Is Sentence Splitting a Solved Task? Experiments to the Intersection Between NLP and Italian Linguistics

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

[1]Università di Parma, Via D'Azeglio, 85, 43125 Parma, Italy

Abstract

Sentence splitting, that is the segmentation of the raw input text into sentences, is a fundamental step in text processing. Although it is considered a solved task for texts such as news articles and Wikipedia pages, the performance of systems can vary greatly depending on the text genre. This paper presents the evaluation of the performance of eight sentence splitting tools adopting different approaches (rule-based, supervised, semi-supervised, and unsupervised learning) on Italian 19th-century novels, a genre that has not received sufficient attention so far but which can be an interesting common ground between Natural Language Processing and Digital Humanities.

Keywords

sentence splitting, text segmentation, literary texts, Italian

1. Introduction

Sentence splitting is the process of segmenting a text into sentences[1] by detecting their boundaries, which, at least for Western languages, including Italian, usually correspond to certain punctuation marks [2]. This means that sentence splitting, for many languages, is a matter of punctuation disambiguation, that is, recognizing when a punctuation mark signals a sentence boundary or not. The importance of sentence splitting is often underestimated because it is considered an easy task, but its quality has a strong impact on the quality of subsequent text processing because errors can propagate reducing the performance of downstream tasks such as Syntactic Analysis [3], Machine Translation [4] and Automatic Summarization [5].

The most popular pipeline models, such as those of Stanza [6] and spaCy[2], have mostly been trained and evaluated on fairly formal texts, such as news articles and Wikipedia pages, so the publicly reported performances tend to be high, i.e. above 0.90 in terms of F1. However, the text genre has a significant impact on the results. For example, in the CoNLL 2018 shared task "Multilingual Parsing from Raw Text to Universal Dependencies", the best system on the Italian ISDT treebank [7] achieved a F1 of 0.99, while on the PoSTWITA treebank, made of tweets [8], the highest result was 0.66.

Given these variations, considering less formal text genres could provide valuable insights into the challenges of sentence splitting. Among these genres are literary texts, which present unique and peculiar stylistic and creative features that can break traditional grammatical norms, including punctuation ones [9]. These features depend on both authorial choices and the cultural context of the time. As a matter of facts, punctuation can vary significantly depending on the historical period; literary texts may follow prevailing trends or oppose them, giving rise to new trends. This phenomenon is particularly evident in 19th century, when the Italian *usus punctandi* began shifting from a primarily syntactic usage, prescribed by grammar books, to a communicative-textual usage of punctuation marks [10]. Since this shift was probably influenced by the reflections and the practical uses of prominent authors such as Alessandro Manzoni [11], our study focuses on his historical novel, "I Promessi Sposi". The author paid meticulous attention to the punctuation of the text, revising it up to the final print proofs, and made specific and personal choices in collaboration with the publisher, alongside more classical ones [12]. Although not always consistent, Manzoni's decisions make the novel particularly complex and interesting from a punctuation perspective. Furthermore, "I Promessi Sposi"

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, 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, 6; Arianna Redaelli is responsible for Sections 1, 4, 8. Section 7 were collaboratively written by the two authors.

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

[1] By "sentence" we mean a coherent set of words constructed according to the general rules of the language, conveying a complete thought that makes sense on its own [1]. A sentence ends with a strong punctuation mark (e.g., full stop, question mark, or exclamation point) and is typically followed by a capital letter. The definition of sentence adopted here, which like any definition is inherently problematic, is motivated by the specific requirements of the present work, as will be seen below.

[2] https://spacy.io

has been a fundamental reference for the development of a common written Italian language: starting from this assumption, many of the author's punctuation choices have been adopted by later grammars for rule-making, though only some of them have become part of the standard. Given that punctuation was still undergoing standardization at the time, and that its use can depend not only on the conventions of the period but also on the writer's personal style, the type of content being addressed (and how it is presented), and even the influence of typography during the printing process, we also decided to broaden our study to include sections from other novels contemporary to Manzoni's (1840-42). Specifically, we analyzed "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.

In this paper, our main contributions are as follows: (i) we provide an estimate of the performance of eight sentence splitting tools adopting different approaches on a specific and challenging text genre, namely historical literary fiction texts, which has not received enough attention so far; (ii) we compare the results considering the point of view of humanities scholars (in particular Italian linguistics) as the main stakeholders in the considered domain, in order to establish a flourishing cross-fertilization between NLP and Digital Humanities; (iii) we release manually split data for four 19th-century Italian novels and a shared notebook where to run many of the tested systems.[3]

2. Related Work

Sentence splitting systems can be categorized into three macro-classes based on the approach used to develop them. There are rule-based systems, such as `Sentence Splitter`[4] and the `Sentencizer` module of spaCy, that use heuristics specific to the various languages and lists of exceptions and abbreviations. Then, there are supervised systems that need datasets in which sentences are already correctly segmented to be trained. For example, `UDPipe` [13] and `Stanza` are trained on Universal Dependencies (UD) treebanks [14]. Finally, unsupervised systems are trained on datasets of non-segmented texts taking advantage of features such as the length of words and collocational information. An example is given by Punkt, available as a module within the NLTK (Natural Language Toolkit) library [15]. In our work, we test these various approaches on a benchmark dataset of historical literary fiction texts by evaluating the performance of eight different systems.

There are several studies that analyze the impact of text genre on sentence splitting, but literary texts are rarely considered. For example, Liu et al. [16] work on speech transcriptions, Sheik et al. [17] on legal texts, and Rudrapal et al. [18] on social media posts. Moreover, a shared task on sentence boundary detection in the financial domain (FinSBD) was organized in 2019, 2020 and 2021 [19].

Most of the available studies concern the processing of English texts while Italian is usually not included in the evaluation. An interesting exception is given by a work on multilingual legal texts that contains a detailed evaluation of the results on Italian documents [20].

Our work draws inspiration from the assessment on English texts provided by Read et al. [21] which includes, among others, the Sherlock Holmes stories, but moving to the Italian context. Furthermore, we focus on the literary context showing how 19th-century novels are a challenge for current sentence splitting systems.

3. 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 tools developed with different approaches. Some tools are modules integrated in larger pipelines, others are systems specifically created to perform only sentence splitting. 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 the colons and semicolons as punctuation marks capable of indicating the boundary of a sentence [22], as anticipated in footnote 1, 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 the colons, 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, yet the diegetic frame follow-

[3] https://github.com/RacheleSprugnoli/Sentence_Splitting_Manzoni

[4] https://github.com/mediacloud/sentence-splitter

ing 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.

Selected tools are the following:

- `CoreNLP`[5]: an NLP pipeline written in Java and developed by Stanford University [23]. 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`[6]: a Python module based on scripts developed for processing the Europarl corpus [24]. It supports several languages with ad-hoc rules.
- `UDPipe`[7]: 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 2.12), seven are for Italian. We evaluated the model trained on the VIT treebank [25] that does not (always) split at colons and semicolons.
- `Stanza`[8]: 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.
- `Ersatz`[9]: 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 [26].
- `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`[10]: 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 [27]. Among the various available models, we adopted the `wtp-canine-s-12l` which, according to the official documentation of the tool, have the best results on languages other than English.

For the evaluation, the tools were used as they are, using their default configurations, without making any customization. For this reason, given the choices motivated above, we did not consider other systems, such as Tint [28], which by default split at colons and semicolons.

4. Dataset

The data used to evaluate the aforementioned tools are taken from "I Promessi Sposi" in its final version published in 1840-1842[11]. 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.[12] 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.[13] Following this division, the test set is made of 324 sentences.

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. 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 role, modifying the sentence's internal intonation without determining its end. Low quotation marks («») 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,

[5] https://stanfordnlp.github.io/CoreNLP/
[6] https://github.com/mediacloud/sentence-splitter
[7] https://ufal.mff.cuni.cz/udpipe
[8] https://stanfordnlp.github.io/stanza/
[9] https://github.com/rewicks/ersatz
[10] https://github.com/segment-any-text/wtpsplit
[11] The text, fully digitized and available online, was collated with the reference edition [29] prior to analysis, to ensure maximum fidelity to the author's punctuation choices.
[12] https://universaldependencies.org/release_checklist.html#data-split
[13] The output of this process was used to train a new Stanza model as reported in Section 6.

Table 1
End-of-sentence markers in the test set.

MARK	# TOTAL	# EOS
.	277	237
»	90	53
?	47	22
!	31	6
…	23	3
–	10	3

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.

5. Results of the Evaluation

Table 2 reports the results of our evaluation in terms of F1. The best performance (0.94) is registered with `Sentence Splitter`, a rule-based system. All other tools do not exceed 0.70, thus having significantly lower performances than those reported on contemporary Italian texts. For example, the official result of `UDPipe 2` on the `VIT` treebank with the 2.12 model starting from a raw text is 0.95, that is almost 30 points more than what is obtained on our test set. The lowest result (0.51) is obtained by the unsupervised `WtP` system. Although the rule-based approach seems to be the most promising, only `Sentence Splitter` has an excellent result even without any adaptation of the existing rules.

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
	`Stanza combined`	0.69
SS	`Ersatz`	0.60
U	`Punkt`	0.68
	`WtP wtp-canine-s-12l`	0.51

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 supervised systems semicolons and colons are sometimes considered as sentence boundary signals. Indeed, in the `VIT` treebank and in those used to train the combined `Stanza` model, sentences are segmented inconsistently: sometimes semicolons and colons are strong punctuation, and sometimes not.
3. Suspension points are always considered strong punctuation marks and the sentence is splitted after them.
4. A sentence is often split after an expressive punctuation mark (?, !) even if it is followed by a lowercase letter.
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.

6. Training a New Stanza Model

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 was trained. 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:

- 300 sentences: 0.97 F1
- 1000 sentences: 0.98 F1
- 2,447 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.

7. What About Other 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:[14] "I Malavoglia" (1881) by Giovanni Verga [30], "Le avventure di Pinocchio. Storia di un burattino" (1883) by Carlo Collodi [31], "Cuore" (1886) by Edmondo de Amicis [32].[15]

[14] The reference edition text was used for the analysis of these novels too.

[15] 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 "Cuore", corresponding to the first three chapters of the novel.

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

TEST GOLD	UDPipe 2 -VIT model	Ersatz
1) «Al sagrestano gli crede?» 2) «Perché?»	1) » «Al sagrestano gli crede?» «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!– 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! ha voglia di scherzare, lei,» disse questa.	1) Anche Agnese, veda; anche Agnese… » «Uh! 2) ha voglia di scherzare, lei,» disse questa. «

Table 4
Results on about 90 sentences taken from other 19th-century novels. `Stanza retr.` refers to the model retrained on Manzoni's novel, as described in Section 6.

	Malavoglia	Pinocchio	Cuore
`spaCy`	0.73	0.35	**0.84**
`CoreNLP ssplit`	0.76	0.72	0.62
`SentenceSplit.`	0.77	0.45	0.68
`UDPipe`	0.75	0.79	0.67
`Stanza`	0.71	0.70	0.61
`Stanza retr.`	**0.90**	**0.89**	0.69
`Ersatz`	0.72	0.75	0.66
`Punkt`	0.73	0.77	0.66
`WtP`	0.53	0.78	0.39

The results obtained are once again lower than those reported for contemporary texts but the model retrained on "I Promessi Sposi" shows improved performance for all novels, especially when applied on "I Malavoglia" and on "Le avventure di Pinocchio" (+19 points with respect to the default `Stanza` combined model in both cases); the improvement is more limited for "Cuore" (+ 8 points).

The rule-based approach is promising but with different systems (`spaCy` for "Cuore" and `ssplit` for "I Malavoglia"). Instead, the VIT model of `UDPipe`, and therefore a supervised approach, is the best on "Le avventure di Pinocchio". Some tools obtain extremely different results depending on the text they process. `spaCy` and `Sentence Splitter` record a very low result on "Le avventure di Pinocchio" (0.35 and 0.45 respectively) while `WtP` has an F1 of only 0.39 on "Cuore", half of what it achieved on "Le avventure di Pinocchio".

This diversified situation is principally due to the fact that each novel presents unique characteristics, even in punctuation.

"I Malavoglia" is a choral novel in which the various styles of speech of the characters and the narrative voice are mixed together. Punctuation marks largely represent this mixture. 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 [33], which sometimes intersperse the diegesis, whether introduced by colons or not, and sometimes isolate a complete enunciative section. The long dash (–), instead, has a number of different functions [34]: 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 [35]. 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. Additionally, Collodi frequently uses punctuation clusters, specifically the exclamation mark followed by suspension points (!...), at the end of sentences [36], a possibility mostly not contemplated by late 19th-century grammars.

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. In "Cuore", the linguistic form is simple and plain: the sentences are mainly short and often end with a standard strong punctuation mark, followed by a capital letter. Direct speech is clearly 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. Since the lines of dialogue are perfectly 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.

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".

8. Conclusions

This paper presents an assessment of the performance of eight sentence splitting tools adopting different approaches on four 19th-century novels: "I Promessi Sposi" by Alessandro Manzoni, "I Malavoglia" by Giovanni Verga", "Le avventure di Pinocchio" by Carlo Collodi, and "Cuore" by Edmondo de Amicis. Although these texts belong to the same historical period, they show specific features depending on the form and content of the novel as well as the author's stylistic choices. Among these features is punctuation, which in the late 19th century had not reached a detectable stability yet and was rather experiencing a paradigmatic change.

Since sentence splitting for Western languages, including Italian, relies heavily on punctuation disambiguation, applying existing tools to the four novels considered has resulted in performances well below the standards. These texts demonstrate that sentence splitting is not a completely solved task.

On the other hand, applying the model retrained on "I Promessi Sposi" to the other three novels showed significant improvements for "Le avventure di Pinocchio" and "I Malavoglia", and a moderate improvement for "Cuore." This result suggests that shared historical context and belonging to the same textual genre may offer sufficient similarities to improve the model's performance. However, the example of "Cuore" is evidence of how this is sometimes not enough: some specific features in form, punctuation and style continue to affect sentence splitting, demonstrating that although retraining may mitigate some problems, it does not completely overcome the inherent variability of these texts.

Philologists have increasingly focused on preserving the original punctuation as a part of the author's creation of the text, providing valuable and reliable supports of study for scholars of linguistics and the history of the Italian language. Their combined knowledge is precious for achieving accurate sentence splitting in these texts. Thus, sentence splitting can be an interesting common ground between different disciplines, potentially leading to the development of tools for the automatic analysis of historical literary texts. This field remains under-explored in the Italian context, offering significant opportunities for further study and cross-disciplinary collaboration.

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

[1] I. Bonomi, A. Masini, S. Morgana, M. Piotti, et al., Elementi di linguistica italiana, volume 103, Carocci, 2010.

[2] D. D. Palmer, Chapter 2: Tokenisation and sentence segmentation, Handbook of natural language processing (2007).

[3] R. Dridan, S. Oepen, Document parsing: Towards realistic syntactic analysis, in: Proceedings of The 13th International Conference on Parsing Technologies (IWPT 2013), 2013, pp. 127–133.

[4] R. Wicks, M. Post, Does sentence segmentation matter for machine translation?, in: Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, pp. 843–854.

[5] Y. Liu, S. Xie, Impact of automatic sentence segmentation on meeting summarization, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 5009–5012.

[6] 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.

[7] 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.

[8] M. Sanguinetti, C. Bosco, A. Lavelli, A. Mazzei, O. Antonelli, F. Tamburini, PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1279.

[9] E. Tonani, Premessa. Tra punteggiatura e tipografia, in: E. Tonani (Ed.), 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, 2010, pp. 13–28.

[10] A. Ferrari, Punteggiatura, in: G. Antonelli, M. Motolese, L. Tomasi (Eds.), Storia dell'italiano scritto. Grammatiche, volume IV, Carocci, Roma, 2018, pp. 169–202.

[11] B. Mortara Garavelli, Prontuario di punteggiatura,

Laterza, Bari, 2003.
[12] A. Manzoni, F. Ghisalberti, A. Chiari, L'ultima revisione dei Promessi Sposi, in: Tutte le opere di Alessandro Manzoni. I Promessi Sposi, volume II, Mondadori, Milano, 1954, pp. 789–989.
[13] M. Straka, UDPipe 2.0 prototype at CoNLL 2018 UD shared task, in: D. Zeman, J. Hajič (Eds.), Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 197–207. URL: https://aclanthology.org/K18-2020. doi:10.18653/v1/K18-2020.
[14] M.-C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational linguistics 47 (2021) 255–308.
[15] T. Kiss, J. Strunk, Unsupervised multilingual sentence boundary detection, Computational Linguistics 32 (2006) 485–525. URL: https://aclanthology.org/J06-4003. doi:10.1162/coli.2006.32.4.485.
[16] Y. Liu, A. Stolcke, E. Shriberg, M. Harper, Using conditional random fields for sentence boundary detection in speech, in: Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05), 2005, pp. 451–458.
[17] R. Sheik, T. Gokul, S. Nirmala, Efficient deep learning-based sentence boundary detection in legal text, in: Proceedings of the Natural Legal Language Processing Workshop 2022, 2022, pp. 208–217.
[18] D. Rudrapal, A. Jamatia, K. Chakma, A. Das, B. Gambäck, Sentence boundary detection for social media text, in: Proceedings of the 12th International Conference on Natural Language Processing, 2015, pp. 254–260.
[19] A. A. Azzi, H. Bouamor, S. Ferradans, The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain, in: C.-C. Chen, H.-H. Huang, H. Takamura, H.-H. Chen (Eds.), Proceedings of the First Workshop on Financial Technology and Natural Language Processing, Macao, China, 2019, pp. 74–80. URL: https://aclanthology.org/W19-5512.
[20] T. Brugger, M. Stürmer, J. Niklaus, MultiLegalSBD: a multilingual legal sentence boundary detection dataset, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, 2023, pp. 42–51.
[21] J. Read, R. Dridan, S. Oepen, L. J. Solberg, Sentence boundary detection: A long solved problem?, in: M. Kay, C. Boitet (Eds.), Proceedings of COLING 2012: Posters, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 985–994. URL: https://aclanthology.org/C12-2096.
[22] A. Ferrari, L. Lala, F. Longo, F. Pecorari, B. Rosi, R. Stojmenova, La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale, Carocci, Roma, 2018.
[23] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.
[24] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, 2005, pp. 79–86. URL: https://aclanthology.org/2005.mtsummit-papers.11.
[25] R. Delmonte, A. Bristot, S. Tonelli, VIT-Venice Italian Treebank: Syntactic and quantitative features., in: Sixth International Workshop on Treebanks and Linguistic Theories, volume 1, Northern European Association for Language Technol, 2007, pp. 43–54.
[26] R. Wicks, M. Post, A unified approach to sentence segmentation of punctuated text in many languages, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 3995–4007. URL: https://aclanthology.org/2021.acl-long.309. doi:10.18653/v1/2021.acl-long.309.
[27] B. Minixhofer, J. Pfeiffer, I. Vulić, Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 7215–7235. URL: https://aclanthology.org/2023.acl-long.398. doi:10.18653/v1/2023.acl-long.398.
[28] A. Palmero Aprosio, G. Moretti, Tint 2.0: an all-inclusive suite for NLP in Italian, in: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Accademia University Press, 2018, pp. 311–317.
[29] A. Manzoni, B. Colli, I Promessi Sposi. Edizione genetica della Quarantana, Casa del Manzoni, Milano, 2024.
[30] G. Verga, F. Cecco, I Malavoglia, Fondazione Verga-Interlinea, Catania-Novara, 2014.
[31] C. Collodi, O. Castellani Pollidori, Le avventure di Pinocchio, Fondazione nazionale Carlo Collodi, Pescia, 1983.
[32] E. De Amicis, L. Tamburini, Cuore. Libro per ragazzi, Einaudi, Torino, 2018 (1° ed. 1972).

[33] G. B. Bronzini, Proverbi, discorso e gesto proverbiale nei «Malavoglia», in: I Malavoglia. Atti del Congresso Internazionale di Studi (26-28 novembre 1981), Biblioteca della Fondazione Verga, Catania, 1982, pp. 637–684.

[34] E. Tonani, Il 'bianco di dialogato' e il trattamento tipografico del discorso diretto, in: E. Tonani (Ed.), 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, 2010, pp. 103–136.

[35] R. Pellerey, Pinocchio tra dialogo e scrittura, Belfagor 60 (2005) 267–284. URL: https://www.jstor.org/stable/26150287.

[36] O. Castellani Pollidori, Introduzione, in: C. Collodi, O. Castellani Pollidori (Eds.), Le avventure di Pinocchio, Fondazione nazionale Carlo Collodi, Pescia, 1983, pp. XIII–LXXXIV.