When Lexicographic Quotations Become a Corpus: To Deduplicate or Not to Deduplicate?

Manuel Favaro[§], Elisa Guadagnini[§], Eva Sassolini[§], Marco Biffi*^, Simonetta Montemagni[§]

[§]Istituto di Linguistica Computazionale "Antonio Zampolli" – CNR
*Università di Firenze
^Accademia della Crusca
[§]name.surname@ilc.cnr.it
*marco.biffi@unifi.it

Abstract

Historical dictionaries are increasingly reused as sources for diachronic language corpora. In this context, lexicographic quotations represent a valuable yet challenging type of data, as they are both editorially curated and diachronically representative. A major issue in their computational reuse is the presence of duplicate and near-duplicate quotations. This paper addresses quotation deduplication in corpora derived from lexicographic resources. We introduce QRD (Quotation Reuse Detection), a multi-stage pipeline designed to identify, compare, and cluster quotations based on graded similarity rather than binary matching. The approach combines string-based similarity measures, iterative threshold analysis, and clustering, enabling both quantitative and qualitative investigation of quotation reuse. Our results show that deduplication in this context cannot be reduced to the automatic elimination of redundant data. The variability observed in the quotations - ranging from OCR-related noise to substantial editorial variation - reflects both technical and structural factors and calls for a more nuanced approach. QRD supports the identification of OCR-related errors and reveals patterns of textual reuse underlying the compilation of the dictionary. We argue that quotation deduplication should be conceived primarily as a task of identification and clustering. This perspective reframes deduplication from a data-cleaning operation into an analytical methodology for historically and editorially curated textual resources.

Keywords: Historical Corpora, Text Deduplication, Data Matching Process, Historical Lexicography

1. Introduction

Historical dictionaries are among the most important resources for the study of language over time, as they provide a structured interpretation of the different layers of lexical information (semantics, register and usage labels, multiword expressions, etc.), grounded in diachronic textual data. In these resources, authorial quotations play a central role: they document actual language use, support lexicographic decisions, and offer a bridge between linguistic description and primary textual sources (Atkins and Rundell, 2008). For this reason, quotation-based resources derived from historical dictionaries are increasingly reused in digital humanities and natural language processing (NLP) research, including diachronic corpus studies (see e.g. Hoffmann, 2004; Rohdenburg, 2013).

However, the reuse of lexicographic quotations as textual corpora raises a number of data quality issues that are often overlooked. One of the most prominent among these is the presence of duplicate or near-duplicate quotations. In historical dictionaries, it is frequently the case that the same textual passage appears multiple times across different entries, possibly in slightly different forms. Such duplication can result from the reuse of canonical sources, from editorial adaptation of quotations, or from the inheritance of citation material across lexicographic traditions (Béjoint, 2010). From a computational perspective, unaddressed duplication can affect quantitative analyses, bias frequency counts, and reduce the reliability of downstream NLP tasks using this type of resources (Lee *et al*., 2022). At the same time, in this specific context, duplication should not be considered merely as noise: repeated quotations reflect the centrality of canonical sources, and therefore need to be preserved and made explicit rather than removed indiscriminately.

In this paper, we address the open issue of how to treat duplicated quotations in the creation of a corpus extracted from the examples in a historical dictionary of Italian. This work is carried out as part of ongoing efforts aimed at the digitization, organization, and computational exploitation of the *Grande Dizionario della Lingua Italiana* ('Great Dictionary of the Italian Language', henceforth GDLI) (Sassolini *et al*., in preparation). In particular, the work presented here builds on and extends Favaro *et al*. (2022), which describes the creation of a diachronic corpus of Italian based on GDLI quotations, focusing on their extraction from a structured digital representation of the dictionary and on subsequent linguistic annotation. While that work primarily addresses corpus construction and annotation, it also implicitly highlights data quality issues that become crucial when quotations are reused as corpora.

The resulting resource, hereafter referred to as the *GDLI Quotations Corpus* (GDLI-QC), shares

Proceedings of LT4HALA 2026 @ LREC 2026, pages 49–57.

the main features of the original lexicographic resource, such as editorial selection, diachronic breadth, and textual inconsistencies, but also exhibits properties introduced by the digitization process, including OCR-related noise. Within the corpus building process, a key issue concerns the treatment of duplicate or near-duplicate quotations. In tackling this issue, we explore why and where duplicates occur, how they can be detected, and what to do about them. This issue is analysed from a threefold perspective, namely with respect to: the identification of OCR-related errors occurring across identical or near-identical textual passages; quantitative analyses of the corpus; and, last but not least, the impact of duplicate examples for training of linguistic annotation models.

Taken together, these considerations indicate that duplication in lexicographic quotation corpora cannot be addressed through simple removal strategies, but instead requires an approach that accounts for both computational and editorial dimensions. In response to the issues outlined above, we adopt a method that does not eliminate redundancy indiscriminately, but aims to identify clusters of quotations deriving from the same underlying textual passage. This strategy allows us to improve corpus quality while preserving editorial variation, maintaining transparency with respect to lexicographic practices, and supporting scholarly inspection. On this basis, we propose a transparent multi-stage duplication processing pipeline specifically tailored to editorially curated lexicographic data.

The remainder of the paper is organized as follows: Section 2 presents the main contributions of this work; Section 3 reviews related work on text deduplication and historical language data; Section 4 describes the GDLI quotation corpus, and Section 5 the proposed methodology; Section 6 discusses the evaluation and qualitative analysis; and Section 7 concludes the paper.

2. Contributions

This paper makes the following contributions:

Contextualized problem definition - We frame the identification of duplicate lexicographic quotations as an issue of data complexity, one that requires both an analysis of the underlying lexicographic and editorial practices and an assessment of how duplication affects the transformation of a historical dictionary into a textual corpus suitable for use as an independent linguistic resource.

Task-driven motivation - We explicitly link this topic to three concrete objectives relevant for the construction and reuse of historical language resources: (i) supporting the identification of recurring OCR-related errors in digitized historical data; (ii) evaluating their effect on quantitative analyses derived from quotation-based corpora; (iii) assessing the impact of repeated quotations on training linguistic annotation models.

Methodological contribution - We propose a transparent multi-stage pipeline tailored to handle duplication of examples in editorially curated historical lexicographic data.

Resource-oriented perspective - Rather than treating deduplication as a purely eliminative step, we demonstrate how it can be used to identify clusters of quotations deriving from the same underlying textual passage, thereby enhancing corpus quality while preserving editorial variation and enabling scholarly inspection.

3. Related Work

Text deduplication can be framed as a specific instance of the more general data matching task (Christen, 2012), which aims at identifying, matching, and possibly merging items that correspond to the same underlying entity across one or more data sources, typically databases. While data matching has most commonly been applied to entities such as individuals (e.g. patients, customers, or taxpayers), it has also been extended to bibliographic records and textual data. In the latter case, the goal is to identify redundant documents or passages within a single collection or across multiple textual collections, which is the focus of the present work.

Since the early work by Broder (1997), text deduplication has been extensively studied in information retrieval and corpus construction, where it is commonly employed to remove redundant documents or passages and to improve overall data quality.

A comprehensive overview of the challenges and solutions in duplicate detection is provided by Elmagarmid *et al*. (2007). As they observe, duplicate records often lack a common key and may contain various types of errors (such as transcription mistakes, incomplete information, or inconsistent formatting) that make matching particularly challenging. To address these issues, a wide range of similarity measures has been proposed, including edit distance, token overlap, and character n-gram-based approaches, often combined with hashing or blocking techniques to ensure computational efficiency and scalability. These methods form the basis of many contemporary deduplication pipelines.

In more recent NLP research, deduplication has gained renewed importance due to its impact on language model training and evaluation. Several studies have shown that duplicate or near-duplicate texts can significantly bias empirical results and model behavior (Lee *et al*., 2022). As a consequence, deduplication is increasingly regarded as a standard pre-processing step in the creation of language resources.

Text deduplication in historical language data poses additional challenges due to orthographic variation, diachronic change, and editorial or transcriptional noise. Prior work in historical NLP emphasizes the need for normalization strategies and noise-robust representations, as well as for multi-stage pipelines that separate candidate generation from fine-grained similarity assessment (Bollmann, 2019; Piotrowski, 2012). In this context, deduplication may also serve as a tool for identifying systematic digitization errors, such as recurring OCR mistakes across repeated textual passages, a potential use that has received limited attention in previous work.

Unlike most previous approaches, which focus on entire documents or extended textual passages, the present work applies duplicate detection to GDLI quotations, that is, relatively short spans of text covering a limited number of sentences and often containing internal interruptions. This specificity introduces additional challenges, which are discussed in the remainder of the paper.

4. The GDLI Quotation Corpus

The GDLI is the most authoritative dictionary of the Italian language. Initially edited by Salvatore Battaglia and subsequently by Giorgio Barberi Squarotti, the *GDLI* comprises 21 volumes published between 1961 and 2002, amounting to more than 23,000 pages and approximately 200,000 headwords.

The lexicographic description provided by the GDLI is grounded in an exceptionally rich apparatus of quotations, ranging from the earliest extant Italian documents, such as the «Placiti capuani» of the 10th century, to contemporary usage. While retaining a strong overall literary orientation, the dictionary exhibits considerable diatextual variation: alongside lyric and narrative texts, it includes hagiographic and technical–theoretical works, as well as practical and legislative texts, in addition to material drawn from periodical publications. It should also be noted that the dictionary, published over a period of 41 years, progressively increased both the number and the range of texts excerpted, thereby significantly expanding the coverage of the Italian language. In particular, this expansion involved a broadening along the dimensions of diatopy (including dialects and regional varieties of Italian), diastraty (extending to popular and semi-educated varieties), and diaphasy (encompassing specialized and domain-specific languages). The corpus of quotes draws on passages extracted from more than 6,000 authors and nearly 14,000 distinct sources (excluding quotations taken from periodical press; cf. Biffi and Guadagnini, 2022; Biffi *et al.*, 2025). Each quotation is selected so as to constitute a syntactically complete unit and often includes a relatively extended textual span. Overall, the corpus is estimated to include more than 2,000,000 cited passages, amounting to over 40 million tokens and covering a chronological range of more than one thousand years.

GDLI quotations were automatically extracted from the TEI XML version of the dictionary, which was obtained through a semi-automatic conversion process designed at structuring the dictionary contents starting from the OCRed source. The overarching goal of this process is to obtain a detailed and articulated representation of GDLI entries by means of a sequence of iterative steps, each progressively refining and organizing the previously identified dictionary structure.

The general approach to the extraction and structuring of GDLI contents (described in Sassolini *et al.*, 2019, 2021, and Biffi *et al.*, 2020) is largely based on pattern-matching techniques. The identification criteria rely on a wide range of features, including page layout characteristics and structural information related to the different components of the lexical entry. The ultimate goal of this process is the conversion of the dictionary contents into structured macro-fields, which are subsequently mapped onto the TEI XML standard.

Quotation extraction constitutes one of the steps of this iterative workflow. The TEI XML encoding of the GDLI quotation macro-field is exemplified in Figure 1, which shows the automatically generated TEI XML representation of the entry *abiatico* ‘grandchild’. It can be noticed that for each sense the set of quotations is encoded using the `<cit>` element, which contains one or more `<bibl>/<quote>` pairs. These elements respectively encode a loosely structured bibliographic reference and the quotation text itself. The extraction process has so far been applied to volumes I–X of the dictionary (out of a total of XXI), for which the manual revision of entry segmentation has been completed.

In transforming GDLI quotations into a textual corpus, each quotation is associated with metadata concerning authorship and the headword under which it appears. The corpus currently includes all the quotations from volumes I–X, as previously noted, for a total of more than 830,000 quotations and more than 18,000,000 tokens. At this stage, quotations are extracted as independent textual units, without attempting to determine whether identical or near-identical passages occur elsewhere.

The extracted quotation set turned out to contain multiple instances of the same underlying textual passage. The most complex and interesting cases, however, involve a substantial number of quotations that differ in certain elements, while still exhibiting a clear degree of similarity. This form of duplication can be attributed to two main factors. On one hand, noise arising from OCR errors introduced near-identical or slightly corrupted strings during the digitization process.

On the other hand, many redundancies stem from editorial practices: the GDLI editors frequently reused the same authorial quotations to illustrate different lexical entries. In these instances, the text was often re-elaborated, truncated, or adapted to better suit the specific requirements of the entry, leading to a complex layer of intentional textual variation.

The resulting GDLI-QC, covering nearly half of the dictionary volumes, therefore constitutes a suitable testbed for investigating the nature and extent of quotation duplication, as well as for developing and evaluating deduplication strategies specifically tailored to historical lexicographic data.

```
<entry>
    <form type="lemma">
        <orth>Abiàtico</orth>
    </form>
    <sense level="1" n="1">
        <def>(abbiàtico, aviàtico) sm. (femm. abiòtica, abbiàtica) Dial.
        Nipote (figlio del figlio o della figlia).</def>
        <cit>
            <bibl> Comp. Antico Testamento [Tommaseo]:</bibl>
            <quote> Seppe... come 10 aveva trovato in lo bosco e comprese
            che questo era suo abiatico.</quote>
            <bibl>Vite di imperatori romani [Tommaseo] :</bibl>
            <quote> Tiberio voleva piuttosto che succedesse questo suo
            abiatico che questo Gaio.</quote>
            <bibl>Idem [Tommaseo]:</bibl>
            <quote> La figlia e le abiatiche fece usare all'esercizio
            della lana.</quote>
            <bibl>Fogazzaro, 5-18:</bibl>
            <quote> A poppa sotto la bandiera, v'era seduto don Franco
            Maironi, l'abiatico della vecchia marchesa.</quote>
        </cit>
    </sense>
    <sense level="1" n="2">
        <def>2. Antenato.</def>
        <cit>
            <bibl>Moretti, 32-116:</bibl>
            <quote> Camere... piene di tante goffaggini, oleografie,
            ritratti dei padroni di casa, d'abiatici, che non sono stati
            neppure Reggenti.</quote>
        </cit>
    </sense>
    <etym>= Lat. mediev. abiatlcus o aviatlcus (nella Lex Longobardorum,
    del sec. XXXX dal lat. class, avus (attraverso l'agg. 'avius :
    dell'avo, nipote dell'avo). È voce dell'Italia settentrionale.</etym>
</entry>
```

Figure 1: TEI XML representation of the GDLI entry *abiatico* 'grandchild'.

5. Methodology

The methodology we devised for dealing with GDLI-QC duplicate quotations consists of three different phases, respectively aimed at:

1. minimizing differences due to systematic OCR errors;
2. carrying out quotation pair comparison;
3. clustering quotations identified as potentially matching at the previous step.

The first stage involved a rigorous data-cleaning process to minimize OCR errors as much as possible. We employed a series of regular expressions (*regex*) to perform systematic normalization and correct recurring OCR errors in the quotation text. This phase was crucial to ensure that subsequent fuzzy matching would not be skewed by predictable mechanical discrepancies.

Table 1 provides a few examples of the systematic corrections implemented during this stage. As one can see, key interventions include resolving the graphic overlap between the digraph <ll> and high-density uppercase characters (e.g., U, H), as well as correcting spurious whitespace following hyphenation. These patterns facilitated a hybrid workflow of automatic and semi-automatic character substitution, significantly improving the lexical integrity of the digital resource.

Regex	Output	Correct string
([a-z])([A-ZÀÈÉÌÒÙ])(["])	"deU'adrie"; "deH'esemplare"; "aU'appianarmeli"	"dell'adrie" "dell'esemplare"; "all'appianarmeli"
(<quote>.*?)(\b[a-z]{2,})-(\s)([a-z]{2,}\b)(.*?</quote>)	"con- clusione; "; "abbas- sare"	"conclusione"; "abbassare"

Table 1: Examples of OCR systematic corrections

In the second step, quotation pairs were compared by computing string similarity using the Levenshtein distance (Levenshtein, 1966; Navarro, 2001). As a starting point, we considered the FuzzyWuzzy algorithm[1], which has been successfully employed in previous work on historical language varieties, such as BERToldo (Palmero Aprosio *et al.*, 2022). We ultimately adopted RapidFuzz, a C++ and Python reimplementation that provides substantially higher computational efficiency, enabling iterative analyses that would otherwise be impractical.

Importantly, we adapted the original tool to function not merely as a filtering mechanism for duplicate removal, but as a diagnostic instrument capable of producing fine-grained similarity information. To this end, punctuation and diacritical marks were removed prior to similarity computation. A key aspect of our approach is the iterative analysis of quotation pairs across discrete similarity intervals of 5%. This choice reflects the observation that exact duplicates (100% similarity) constitute a relatively small and easily manageable portion of the data, whereas the methodological challenges primarily arise in cases of partial similarity (near-duplicates).

Accordingly, we focused our analysis on similarity ranges between 70% and 99.99%, which proved crucial for identifying and characterizing different types of near-duplicate quotations. This subdivision allowed for a qualitative investigation going beyond the binary "duplicate/non-duplicate" dichotomy, leading to the definition of an articulated typology of duplicate examples, described in detail in Section 6. The 70% threshold was established through preliminary empirical testing on a random sample of 500

[1] https://github.com/xdrop/fuzzywuzzy

pairs. Lower thresholds (e.g., 60–69%) were found to introduce excessive noise, primarily due to short and frequent formulaic expressions.

In the final phase, we integrated the NetworkX[2] library into the pipeline to identify clusters of quotations based on configurable similarity ranges. In addition, we implemented a difference extraction function (*diff_words*) to identify and inspect divergent string segments within clustered quotations, supporting qualitative analysis of variation and error patterns. An example from Boccaccio is reported in Table 2.

Entry	Quote
Abbattere	Commossa adunque la santa dea per le costui opere, propose di riducerlo a niente, abbattendo la infiammata sua superbia, come quella degli antecessori aveva altra volta abbattuta.
Antecessore	Commossa adunque la santa dea per le costui opere, propose di riducerlo a ***mente***, abbattendo la infiammata sua superbia, come quella degli antecessori aveva altra volta abbattuta.
Degno	Commossa adunque la santa dea per le costui opere, propose di riducerlo a niente, abbattendo la infiammata sua superbia, come quella degli antecessori aveva altra volta abbattuta, ***con degno mezzo***.

Table 2: Example of a cluster of quotations characterized by minimal variations

As shown in the table, a comparison between the first two instances, cited under the entries *abbattere* and *antecessore*, reveals a near-perfect match disrupted only by the discrepancy between *niente* and *mente*: «mente» is an OCR error which, however, results in a form that is in itself admissible in Italian. The third instance differs from the previous two in that it includes an additional segment of text—which, in this specific case, justifies the inclusion of the quotation under the entry *degno*. To provide a proper context for the target adjective, the textual span was extended to include the phrase *con degno mezzo*, an addition absent in the previous two entries.

Overall, we identified a set of approximately 91,000 quotations, including both exact duplicates and near-duplicates, which represent nearly 11% of the total number of quotations in the corpus. Table 3 highlights cluster size in relation to the number of quotations involved.

Cluster Size	Number of Clusters	Involved Quotations
2	59275	118550
3	9783	29349
4	2265	9060
5	723	3615
6	240	1440
7	94	658
8	40	320
9	18	162
10	5	50
11	3	33
13	1	13
14	1	14

Table 3: Duplicates and near-duplicates cluster distribution

The pipeline implementing the three steps described above will hereafter be referred to as QRD (Quotation Reuse Detection).

6. Typology of duplicated quotations in GDLI-QC

In this section, we present a qualitative analysis of automatically clustered quotations with a similarity score ≥ 70%, including both exact (100% similarity) and near-duplicates (similarity ranging between 70% and 99.99%). A quantitative evaluation of near-duplicate identification in terms of precision and recall is not provided, as it would require manual validation of each pair. Moreover, the current corpus - limited to 10 of 21 volumes and with source mapping still ongoing - does not support yet the construction of a representative quotation sample. The analysis therefore focuses on qualitative evidence, with quantitative assessment deferred to future work on a complete and stabilized dataset.

Exact duplicates correspond to cases in which an identical string of text is cited under different headwords: for example, the quotations of Giovan Battista Marino (the foremost Italian Baroque writer) «Tu tra questi deserti, / ond'uscir mai non spero, / inculti, abbandonati, / disleal, m'abbandoni» are cited - exactly in this form - s.vv. *abbandonare*, *abbandonato*, and *deserto*[2].

As far as near-duplicate quotations are concerned, the QRD pipeline allows us to extract different types of variation, which are illustrated below. We identified several categories of near-duplicates, ranging from simple spelling differences due to OCR errors or oversights introduced directly by the editors of the printed dictionary, to quotations that share the same semantic core but result from different excerpting strategies (textual extensions and/or reductions).

[2] https://networkx.org/en/

The first type consists of cases in which quotations that are identical in the printed dictionary appear as different strings in the digitized version due to errors introduced by OCR processing. For example, the lines by Ciro da Pers (another Baroque author) «O sonno, tu ben sei fra i doni eletti / dal **ciel** concesso ai miseri mortali; / tu l'agitato sen placido assali / e tregua apporti ai combattuti affetti» cited s.vv. *agitato*, *apportare*[1]*, assalire*, and *combattuto*, and «O sonno, tu ben sei fra i doni eletti / dal **del** concesso ai miseri mortali; / tu l'agitato sen placido assali / e tregua apporti ai combattuti affetti» s.v. *affetto*[1] (in which, as can be seen, «ciel» has been erroneously OCRed as «del»). Note that this type of OCR error was not dealt with in the pre-processing stage, where the focus was on systematic OCR errors, which could be automatically corrected.

Another type of near-duplicate quotations is represented by those that already appear as such in the *GDLI* original printed version, where the difference is not simply due to an OCR error but because of a transcription error by the entry editor. For example, again citing Marino, the quote «Gli sguardi e l'orme / a le mura superbe intento gira, / e mentre queste ed altre illustri forme, / di cui son tutte effigiate, ammira, / sembra, né sa s'ei **vegghia** o pur s'ei dorme, / statua animata; imagine che spira» (s.v. *animato*) appears s.v. *ammirare* with the reading «veglia» instead of «vegghia». The cited source gives the reading «vegghia,» but both forms are admissible in Italian for the third-person singular of the verb *vegliare* ('to be awake'), and this explains the quotation error introduced s.v. *ammirare*. This type of case is also automatically extracted: the variant that distinguishes the quotes correctly detected as repeated is present in the printed original (thus raising the question of whether it would be legitimate to modify it), and, naturally, in order to determine the correct reading, one must consult the cited source.

A formally similar case, yet profoundly different in its lexicographical implications, is the series of instances in which near-duplicate quotations differ from a single reading not because of an OCR error nor because of a transcription error by the lexicographer, but because of an error in the attribution of the cited source. For example, a passage from Giovanni Villani's *Cronica* reading «Le dette sue prediche... erano molto efficaci e d'una buona loquela e di sante parole, dicendole molto dubbiose e accentive a commuovere genti» (s.v. *accentivo*) presents the same quote s.v. *accettevole*, except that the reading «accettevole» appears in place of «accentivo». Both quotes are traced back to the same source, the Moutier edition of the *Cronaca*: here the reading «accentive» is printed in the text, while the reading «accettevoli» is recorded in a note (and appears in the *Vocabolario della Crusca*, which has cited this passage since its first edition, s.v. *accettevole*) (Moutier 1845: 31, note 1)[3].

A different case is represented by identical quotes that in the dictionary are associated with two different identifying strings for the cited source: for example, the quotation «È permesso di fabbricare aceto per diluizione dell'acido acetico puro e di buon gusto purché si venda col nome di aceto artificiale» is cited s.v. *acetico* as an extract from the «Leggi sanitarie, 74-153» and s.v. *aceto* as an extract from the «Leggi industriali e commerciali, 810-153». In this instance, research into the sources would be necessary to determine whether this is a "duplicated quotation" (in which the declared source differs owing to an inconsistency in the dictionary) or a passage that appears identically in two distinct texts. At times, however, consultation of the dictionary alone is sufficient to establish that identical quotes (identified as such by the pipeline) occur in different texts and that we are therefore dealing with a case of intertextual reuse: see, for example, the celebrated incipit of Boccaccio's *Decameron*, «Umana cosa è l'avere compassione agli afflitti», which is cited directly from Boccaccio s.v. *a2* and from Torquato Tasso's reuse s.v. *compassione*. In both cases, as can be seen, the match flagged by the QRD pipeline is not sufficient to determine whether we are truly dealing with replicated quotes: the judgment of a human user is necessarily required.

So far, we have examined cases in which the text strings correspond (punctuation marks notwithstanding, as noted above). The majority of near-duplicates, however, do not present identical text strings: the same passage may be excerpted differently from one entry to another, including additional words at the beginning or end of the quote, or even within it. For example, from the same poem by Vincenzo Cardarelli are drawn the quote «Il Disinganno, / che si nutre di sottigliezze / acerrime e conclusive» (cited s.v. *acerrimo*) and the quote «Il più frettoloso figliolo / del Tempo, il Disinganno, / ...si nutre di sottigliezze / acerrime e conclusive» (cited s.vv. *disinganno* and *frettoloso*). In this case as well, the analysis of the quote strings alone is insufficient: the pipeline detects the quotation similarity, but only the human reader can discriminate between cases of duplicated quotations and those of merely similar passages; cf. «Se mi s'accostava un passo di più, ...l'infilavo addirittura, prima che avesse tempo d'accomodarmi me, il birbone», cited s.v. *accomodare* from Alessandro Manzoni's *Promessi sposi*, matched by the tool with the quote «Se mi s'accostava un passo di più..., l'infilavo addirittura, prima che avesse tempo d'accomodarmi me, il birbone» (s.v. *infilare*),

[3] I. Moutier (ed.), *Cronica di Giovanni Villani*, t. 3, Firenze, Sansone, 1845.

which is likewise drawn from Manzoni but from an earlier redaction of the *Promessi Sposi* (the so-called “Ventisettana”).

Note that the different phenomena we have presented can occur simultaneously: within clusters of duplicated quotations, we may find strings of varying length due to additional elements appearing at any point in the sentence, possible OCR errors, and readings that already diverge in the original printed text, whether drawn from the same source or from different sources.

Overall, as can be seen, the picture of the quotes identified by the pipeline as potentially duplicated quotations is extremely complex, above all because of the characteristics of the original resource: in order to be certain that one is truly dealing with replicated quotes, the judgment of a human expert would almost always be required.

7. Conclusion

The choice between a minimalist solution - where each quotation is considered only once (or a balanced number of times) - and an approach that preserves all occurrences entails both advantages and drawbacks. On the one hand, the noise generated by repeated quotations may be problematic when the corpus of quotations is used for the construction of lexicographic resources. On the other hand, from the perspective of representativeness and quantitative analysis, the recurrence of cited authors and lexical items is far from negligible.

To maximize the flexibility of the resource and broaden its potential applications, it is therefore preferable to keep both options available, allowing users to tailor corpus queries to their specific needs (cf. Biffi *et al*., 2026). In this perspective, the extreme variability observed in the data, ranging from purely technical noise to substantial structural variation in the reuse of textual material, makes the automatic and indiscriminate removal of duplicate quotations both impractical and scientifically undesirable. A “blind” deduplication strategy would entail a systematic risk of eliminating data that are essential to the integrity and interpretability of historical lexicographic resources.

By contrast, the methodology implemented in QRD goes beyond mere dataset cleaning, enabling a deeper and more systematic understanding of the resource itself. In particular, the analysis of similarity-based quotation clusters makes it possible to uncover editorial practices underlying the compilation of the GDLI, shedding light on recurrent patterns of textual reuse, selection, and adaptation over time.

More generally, our findings highlight the intrinsic complexity of quotation duplication in lexicographic data and show that it cannot be addressed through a single, univocal solution. At the same time, QRD effectively supports the identification of OCR-related errors in lexicographic quotations, encompassing both systematic patterns and isolated cases. These errors can subsequently be corrected both in the TEI XML representation of the GDLI and in the derived GDLI Quotations Corpus.

With respect to citation deduplication in corpora derived from lexicographic resources, we contend that it must be conceived primarily as a task of identification and clustering rather than as a straightforward elimination procedure. By explicitly encoding the membership of reused quotations within similarity clusters, our approach enables the adoption of task-specific strategies, ranging from the preservation of all quotation variants for philological and lexicographic analysis to the selection of representative instances for computational processing.

This flexible framework allows multiple requirements to be satisfied simultaneously: corpus quality is improved, editorial variation is preserved and made accessible for scholarly analysis, and - when necessary - linguistically informed deduplication strategies can be applied to support downstream computational tasks, such as the training or fine-tuning of linguistic annotation models.

Finally, the in-depth analysis of repeated quotations has made it possible to characterize the patterns underlying quotation reuse in the GDLI and to inform the design of clustering-based solutions for advanced dictionary consultation and exploration.

Looking ahead, several directions will be pursued to extend this work. First, the GDLI Quotations Corpus will be completed to cover all dictionary volumes, thereby enabling large-scale and fully diachronic analyses. Second, ongoing research is focusing on the systematic evaluation of the impact of duplicated quotations on the training and fine-tuning of linguistic annotation tools. Third, clustering-based methods will be further developed to support advanced functionalities in the structured digital dictionary, including the automatic detection, navigation, and exploration of reused quotations. Together, these directions aim to consolidate QRD as both a methodological framework and an operational tool for the analysis and reuse of historical lexicographic data.

In this perspective, deduplication is no longer a data-cleaning operation, but an analytical methodology for understanding and exploiting historically and editorially curated textual resources.

8. Acknowledgments

This contribution forms part of the ongoing digitization and structuring efforts concerning the GDLI, carried out since 2018 by the “Istituto di

Linguistica Computazionale" of CNR in collaboration with the "Accademia della Crusca". The current implementation of the GDLI-QC constitutes a primary objective of the "GDLIplus" project, funded under the Regione Toscana FSE+ Programme 2021–2027 (Action 4.a.5) and co-financed by the "Accademia della Crusca" (CUP B53C24004090007).

9. Bibliographical References

Atkins, B.T.S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Béjoint, H. (2010). *The Lexicography of English. From Origins to Present.* New York, Oxford University Press.

Biffi, M., Favaro, M., Guadagnini, E., Montemagni, S., Sassolini, E. (2026). Gli interventi redazionali negli esempi citati nelle voci del «Grande Dizionario Della Lingua Italiana». *Studi di Lessicografia italiana*, XLIII.

Biffi, M., Guadagnini, E., Montemagni, S., Sassolini, E. (2025). La stampa periodica citata nel GDLI: il rapporto tra voci e indice bibliografico e le prospettive per il dizionario strutturato. *Studi di Lessicografia Italiana*, XLII, pp. 267-94

Biffi, M., Guadagnini, E., Montemagni, S., Sassolini, E. (2023). Il lemmario del «GDLI»: dati quantitativi e prime osservazioni. *Studi di Lessicografia Italiana*, XL, pp. 331-51

Bollmann, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp. 3885–3898. Association for Computational Linguistics.

Broder, A. (1997). On the Resemblance and Containment of Documents. In *Proceedings of the International Conference on Compression and Complexity of Sequences*, Salerno, Italy, pp. 21-29.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science and Business Media, Berlin.

Elmagarmid, A.K., Ipeirotis, P.G. and Verykios, V.S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 1-16, Jan. 2007.

Favaro, M., Guadagnini, E., Sassolini, E., Biffi, M., Montemagni, S. (2022). Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, pp. 94–100, Marseille, France. European Language Resources Association (ELRA).

Hoffmann, S. (2004). Using the OED quotations database as a corpus: A linguistic appraisal. *ICAME Journal*, 28, pp. 17-30.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C. and Carlini, N. (2022). Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pp. 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, 33(1):31–88.

Palmerio Aprosio, A., Menini, S., and Tonelli, S. (2022). The "Bertoldo" Corpus: An Annotated Corpus of 17th-Century Italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, pages 78–82, Marseille, France. European Language Resources Association (ELRA).

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, San Raphael, California.

Rohdenburg, G. (2013). Using the OED quotations database as a diachronic corpus. In Krug M. *et al.* (Eds.), *Research Methods in Language Variation and Change*, Cambridge University Press, pp 136-157.

Sassolini, E. *et al.* (in preparation). Historical Dictionaries as Internally Hybrid Resources: The Case of the "Grande Dizionario della Lingua Italiana".

Sassolini, E., Biffi, M., De Blasi, F., Guadagnini, E. e Montemagni, S. (2021). La digitalizzazione del GDLI: un approccio linguistico per la corretta acquisizione del testo? In *Proceedings di AIUCD 2021: DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale*. Raccolta degli abstract estesi della 10a conferenza nazionale, pp. 159-166.

Sassolini, E., Biffi, M. (2020). Strategie e metodi per il recupero di dizionari storici. In *Proceedings of the AIUCD 2020 Conference*, 15-17 gennaio

2020, Università Cattolica del Sacro Cuore. Milano, pp. 235-239.

Sassolini, E., Khan, A. F., Biffi, M., Monachini, M., Montemagni, S. (2019). Converting and structuring a Digital Historical Dictionary of Italian: a case study. Electronic lexicography in the 21st century. In *Proceedings of the eLex 2019 conference*, 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o. Eds., pp. 603-621.