TEI Encoding as a Unified Structure for Multilingual Digital Editions: The *LeggoManzoni* Case Study

Mariia Levchenko[1], Beatrice Nava[2], Ersilia Russo[3]
[1] Università "Alma Mater" di Bologna, Italy- maria.levchenko@studio.unibo.it
[2] Universität Wien, Austria - beatrice.nava@univie.ac.at
[3] Università degli Studi di Firenze, Italy- ersilia.russo@unifi.it

ABSTRACT (ENGLISH)

This paper presents *LeggoManzoni*, a digital edition of Alessandro Manzoni's *I promessi sposi* that tests the effectiveness of TEI encoding as a unified framework for managing complex textual relationships. The project aligns forty Italian commented editions from 1893 to 2021 and fourteen translations in ten languages (1845-2022) with the original text of the 1840-42 edition. We describe an investigative stand-off markup approach that keeps the source text separate from commentaries and translations while maintaining accurate references through automated alignment pipelines. The commentary pipeline processes digitized historical editions and implements a flexible text-matching algorithm with a 70% match threshold, achieving 87% accuracy on 2.441 processed comments. For translations, we use the Bertalign algorithm, configuring broader alignment parameters to account for cross-linguistic variation. Both pipelines generate TEI-compliant files using the same encoding pattern based on <note> elements with word-level references, allowing easy integration through XSLT transformations. The outcome of the project demonstrates TEI's ability to serve as a sustainable foundation for multilingual digital editions, bridging scholarly research and educational applications while ensuring long-term preservation through convertibility to emerging formats.

Keywords: *I promessi sposi*; XML/TEI; Multilingual Digital Editions; stand-off markup; text alignment

ABSTRACT (ITALIANO)

La codifica TEI come struttura unificata per le edizioni digitali multilingue: Il caso di studio LeggoManzoni*.*
Questo articolo presenta *LeggoManzoni*, un'edizione digitale de *I promessi sposi* di Alessandro Manzoni che testa l'efficacia della codifica TEI come framework unificato per la gestione di relazioni testuali complesse. Il progetto allinea quaranta edizioni commentate in italiano dal 1893 al 2021 e quattordici traduzioni in dieci lingue (1845-2022) con il testo originale dell'edizione del 1840-42. Adotta un markup stand-off investigativo che tiene separato il testo di partenza dai commenti e dalle traduzioni, mantenendo riferimenti accurati attraverso pipeline di allineamento automatizzate. La pipeline dei commenti elabora le edizioni storiche digitalizzate e implementa un algoritmo flessibile di corrispondenza del testo con una soglia di corrispondenza del 70%, raggiungendo un'accuratezza dell'87% su 2.441 commenti. Per le traduzioni, utilizziamo lo strumento Bertalign, configurando parametri di allineamento più ampi per tenere conto delle variazioni linguistiche. Entrambe le pipeline generano file conformi a TEI utilizzando lo stesso modello di codifica basato su elementi <note> con riferimenti a livello di parola, consentendo una facile integrazione attraverso XSLT. Il risultato del progetto dimostra la capacità di TEI di fungere da base sostenibile per le edizioni digitali multilingue, stabilendo un legame tra la ricerca scientifica e le applicazioni didattiche e garantendo al contempo la conservazione a lungo termine attraverso la convertibilità in formati in espansione.

Parole chiave: *I promessi sposi*; XML/TEI; edizioni digitali multilingue; stand-off markup; allineamento testuale

1. INTRODUCTION. *LEGGOMANZONI*: OLD ROOTS, NEW LEAVES

Our community is nowadays deeply aware of the potential of digital editions and knowledge sites (Tomasi, 2016, p. 133). These resources serve as repositories of vast amounts of information, presented at various levels and tailored to diverse users - humans or machines. We are of course also fully committed to addressing the challenges that these resources pose. This proposal seeks to highlight the constructive potential of digital editions by presenting *LeggoManzoni* as a successful case study of an edition project. The project is examined from a twofold perspective. First, we provide an overview of its architecture and content, focusing on the output of the project, the digital environment *LeggoManzoni*.[1] Second, we explore

[1] https://projects.dharc.unibo.it/leggomanzoni/ (cons. 25/01/2025).

how TEI encoding can serve as a unified framework for creating Digital Editions with commentaries, as well as Multilingual Digital Editions.

LeggoManzoni was conceived by Paola Italia and Francesca Tomasi (FICLIT Department of the University of Bologna) and launched in 2019 as part of the PRIN project *Manzoni Online: unpublished manuscripts and documents, tradition and translation*.[2] Built upon a collaboration between the University and 20 Italian high schools, the project has resulted in a comprehensive digital environment that provides access to the complete text of the 1840–42 edition of *I promessi sposi* (*The Betrothed*). The edition is enriched with critical commentaries, translations and a digitized version with metadata of the original edition. *LeggoManzoni* therefore provides access to the novel from multiple perspectives, allowing users to engage with the masterpiece both as it was originally conceived by the author and as it has been interpreted by scholars over time. The purpose of this resource is to present Manzoni's well-known work and his cultural legacy in a new format, integrating various elements - such as commentaries, translations, and illustrations - to enhance the experience of reading and studying the text. The very construction of the project, which has evolved over time, has in fact also enabled technical experimentations that could serve as a starting point for the development of similar projects, adhering to best practices for reusing data, ideas, and code. Specifically, it has created opportunities to develop and apply digital methodologies and technologies for encoding, integrating, and aligning texts, through the design of an automated and reproducible pipeline. What the project actually aims to achieve is to create new leaves from old, established roots, both in terms of knowledge of Manzoni's work - for scholars as well as for schools - and in terms of digital technologies and methodologies.

2. THE TEI ENCODING FRAMEWORK

At the core of the edition lies a TEI framework. Both the text of the novel and the commentaries are encoded according to this standard, which serves as the foundation of the whole infrastructure, informing the entire pipeline. The main concern when modelling the connection between the novel's text and the various critical commentaries was, of course, managing overlaps. In fact, the notes from different commentaries often refer, as expected, to the same portions of text or to overlapping sections. To address this, we chose a stand-off markup approach, which keeps the novel’s text separate from the commentaries while allowing references between the text file and each individual commentary encoded as a separate file, thus avoiding prohibited nesting and intersection of tags. The encoding model is therefore based on creating separate XML/TEI files for each chapter of the novel along with a rough tokenization of each chapter’s text using a simple XSLT stylesheet.[3] The stylesheet tags each word in the novel (keeping punctuation with the preceding word) as a <w> element, which includes a unique identifier (@xml:id attribute), enabling cross-referencing from the commentary files. An encoded sample word appears as: <w xml:id="c1_10001">Quel</w>, where the first part of the @xml:id refers to the chapter, and the second part numbers the word in sequence.

For the commentaries, the encoding model uses <note> elements to store essential identification information (@xml:id and @type attributes) and links to the corresponding text (@target and @targetEnd attributes). Each note contains two child elements: a <ref> that holds the plain-text reference to the portion of text being commented, followed by a colon, and the plain-text commentary itself. A commentary entry looks as follows:

```
<note xml:id="Russo_cap1-n51"
      type="comment"
      target="quarantana/cap1.xml#c1_13095"
      targetEnd="quarantana/cap1.xml#c1_13101">
    <ref rend="bold">chiudendo il libro con le due mani</ref>:
    Il breviario era rimasto dunque aperto, durante il colloquio…
</note>
```

More in detail, each <note> element includes a unique identifier (@xml:id) combining the commentator's name, chapter, and entry number; the type attribute specifying it as commentary (type="comment"); a precise reference link to the source text through @target and @targetEnd attributes, using word-level IDs

[2] https://www.alessandromanzoni.org/ (cons. 25/01/2025).

[3] We didn’t use a specific tokenization software since a perfect accuracy of this step was not required (see below the commentary pipeline) and a custom XSLT easily served the scope.

from the original TEI files and, finally, the quoted text wrapped in a <ref> element with bold rendering, plus the commentary entry following the quoted passage.

The commentaries' texts have been enriched with other information such as place names, bibliography references, placeholders for images, internal references from one note to another and so on.[4] However, the simple structure described above is the key concept we want to focus on, since it allowed us to define and build the automatic alignment mechanism described in the next section.

The alignment system has been created for the massive corpus of forty commented editions of *I promessi sposi*, representing almost a century of Italian literary scholarship, from Policarpo Petrocchi's 1893-1902 to Raimondi and Bottoni's 2021. The collection represents the work of major Italian literary scholars and includes both academic and educational editions from prominent Italian publishers.

Currently the project has completed the alignment and encoding of the Introduction and Chapter I for all forty editions, with two editions - Luigi Russo (1935) and Mariarosa Bricchi (1996) - fully completed. Work is underway to extend the alignment to the remaining chapters in all editions. The temporal distribution of these editions is significant: from the historical publications beginning with Petrocchi in 1893 through to contemporary critical works in the 2020s, demonstrating the breadth of critical perspectives and interpretative approaches to the text.

3. COMMENTARY PIPELINE

The preparation of the commentaries follows a modular approach in which each commented edition of *I promessi sposi* is processed in parallel, maintaining a clear separation between the source text and its commentaries. The process starts with the digitised historical editions, which are OCR'd into plain text files, organised chapter by chapter. Each chapter's commentary is stored in a separate .txt file, with each line following this format: "commented text": corresponding commentary.

The alignment process uses the TEI-encoded chapters of the novel as reference points but maintains the separation of content. The system reads the TEI files to create a normalised word index, which serves as the basis for identifying the exact locations of annotated passages. The alignment algorithm uses a flexible text-matching technique that accounts for OCR variations and encoding differences common in digitised historical texts. It can in fact tolerate small gaps (up to two words) and requires a match rate of 70%, making it robust enough to handle textual variation while maintaining accuracy.

Once the text fragments have been aligned, the system generates new TEI files for the chapter commentaries. Each file contains a series of note elements structured according to the encoding pattern defined in the encoding model, namely a note with a starting and ending point (see above section 2). This encoding structure maintains precise alignment with the source text while preserving the scholarly apparatus of the commentary. The reference system allows the exact location of commented passages while maintaining the independence of the commentary files from the base text.

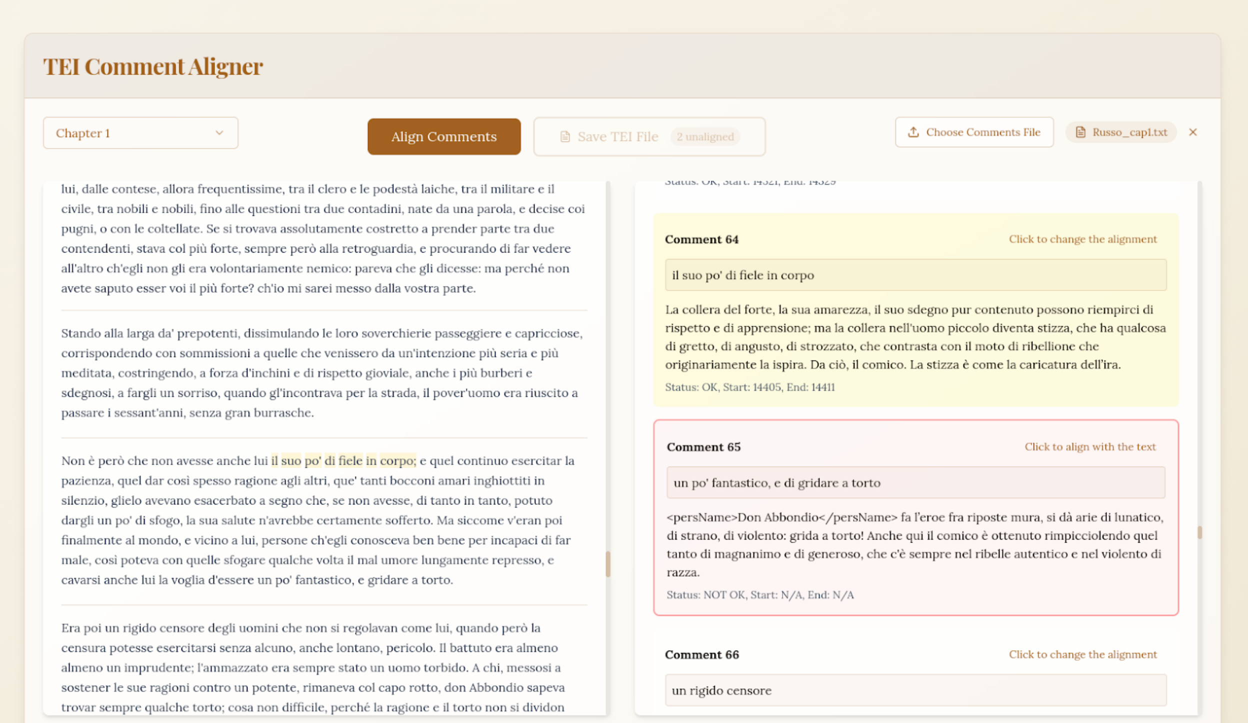
The resulting TEI files contain full metadata about the source edition, including bibliographic and editorial information. This separation of content - retaining the original TEI-coded novel text while creating separate, linked commentary files - has several advantages. It allows for easier management of multiple commented editions, independent updating of either text or commentary, and flexible presentation options in the digital edition interface using XSLT files.

Alignment Accuracy Analysis. The performance of the alignment algorithm was evaluated across all forty commented editions, processing a total of 2,441 comments. Analysis of the alignment results showed an overall error rate of 13.03%. The errors fell into several different categories: 44 cases (1.8%) where the start position could not be identified, 44 cases where the end position was missing, 152 cases (6.2%) where both start and end positions were not identified, and 78 cases (3.2%) where the algorithm linked to incorrect occurrences of text spans appearing more than once in the chapter.

These results highlighted specific challenges in processing comments from the raw text. This error rate motivated the development of a web-based validation interface designed to facilitate the efficient identification and correction of misaligned comments. The interface provides specific tools for handling each type of error, allowing annotators to quickly locate problematic alignments and make the necessary corrections while preserving the TEI structure.

[4] The model was initially tested through the manual encoding of a selection of commentaries, thanks to the collaboration with the high schools previously mentioned. Students were tasked with encoding the structure of the commentary for a single chapter after receiving an introduction to Digital Scholarly Editing, as well as encoding languages and practices, during three PCTO (Pathways for Soft skills and Orientation) organized by the University of Bologna. Details on the schools involved in the project can be found here: https://projects.dharc.unibo.it/leggomanzoni/progetto.

Web-Based Alignment Management System. The alignment system was implemented as a Next.js web application that provides an intuitive interface for both automatic and manual alignment of comments (see Fig. 1). This interface integrates with the backend alignment service, while giving annotators full control over the alignment process and validation.



lui, dalle contese, allora frequentissime, tra il clero e le podestà laiche, tra il militare e il civile, tra nobili e nobili, fino alle questioni tra due contadini, nate da una parola, e decise coi pugni, o con le coltellate. Se si trovava assolutamente costretto a prender parte tra due contendenti, stava col più forte, sempre però alla retroguardia, e procurando di far vedere all'altro ch'egli non gli era volontariamente nemico: pareva che gli dicesse: ma perché non avete saputo esser voi il più forte? ch'io mi sarei messo dalla vostra parte.

Stando alla larga da' prepotenti, dissimulando le loro soverchierie passeggiere e capricciose, corrispondendo con sommissioni a quelle che venissero da un'intenzione più seria e più meditata, costringendo, a forza d'inchini e di rispetto gioviale, anche i più burberi e sdegnosi, a fargli un sorriso, quando gl'incontrava per la strada, il pover'uomo era riuscito a passare i sessant'anni, senza gran burrasche.

Non è però che non avesse anche lui il suo po' di fiele in corpo; e quel continuo esercitar la pazienza, quel dar così spesso ragione agli altri, que' tanti bocconi amari inghiottiti in silenzio, glielo avevano esacerbato a segno che, se non avesse, di tanto in tanto, potuto dargli un po' di sfogo, la sua salute n'avrebbe certamente sofferto. Ma siccome v'eran poi finalmente al mondo, e vicino a lui, persone ch'egli conosceva ben bene per incapaci di far male, così poteva con quelle sfogare qualche volta il mal umore lungamente represso, e cavarsi anche lui la voglia d'essere un po' fantastico, e gridare a torto.

Era poi un rigido censore degli uomini che non si regolavan come lui, quando però la censura potesse esercitarsi senza alcuno, anche lontano, pericolo. Il battuto era almeno almeno un imprudente; l'ammazzato era sempre stato un uomo torbido. A chi, messosi a sostener le sue ragioni contro un potente, rimaneva col capo rotto, don Abbondio sapeva trovar sempre qualche torto; cosa non difficile, perché la ragione e il torto non si dividon

Status: OK, Start: 14321, End: 14329

Comment 64

il suo po' di fiele in corpo

La collera del forte, la sua amarezza, il suo sdegno pur contenuto possono riempirci di rispetto e di apprensione; ma la collera nell'uomo piccolo diventa stizza, che ha qualcosa di gretto, di angusto, di strozzato, che contrasta con il moto di ribellione che originariamente la ispira. Da ciò, il comico. La stizza è come la caricatura dell'ira.

Status: OK, Start: 14405, End: 14411

Comment 65

un po' fantastico, e di gridare a torto

<persName>Don Abbondio</persName> fa l'eroe fra riposte mura, si dà arie di lunatico, di strano, di violento: grida a torto! Anche qui il comico è ottenuto rimpicciolendo quel tanto di magnanimo e di generoso, che c'è sempre nel ribelle autentico e nel violento di razza.

Status: NOT OK, Start: N/A, End: N/A

Comment 66

un rigido censore

Figure 1. User Interface for Comment Alignment (https://manzoni-comments-aligner.vercel.app)

The workflow begins with the annotator selecting a chapter and uploading a text file containing comments for that chapter. Each comment in the file follows the standard format of quoted text followed by commentary text, separated by a colon. The system processes this input through the alignment algorithm and presents the results in an interactive interface. The interface displays the TEI text of the novel chapter alongside the processed comments, with clear visual indicators of alignment status. Comments requiring manual intervention are visually highlighted in red. For each comment, the system displays both the referenced text and the corresponding comment, along with the precise alignment positions derived from the TEI word identifiers.

When manual alignment is required, annotators can use an interactive text selection mode. They click on a comment that needs correcting, then select the correct text passage directly in the TEI content view. The system automatically updates the alignment references, maintaining the proper TEI linking structure. This combination of automatic alignment and manual validation ensures both efficiency and accuracy.

Once all alignments have been verified, the system generates a new TEI file for the chapter commentary. A metadata dialogue allows annotators to select the specific edition details from the *LeggoManzoni* project catalogue and provide their attribution information. The resulting TEI file contains proper namespaces, reference systems and metadata according to the project's specifications.

This web-based tool significantly streamlined the processing of forty historical commentaries by providing an efficient interface for both automatic alignment and manual verification, while ensuring consistent TEI encoding across all processed editions.

4. TRANSLATION ALIGNMENT PIPELINE

The translation alignment pipeline for the *LeggoManzoni* project has been implemented on a multilingual corpus of translations from 1845 to 2022, including English (1845, 1876, 1972, 2022), French (1874), German (1880), Spanish (1858), Dutch (1849), Polish (1882), Russian (1854, 1936, 1999), Ukrainian (1982), Chinese (1998), Finnish (1910). The process uses Bertalign (Liu & Zhu, 2022), a sentence alignment module that uses multilingual sentence transformer models, in this case the Language-agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2020) model, to produce accurate alignments between the Italian source text and its translations.

The pipeline starts with text pre-processing, where both the Italian source text (Quarantana) and each translation are normalised and segmented. The core alignment process takes place in two stages. The first stage identifies basic sentence pairs, establishing initial correspondences between the source and translated texts. The second stage implements a more sophisticated alignment algorithm that can handle complex translation patterns, including one-to-many, many-to-one and many-to-many sentence relationships. This two-stage approach allows the system to handle structural differences between languages while maintaining semantic correspondence. The alignment system uses a maximum alignment window of 8 sentences, 5 nearest target neighbors and a sliding window of 5 sentences for context, ensuring both precision in local alignments and consideration of the wider textual context. While Sprugnoli & Sartor (2023) used more restrictive parameters (maximum alignment types: 6, k nearest neighbors: 3, search window: 5) for aligning the Ventisettana with its English translation, our multilingual pipeline requires broader settings to accommodate greater syntactic variation across languages.

Following the same TEI encoding pattern established for commentaries, the system generates XML files in which each aligned segment is encoded as a <note> element with a unique identifier that combines language, chapter, and segment number (see Levchenko, 2024 for text segmentation details). Using the same target attribute structure to reference the source text enables precise word-level linkage between the original text and the translations. Like DiScEPT's approach using the <annotation> element for stand-off markup (Hohenegger et al., 2024), this unified TEI approach allows translations to be seamlessly integrated into the existing web interface, while maintaining comprehensive metadata about each edition and taking advantage of independent file management and flexible presentation options through XSLT transformations.

5. CONCLUSIONS

In conclusion, the project led to the creation of an enriched edition of Manzoni's text and gave the opportunity for experimentation with automated pipelines for text alignment, starting from a simple XML/TEI structure. Despite common critiques of TEI's complexity, our implementation proved highly effective in two key areas: (1) enabling precise and flexible alignment between source text, commentaries, and translations through automated pipelines, and (2) bridging academic research and classroom education through an accessible digital interface, thanks to the mentioned PCTO (and, hopefully, the edition will remain a useful teaching tool: see Levchenko, Menna & Nava, 2024). The automated alignment pipelines and web-based validation tools we developed provide a reproducible methodology that can serve as a model for similar digital edition projects. While digital humanities may shift toward newer formats like JSON, TEI's well-structured nature ensures easy conversion of editions built over past decades to emerging standards, preserving the scholarly value of digital cultural heritage.

REFERENCES

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. Muresan, S., Nakov, P., & Villavicencio, A. (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: Association for Computational Linguistics, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62.

Hohenegger, H., Mancinelli, T., Ciotti, F., Boschetti, F., Del Grosso, A. M., & De Longis, E. (2024). Paul Klee, Tunisreise e Bildnerische Formlehre: un caso studio di DiScEPT (Digital Scholarly Editions Platform and Aligned Translations). Di Silvestro, A., & Spampinato, D. (Eds.), Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings of the 13th Annual Meeting of AIUCD 2024, Catania 28-30 May 2024. Catania: AIUCD2024, 186-190. https://doi.org/10.6092/unibo%2Famsacta%2F7927.

Levchenko, M. (2024). Automatic Translation Alignment Pipeline for Multilingual Digital Editions of Literary Works. Proceedings of the Computational Humanities Research Conference 2024. Aarhus, Denmark, December 4-6, 2024, 1086-1104. https://doi.org/10.48550/arXiv.2410.13255.

Levchenko, M., Menna, G., & Nava, B. (2024). *LeggoManzoni*. Quaranta commenti alla Quarantana. Russo, E. (Ed.), Manzoni e Leopardi in digitale. Idee e proposte per la scuola. Bologna: Clueb, 73-86. ISBN: 978-88-491-5800-7.

Liu, L., & Zhu, M. (2023, june). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. Digital Scholarship in the Humanities, 38, 2, 621-634. https://doi.org/10.1093/llc/fqac089.

Sprugnoli, R., & Sartor, M. (2024). That branch of the lake of Como...: Developing a new resource for the analysis of *I Promessi Sposi* and its historical translations. Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023). Torino: Accademia University Press, 421–427. ISBN: 9791255000846. https://ceur-ws.org/Vol-3596/paper48.pdf.

Tomasi, F. (2016). Edizioni o archivi digitali? Knowledge sites e apporti disciplinari. Bonsi, C., & Italia, P. (Eds.), Edizioni Critiche Digitali Digital Critical Editions: Edizioni a confronto / Comparing Editions. Roma: Sapienza Università Editrice, 130-136. https://doi.org/10.13133/9788893770033.