surpass retrained supervised models. `GPT 5.2` shows more balanced but overall lower performance, with greater variability across runs. Interestingly, even high-performing LLMs exhibit a drop in accuracy on "Cuore", confirming that stylistic and punctuation-related idiosyncrasies remain challenging even for these models. Results suggest that, while LLMs can substantially improve sentence segmentation in historical literary texts, they are not immune to genre-specific complexity.

Importantly, this work demonstrates the value of integrating philological and linguistic expertise into NLP. In literary texts, where punctuation is a meaningful part of the author's stylistic repertoire, collaboration with scholars in Italian linguistics and the Digital Humanities is essential for developing tools that respect both the form and function of the original writing. Sentence splitting, therefore, emerges not only as a pre-processing step for downstream applications, but also as a site of interdisciplinary inquiry with the potential to inform the design of historically aware and genre-sensitive NLP models. By releasing annotated data and models, we hope to encourage further research on sentence segmentation in literary texts and promote cross-disciplinary efforts that bridge computational and humanistic approaches to language. This area, particularly in the context of Italian literature, holds rich opportunities for future work.

Acknowledgments

Questa pubblicazione è stata realizzata da ricercatrice con contratto di ricerca cofinanziato dall'Unione europea - PON Ricerca e Innovazione 2014-2020 ai sensi dell'art. 24, comma 3, lett. a, della Legge 30 dicembre 2010, n. 240 e s.m.i. e del D.M. 10 agosto 2021 n. 1062.

References

Alighieri, Dante. 1994. *La Commedia secondo l'antica vulgata voll.* i–iv. Number 7 in Edizione nazionale delle Opere di Dante Alighieri a cura della Società Dantesca Italiana. Le Lettere, Florence, Italy. Editor: Giorgio Petrocchi.

Antonelli, Giuseppe. 2008. Dall'ottocento a oggi. In Bice Mortara Garavelli, editor, *Storia della punteggiatura in Europa*. Laterza, Roma, pages 178–210.

Au, Willy, Abderrahim Ait-Azzi, and Juyeon Kang. 2021. FinSBD-2021: The 3rd Shared Task on Structure Boundary Detection in Unstructured Text in the Financial Domain. In *Companion Proceedings of the Web Conference 2021*, WWW '21, pages 276—279, New York, NY, USA, 19-23 April 2021. Association for Computing Machinery.

Au, Willy, Bianca Chong, Abderrahim Ait Azzi, and Dialekti Valsamou-Stanislawski. 2020. FinSBD-2020: The 2nd shared task on sentence boundary detection in unstructured text in the financial domain. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors, *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 47–54, Kyoto, Japan, 5 January 2020.

Azzi, Abderrahim Ait, Houda Bouamor, and Sira Ferradans. 2019. The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors, *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 74–80, Macao, China, 12 August 2019.

Bonomi, Ilaria, Andrea Masini, Silvia Morgana, and Mario Piotti. 2010. *Elementi di linguistica italiana*, volume 103. Carocci, Roma.

Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper, editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, 8-9 August 2013. Association for Computational Linguistics.

Bronzini, Giovanni Battista. 1982. Proverbi, discorso e gesto proverbiale nei «Malavoglia». In *I Malavoglia. Atti del Congresso Internazionale di Studi*, pages 637–684, Catania, 26-28 November 1981. Biblioteca della Fondazione Verga.

Brugger, Tobias, Matthias Stürmer, and Joel Niklaus. 2023. Multilegalsbd: A multilingual legal sentence boundary detection dataset. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 42–51, New York, NY, USA, 19-23 June 2023. Association for Computing Machinery.

Castellani Pollidori, Ornella. 1983. Introduzione. In Carlo Collodi and Ornella Castellani Pollidori, editors, *Le avventure di Pinocchio*. Fondazione nazionale Carlo Collodi, Pescia, pages XIII–LXXXIV.

Collodi, Carlo and Ornella Castellani Pollidori. 1983. *Le avventure di Pinocchio*. Fondazione nazionale Carlo Collodi, Pescia.

Corbetta, Claudia, Marco Passarotti, Flavio Massimiliano Cecchini, and Giovanni Moretti. 2023. Highway to hell. towards a Universal Dependencies treebank for dante alighieri's comedy. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, and Nicole Novielli, editors, *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 154–161, Venice, Italy, 30 November - 2 December 2023. CEUR Workshop Proceedings.

Crystal, David. 2015. *Making a Point: The Persnickety Story of English Punctuation*. Profile Books, London.

De Amicis, Edmondo and Luciano Tamburini. 2018 (1° ed. 1972). *Cuore. Libro per ragazzi*. Einaudi, Torino.

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational linguistics*, 47(2):255–308.

Delmonte, Rodolfo, Antonella Bristot, and Sara Tonelli. 2007. VIT-Venice Italian Treebank: Syntactic and quantitative features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54, Bergen, Norway, 7-8 December 2007. Northern European Association for Language Technologies.

Dridan, Rebecca and Stephan Oepen. 2013. Document parsing: Towards realistic syntactic analysis. In Harry Bunt, Khalil Sima'an, and Liang Huang, editors, *Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013)*, pages 127–133, Nara, Japan, 27–29 November 2013. Association for Computational Linguistics.

Ferrari, Angela. 2018. Punteggiatura. In Giuseppe Antonelli, Matteo Motolese, and Lorenzo Tomasi, editors, *Storia dell'italiano scritto. Grammatiche*, volume IV. Carocci, Roma, pages 169–202.

Ferrari, Angela. 2020. Le virgole dei promessi sposi. In Angela Ferrari, Letizia Lala, Filippo Pecorari, and Roska Stojmenova Weber, editors, *Capitoli di storia della punteggiatura italiana*. Edizioni Dell'Orso, Alessandria, pages 3–17.

Ferrari, Angela, Letizia Lala, Fiammetta Longo, Filippo Pecorari, Benedetta Rosi, and Roska Stojmenova. 2018. *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*. Carocci, Roma.

Gherardini, Giovanni. 1843. *Lessigrafía italiana o sia maniera di scrivere le parole italiane proposta e messa a confronto con quella insegnata dal Vocabolario della Crusca*. Bianchi di Giacomo, Milano.

Griffis, Denis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M. Lai. 2016. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. In *Proceedings of the AMIA 2016 Joint Summits on Translational Science*, pages 88–97, San Francisco, CA, USA, March. American Medical Informatics Association.

Kiss, Tibor and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, 13-15 September 2005.

Liu, Yang, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 451–458, Ann Arbor, Michigan, USA, 25-30 June 2005. Association for Computational Linguistics.

Liu, Yang and Shasha Xie. 2008. Impact of automatic sentence segmentation on meeting summarization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, page 1262, Las Vegas, Nevada, USA, 31 March - 4 April 2008. IEEE.

Lorenceau, Annette. 1980. La ponctuation au xix^{e} siècle. *Langue française*, (45). La ponctuation.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Kalina Bontcheva and Jingbo Zhu, editors, *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA, 23-24 June 2014. Association for Computational Linguistics.

Manzoni, Alessandro and Barbara Colli. 2024. *I Promessi Sposi. Edizione genetica della Quarantana*. Casa del Manzoni, Milano.

Manzoni, Alessandro, Fausto Ghisalberti, and Alberto Chiari. 1954. L'ultima revisione dei Promessi Sposi. In *Tutte le opere di Alessandro Manzoni. I Promessi Sposi*, volume II. Mondadori, Milano, pages 789–989.

Minixhofer, Benjamin, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada, 9-14 July 2023. Association for Computational Linguistics.

Mortara Garavelli, Bice. 2003. *Prontuario di punteggiatura*. Laterza, Bari.

Mortara Garavelli, Bice, editor. 2008. *Storia della punteggiatura in Europa*. Laterza, Roma.

Nunberg, Geoffrey. 1990. *The Linguistics of Punctuation*. Number 18 in CSLI Lecture Notes. Center for the Study of Language and Information, Stanford, CA.

Palmer, David D. 2000. Chapter 2: Tokenisation and sentence segmentation. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of natural language processing*. Marcel Dekker, New York, chapter 2.

Palmero Aprosio, Alessio and Giovanni Moretti. 2018. Tint 2.0: an all-inclusive suite for NLP in Italian. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 312–318, Turin, Italy, 10-12 December 2018. CEUR Workshop Proceedings.

Parkes, Malcolm B. 1992. *Pause and Effect: An Introduction to the History of Punctuation in the West*. University of California Press, Berkeley and Los Angeles.

Pellerey, Roberto. 2005. Pinocchio tra dialogo e scrittura. *Belfagor*, 60:267–284.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, 5-10 July 2020. Association for Computational Linguistics.

Read, Jonathon, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In Martin Kay and Christian Boitet, editors, *Proceedings of the 24th International Conference on Computational Linguistics: Posters*, pages 985–994, Mumbai, India, 8-15 December 2012. The COLING 2012 Organizing Committee.

Rudrapal, Dwijen, Anupam Jamatia, Kunal Chakma, Amitava Das, and Björn Gambäck. 2015. Sentence boundary detection for social media text. In Dipti Misra Sharma, Rajeev Sangal, and Elizabeth Sherly, editors, *Proceedings of the 12th International Conference on Natural Language Processing*, pages 254–260, Trivandrum, India, 11-14 December 2015. NLP Association of India.

Sanguinetti, Manuela, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 7-12 May 2018. European Language Resources Association (ELRA).

Sheik, Reshma, Gokul T. Adethya, and S. Jaya Nirmala. 2022. Efficient deep learning-based sentence boundary detection in legal text. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro, editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 208–217, Abu Dhabi, United Arab Emirates (Hybrid), 8 December 2022. Association for Computational Linguistics.

Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In Daniel Zeman and Jan Hajič, editors, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, 31 October - 1 November 2018. Association for Computational Linguistics.

Tonani, Elisa. 2010a. Il 'bianco di dialogato' e il trattamento tipografico del discorso diretto. In Elisa Tonani, editor, *Il romanzo in bianco e nero. Ricerche sull'uso degli spazi bianchi e dell'interpunzione nella narrativa italiana dall'Ottocento a oggi*. Franco Cesati, Firenze, pages 103–136.

Tonani, Elisa. 2010b. Premessa. Tra punteggiatura e tipografia. In Elisa Tonani, editor, *Il romanzo in bianco e nero. Ricerche sull'uso degli spazi bianchi e dell'interpunzione nella narrativa italiana dall'Ottocento a oggi*. Franco Cesati, Firenze, pages 13–28.

Verga, Giovanni and Ferruccio Cecco. 2014. *I Malavoglia*. Fondazione Verga-Interlinea, Catania-Novara.

Wicks, Rachel and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online, 1-6 August 2021. Association for Computational Linguistics.

Wicks, Rachel and Matt Post. 2022. Does sentence segmentation matter for machine translation? In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 843–854, Abu Dhabi, United Arab Emirates (Hybrid), 7-8 December 2022. Association for Computational Linguistics.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In Daniel Zeman and Jan Hajič, editors, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, 31 October – 1 November 2018. Association for Computational Linguistics.