From Print to Digital and Beyond: The Retrodigitization of a Historical Dictionary of Italian as a Hybrid Lexical Resource

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

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

Abstract

This paper presents the retrodigitization project of the *Grande Dizionario della Lingua Italiana* (GDLI), the largest historical dictionary of the Italian language. The GDLI's 23,000 pages - originally designed for human consultation - constitute an exceptional repository of linguistic and cultural-historical information, while posing significant challenges to large-scale digitization and data structuring. The project, still ongoing, will result in the development of a set of interoperable and interlinked resources: (i) a TEI-XML edition of the dictionary text, encoding its complex lexicographic structure; (ii) an annotated corpus of the quoted examples, enabling linguistic and historical research across centuries; and (iii) a database of quoted authors and works. Together, these components form a hybrid lexical resource that establishes the foundations for innovative and advanced modes of accessing and exploring the rich and multifaceted content of this historical dictionary.

Keywords: Historical Dictionary, Retro-digitization, Knowledge Organization, e-Lexicography

1. 1. Introduction

Over the last decade, research in e-lexicography has increasingly focused on the design of human-oriented online dictionaries that enable efficient, multidimensional access for diverse user groups and can be seamlessly integrated with other linguistic resources. Initiatives such as the *European Network of e-Lexicography* (ENeL, 2012–2017)[1], followed by the *ELEXIS* European project (2018–2022; Krek et al., 2018) and the subsequent creation of the *Elexis Association*[2], testify to the community's strong interest in digitizing existing dictionaries, standardizing their formats to ensure interoperability, and enriching them through the use of advanced language technologies and tools. The main challenges to be addressed range from facilitating user access to scholarly dictionaries and bridging the gap between the general public and academic lexicography, to promoting a broader and more systematic exchange of expertise, standards, and solutions.

The paper reports on the ongoing retro-digitization and structuring of one of the most authoritative historical dictionaries of the Italian language, the *Grande Dizionario della Lingua Italiana* ('Great Dictionary of Italian Language', in short GDLI). The project aims to establish the foundations for advanced human-centred querying of multiple data types - lexicographic, textual, and encyclopaedic - while ensuring interoperability with existing standards and lexicographic infrastructures. The work is carried out in collaboration with the *Accademia della Crusca*, one of world's oldest linguistic academies and the main authority on the Italian language, which has acquired the dictionary thanks to an agreement with the publisher UTET Grandi Opere.

The GDLI, edited initially by Salvatore Battaglia and later by Giorgio Barberi Squarotti, is the most comprehensive historical dictionary of the Italian language. Published by UTET in 21 volumes between 1961 and 2002, with two supplementary volumes in 2004 and 2009, it documents the evolution of Italian from its origins in the 10th century to contemporary usage. Conceived as an update to Tommaseo and Bellini's *Dizionario della lingua italiana* (1861–1879) and, indirectly, as a continuation of the *Vocabolario degli Accademici della Crusca*, the GDLI retains and expands the tradition of quotation-based lexicography. Each entry is supported by a rich and chronologically broad set of quotations, encompassing both canonical and lesser-known authors, and covering a variety of textual genres—from literature (which constitutes the majority of the quotes cited) to letters, treatises, translations, and even legal and administrative documents.

Today, a digital version of the GDLI is available through the *Scaffali Digitali* (lit. 'digital shelves') of the *Accademia della Crusca*[3]. This version includes all volumes in an unstructured textual form, produced via an OCR-based workflow (ABBYY FineReader) with minimal cleaning. Despite its preliminary state, the resource allows full-text string searches with direct access to facsimile pages (for further details see Biffi 2024). The work presented in this paper aims to create a

[1] https://www.cost.eu/actions/IS1305/
[2] https://elex.is/
[3] https://www.gdli.it/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 770–777
11–16 May 2026.

new, fully enhanced digital version of the GDLI, accessible to both scholars and the general public, and aligned with the standards of contemporary e-lexicography.

The paper describes the approach and technical framework adopted for the digitization and structuring of GDLI contents, highlighting the main challenges encountered, the solutions implemented, and the encoding strategies applied using internationally recognized standards. In what follows, we focus on:

1. The segmentation of GDLI entries and their internal structuring from OCRed text, followed by the conversion into an XML TEI Lex0 format, with soft markup;
2. The linguistic annotation of a representative sample of quotations, including both prose and poetry, as a preliminary step toward automatic annotation of the entire corpus of examples;
3. The segmentation and structuring of the *Index of Cited Authors*, represented in XML-TEI format.

Taken together, these results make the digital version of the dictionary under development an inherently hybrid resource, laying the groundwork for advanced methods of querying and exploring its complex and multidimensional contents. The structured data model will not only move beyond the traditional alphabetical organization of entries but also enable new ways of accessing the dense network of lexical and textual relations made explicit through structured encoding.

The paper is organized as follows: Section 2 presents the segmentation and internal structuring of GDLI entries; Section 3 describes the linguistic annotation of a sample of GDLI quotations; Section 4 illustrates the structuring of the index of sources; Section 5 discusses the integration of these components into a hybrid lexical resource; and Section 6 outlines current and future research directions.

2. Segmentation and Structuring of GDLI Entries

Dictionary segmentation and structuring were carried out implementing an automated parsing tool based on recognition rules targeting both formal layout features and expected structural sequences. The input data consisted of the OCR-derived version of the text (see above). Starting from the identification of entry boundaries, the process progressively reconstructed and marked the internal organization of each entry through a series of iterative refinement phases, interspersed with targeted manual correction campaigns on specific portions of the text.

A machine learning–based approach, such as that adopted in GROBID-Dictionaries (Khemakhem et al., 2017), proved unsuitable in this case due to the structural complexity of the source text and the presence of OCR-generated errors. These inconsistencies were not confined to spelling errors but also involved the correct division into paragraphs and typographical style. A particularly significant case concerns the quotations field: in the original layout, the beginning of this field differs from the preceding definition only by a reduced font size and the use of italics, which mark the string identifying the author and/or work. In the absence of this combined typographical marking, no explicit structural cues are available to support automatic recognition of quotations.

To handle such cases, a set of heuristics was developed based on pattern-matching strategies that consider a broader textual context and, building on previously acquired information, test and validate interpretative hypotheses. Furthermore, as previously noted, in addition to issues arising from various types of errors, the intrinsic complexity of the data must also be taken into account. The information contained in the entries does not always follow a strictly defined structure, as it is often organized to enhance readability and comprehension for human users. This characteristic poses a significant challenge for automatic information extraction systems. An example is shown in Figure 1, where the entry covers two distinct lemmas (a noun and the adjective derived from it).

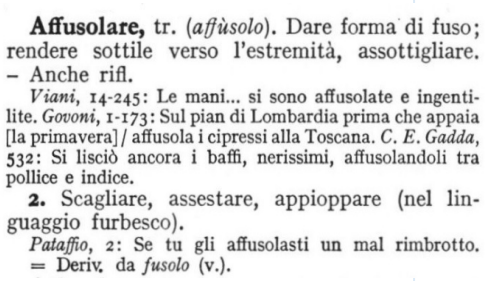

Antimaterialista, sm. e f. (plur. m. *-i*) e **antimaterialistico,** agg. (plur. m. *-ci*). Che è contrario alla considerazione materialistica dell'essere.
B. Croce, II-2-80: Altamente speculativo il suo metodo [del Vico] e disdegnoso dell'empirismo; idealistico, e perciò antimaterialistico e antiutilitaristico, il suo spirito.
= Comp. da *anti-* ' contro ' e *materialistico* (v.).

Figure 1: The GDLI entry for *antimaterialista* – *antimaterialistico* ('anti-materialistic').

The overall methodology - described in detail in Sassolini et al. (2019, 2021) and Biffi and Sassolini (2020) - resulted in a comprehensive segmentation of the data, with limited residual gaps, which were subsequently addressed through targeted manual interventions. A further source of complexity in implementing automatic structuring for the GDLI stems from the nature of the dictionary itself. As a monumental project developed over approximately forty years, the GDLI reflects editorial practices (metalexicographic conventions) that are not always uniform. Moreover, its design for exclusive use by human readers and in particular historical linguist - who are able to interpret and connect complex, distributed information - has made automated text structuring particularly demanding.

The inevitable inconsistencies present in the printed dictionary, often scarcely noticeable to human readers, however, have significant consequences for computational processing. Addressing these challenges has required continuous refinement and adaptation of both segmentation and structuring strategies.

The final objective of this phase is to transform the dictionary content into a set of macrostructural and microstructural components encoded according to a standardized format. Among the available standards for representing lexicographic resources, the Text Encoding Initiative (TEI) was selected, specifically its customization TEI Lex-0 (Tasovac et al., 2018), currently the most widely adopted framework for dictionary encoding. TEI Lex-0 provides a flexible yet interoperable model, capable of capturing not only the printed structure of the source text but also its conceptual, bibliographic, and linguistic dimensions. Nevertheless, the specific features of the GDLI pose significant challenges to the standard. These challenges stem both from the dictionary's complex structural and organizational features and from the need to generate the encoding automatically and systematically.



Affusolare, tr. (*affùsolo*). Dare forma di fuso; rendere sottile verso l'estremità, assottigliare. – Anche rifl.

Viani, 14-245: Le mani... si sono affusolate e ingentilite. *Govoni*, 1-173: Sul pian di Lombardia prima che appaia [la primavera] / affusola i cipressi alla Toscana. *C. E. Gadda*, 532: Si lisciò ancora i baffi, nerissimi, affusolandoli tra pollice e indice.

2. Scagliare, assestare, appioppare (nel linguaggio furbesco).

Pataffio, 2: Se tu gli affusolasti un mal rimbrotto. = Deriv. da *fusolo* (v.).

```
<entry>
    <form type="lemma">
        <orth>Affusolare</orth>
    </form>
    <sense level="1" n="1">
        <gramGrp>
            <gram type="POS">verbo</gram>
            <gram type="subc">tr.</gram>
            <gram type="subc">rifl.</gram>
        </gramGrp>
        <def>tr. (affùsolo) Dare forma di fuso; rendere sottile
        verso l'estremità, assottigliare. - Anche rifl.</def>
        <cit>
            <bibl>Viani, 14-245:</bibl>
            <quote> Le mani... si sono affusolate e ingentilite.
            </quote>
            <bibl>Govoni, 1-173:</bibl>
            <quote> Sul pian di Lombardia prima che appaia [la
            primavera] / affusola i cipressi alla Toscana.</quote>
            <bibl>C. E. Gadda, 532:</bibl>
            <quote> Si lisciò ancora i baffi, nerissimi,
            affusolandoli tra pollice e indice.</quote>
        </cit>
    </sense>
    <sense level="1" n="2">
        <def>2. Scagliare, assestare, appioppare (nel linguaggio
        furbesco).</def>
        <cit>
            <bibl>Pataffio, 2:</bibl>
            <quote> Se tu gli affusolasti un mal rimbrotto.
            </quote>
        </cit>
    </sense>
    <etym>= Deriv. da fusolo (v.).</etym>
</entry>
```

Figure 2: The GDLI entry for *affusolare* ('to streamline') and its TEI Lex0 representation

To illustrate the conversion process, Fig. 2 presents the original GDLI entry alongside its automatically generated XML TEI Lex0 counterpart for the lemma *affusolare* 'to streamline'. Lemma information is encoded using the TEI `<form>` element, with the attribute `@type` set to "lemma". The entry contains two senses - represented at the first level of nesting - each marked by a `<sense>` element and the related attributes `@level` and `@n`. Each sense includes quotations drawn from the corpus of historical Italian texts referenced by the GDLI. For each sense, the set of quotations is annotated using the `<cit>` element, which contains a `<bibl>/<quote>` pair encoding, respectively, a loosely structured bibliographic reference and the quotation text. Finally, the etymological information is encoded within the `<etym>` element.

At the time of writing, the status of the conversion process is as follows. The complete list of entries, consisting of 213,519 lemmas, has been successfully reconstructed from the 21 volumes of the GDLI. This represents a significant milestone: for the first time, the entire set of lemmas can be both quantified and individually identified (Biffi et al., 2023). Furthermore, for the first ten volumes, the internal microstructure of the entries has been manually validated for all main and internal definitions of each entry and further enriched

through the annotation of subentries and usage labels.

From the TEI Lex-0 XML version of the GDLI, a prototype resource was generated and organized into multiple interconnected modules, including a database specifically designed to reflect the internal structure of dictionary entries. This multidimensional architecture supports the complex relationships among dictionary fields while preserving the richness and granularity of the printed edition (Sassolini et al., 2025).

The implementation of this resource is integrated with a dedicated query system that currently enables structured searches across macro-fields, corresponding to the lemma, the definition (currently integrated with other information types, such as grammatical features), and the set of quotations (each articulated into source and example). At this initial stage of structural validation, the query system has been designed as modular and extensible, allowing for the iterative refinement of both data modeling and query functionalities. Owing to its relational architecture, the query system also functions as a tool for evaluating the quality and reliability of the results delivered to users. A targeted testing campaign is currently supporting the progressive refinement of both extraction and retrieval procedures. Once the extraction process has reached a stable stage with satisfactory levels of accuracy, the resource and its query interface will be made publicly available online.

3. Extracting and Annotating the GDLI Quotation Corpus

Quotations represent the “bedrock” of any historical dictionary (Hawke, 2016), as also testified by studies carried out on quotation databases (see e.g. Hoffman, 2004; Rohdenburg, 2013), which demonstrate how they can be used as a valuable information source for different typologies of studies.

Within the large set of GDLI entries, each definition and subdefinition is accompanied by a rich set of quotations covering the widest possible chronological span and arranged from the oldest to the most recent ones. We estimate that the entire GDLI quotation collection comprises between 2 and 2.5 million examples, corresponding to a corpus of roughly 40 million tokens.

Each quotation preserves the syntactic autonomy of the original passage or restores its overall meaning, often extending beyond sentence boundaries. As such, GDLI entries can be viewed as small anthologies of authorial quotations illustrating the diachronic uses of each lexical item in Italian writing, particularly within (mostly) literary contexts. Combined with the extraordinary range of quoted authors and works (see Section 4), these features make the GDLI quotation set an exceptionally rich textual collection that can be exploited as a resource in its own right.

Given the history of Italian as a primarily written and literary language up to the twentieth century, the collection of all quotations found in GDLI entries - hereafter referred to as the GDLI *Quotations Corpus* (GDLI-QC) - can be regarded as a representative diachronic corpus of Italian (Biffi, 2018). Representativeness here refers to the corpus’s ability to provide evidence of word usage within the literary tradition of Italian, as documented by transmitted texts (often mediated by earlier dictionaries) (Burgassi and Guadagnini, 2017; Kabatek, 2013). It should be noted, however, that GDLI quotations derive almost exclusively from printed sources, which are not always modern critical editions: early or pre-normative texts may thus reflect nineteenth-century editorial interventions affecting spelling or morphology.

Favaro et al. (2022) describe the initial steps toward creating a linguistically annotated version of the diachronic GDLI-QC, adopting the Universal Dependencies (UD; De Marneffe et al., 2021) standard. This first phase involved:

1. developing an initial nucleus of manually revised linguistic annotation (POS-tagging and lemmatization) representative of the overall corpus;
2. retraining POS-tagging and lemmatization components to reliably process the diachronic language varieties attested in GDLI-QC.

The first step consisted of the design and development of a manually revised GDLI quotation corpus, POS-tagged and lemmatized, including approximately 45,000 tokens spanning from the 14th to the 20th century. The most frequently cited authors in the dictionary were selected (see Biffi and Guadagnini, 2022), prioritizing those whose works cover the broadest chronological range. These writers represent key milestones in both Italian literature and the history of the Italian language, and include, among others, Dante, Boccaccio, Petrarca, Ariosto, and Manzoni. Their linguistic diversity, shaped by diachronic as well as stylistic factors, makes the corpus highly valuable for testing and retraining linguistic annotation tools.

The annotated GDLI quotations corpus is organized into two balanced subcorpora: one comprising 1,500 prose quotations from 15 authors (100 each), and the other containing 500 poetry quotations from 10 authors (50 each). The prose subcorpus (GDLI-QC_prose) thus includes approximately 35,000 tokens, while the poetry subcorpus (GDLI-QC_poetry) comprises about

10,000 tokens. GDLI-QC is currently available through the CLARIN-IT repository[4].

The second step consisted of retraining the Stanza annotation tool (Qi et al., 2020), using the biggest UD treebank for contemporary Italian (ISDT, Bosco et al., 2013) in combination with corpora representative of historical varieties of Italian, including portions of GDLI-QC. For the evaluation of both POS-tagging and lemmatization, 5-fold cross validation was used: retrained tools achieved high accuracy: 97% for prose and 94% for poetry. Further improvements for what concerns lemmatization are reported by Alzetta and Montemagni (2025).

The next step will be automatically annotating the whole set of quotations in the GDLI. We are currently working in this direction.

4. Structuring the Index of the Sources

GDLI also includes an additional volume containing the index of quoted authors and works. Note that the *Indice* deliberately does not include sources originating from periodical publications; a first survey of these sources is reported in Biffi et al. (2025).

After the manual revision of the OCRed text, the content was encoded in XML. The hierarchical structure comprises bibliographic items (`<item>`), cross-references (`<re>`), and metadata fields for names, dates, and editions. Temporal information was normalized under a `<date>` tag using standardized numerical ranges, while bibliographic references were encoded within `<bibl>`.

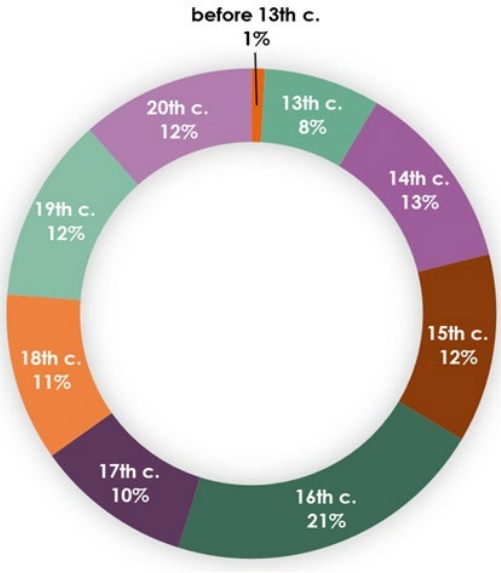


Figure 3: Distribution of Quoted Authors by Century.

As reported in Biffi and Guadagnini (2022), the range of cited authors and works is remarkable: GDLI quotes 6,226 authors and 13,848 sources. Fig. 3 shows the chronological distribution of quoted authors.

Although the data is structured through XML encoding - making it possible to process the content and generate general statistical information previously unavailable - automatic linking between the list of authors/works and the dictionary entries remains problematic at the moment since the strings used to identify authors/works in the index may differ from those appearing in the examples. Moreover, as pointed out above, certain authors/works as well as periodical publications are not included in the index at all. Strategies to enable such linking are currently under development, and further experimentation will determine the degree of precision that can be achieved.

5. Towards a Hybrid Resource

Among the major changes brought about by the integration of computer technology into lexicography, Granger (2012) highlights corpus integration and hybridization. One of the most notable outcomes of this electronic revolution is the gradual blurring of boundaries between different types of language resources, including dictionaries, encyclopaedias, term banks, lexical databases, and annotated corpora. Hartmann (2005) describes this trend as hybridization, defining it as the "combination of one or more types of reference work in a single product."

Hybrid dictionary genres—such as dictionary-cum-encyclopaedia or dictionary-cum-grammar—are increasingly emerging to meet new user needs. Despite their growing presence, these blended forms have received limited scholarly attention, and research on hybrid dictionaries remains sparse, focusing mainly on general-purpose dictionaries.

To our knowledge, this notion has not yet been applied to historical dictionaries. Although not explicitly framed in these terms, the following two examples can be considered special cases of hybridization applied to historical dictionaries.

The first is the integration of the *Oxford English Dictionary* (OED), the accepted authority on the history of English, with the *Historical Thesaurus of the OED* (HTOED), which reorganizes OED entries by meaning. In HTOED, words are classified under three main semantic categories (external world, mind, and society) and further subdivided into nested subcategories with dates of first usage and historical evidence. The online integration of HTOED and OED enables efficient electronic access, allowing users to navigate

[4] http://hdl.handle.net/20.500.11752/ILC-984

semantically related words quickly via direct links from dictionary entries.

The second example is provided by Stein (2020), who presents a case of historical dictionary hybridization in the *Altfranzösisches Wörterbuch* (Tobler and Lommatzsch, 1925ff), achieved by linking sense descriptions to a standardized conceptual hierarchy (GermaNet) and to the English WordNet. This integration allows conceptual access to the dictionary, enabling retrieval of synonyms, related meanings, hyponyms, and semantic classes.

In these cases, hybridization involves external resources. In contrast, we propose the notion of "internal hybridization." The retrodigitization of GDLI contributes to this development, creating an ecosystem of integrated yet heterogeneous linguistic resources, including the dictionary, the annotated GDLI quotations corpus, and the index of sources. The development of an integrated management and querying system for the extracted resources is still in progress. Particular attention is being devoted to the query system designed for the complex set of quotations. At present, each resource is supported by a separate query interface that enables searches across the main dictionary fields, queries within the Index of Quoted Authors, and text analysis functionalities applied to the quotations corpus. Future developments will focus on integrating these resources and enabling and powering advanced queries on the data. By explicitly linking the examples cited in the dictionary entries to the bibliographic data of their sources through the structured version of the Index of Quoted Authors, it will be possible, for instance, to produce historically oriented visualizations of the dictionary's contents. The explicit connection between the quoted examples and their linguistically annotated counterparts will allow the corpus to be queried not only as raw text but also as a structured resource searchable by abstract linguistic features such as lemma, part of speech, and their combinations. Note, however, that both the digitally structured dictionary and the annotated quotations corpus will also maintain their status as independent resources.

6. Conclusion

We illustrated the retrodigitization project of the most important historical dictionary of Italian, the GDLI, covering the entire chronological span of the language, from its origins in the 10th century to the present day. The project, still ongoing, has already produced promising and original results.

First, on the methodological side, the entire process of structuring the dictionary and extracting resources from the original data represents, to our knowledge, a unique achievement in the scientific landscape for comparable resources, encompassing all the challenges, uncertainties, and responsibilities inherent in such frontier work.

Second, the project has led to the design and development of three distinct resources:

- the XML TEI Lex0 version of the entire dictionary, with structuring work continuing progressively.
- a gold-standard annotated corpus of dictionary quotations, POS-tagged and lemmatized, intended for retraining and testing annotation tools specialized for historical varieties of Italian;
- the index of quotation sources.

The design and representation of these resources allow them to function both as independent entities and as an integrated ecosystem, enabling queries that combine information across the different components.

From a broader perspective, the project has significant implications both for the methodological advancement of historical-linguistic disciplines and for the specific results, which extend range and volume of data which can be explored systematically, accurately, and effectively. Last but not least, the outcomes are particularly relevant for the Accademia della Crusca, given its dual role as a research institution and a body committed to public dissemination. In this way, the project can reach a broader audience, including non-specialists, with particular attention to the education sector (on this aspect see Biffi et alli 2022).

7. Acknowledgements

This work was supported by the project "Una nuova risorsa per la storia dell'italiano: il corpus degli esempi citati nel «Grande dizionario della lingua italiana» (GDLI)" ('A New Resource for the History of Italian: The Corpus of Citations in the *Grande Dizionario della Lingua Italiana'*, GDLIplus) funded by Regione Toscana (PROGETTI FSE+ 2021-2027) with the financial support of Accademia della Crusca

8. Bibliographical References

Alzetta C. and Montemagni S. (2025). Low- vs High-level Lemmatization for Historical Languages. a Case Study on Italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, 24-26 September 2025, Cagliari, Italy.

Biffi, M. (2018). Tra fiorentino aureo e fiorentino cinquecentesco. Per uno studio della lingua dei lessicografi. In G. Belloni and P. Trovato (Eds.), *La Crusca e i testi. Lessicografia, tecniche editoriali e collezionismo librario intorno al Vocabolario del 1612*, Padova, libreriauniversitaria.it edizioni, pp. 543-560.

Biffi, M. (2024). Il GDLI in rete. In L. Ambrogio and M. Bardi (Eds.), *Il Grande Dizionario della Lingua Italiana Utet: un monu-mento aperto al futuro*, Alessandria, Edizioni dell'Orso, pages 51-67.

Biffi, M. and Sassolini, E. (2020). Strategie e metodi per il recupero di dizionari storici. In Marras, C., Passarotti, M., Franzini, G. and Litta, E. (Eds.), *La svolta inevitabile: sfide e prospettive per l'informatica umanistica. Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale* (AIUCD, 15-17 gennaio 2020), Milan: Associazione per l'Informatica Umanistica e la Cultura Digitale, pages 235-239.

Biffi, M. and Guadagnini, E. (2022). Le citazioni riconducono il dizionario nell'ambito della letteratura e della vita»: un primo sguardo d'insieme sui citati del GDLI. *Studi di Lessicografia Italiana*, vol. XXXIX, pages 351-386.

Biffi, M., De Blasi, F., Favaro, M., Guadagnini, E., Montemagni, and S., Sassolini, E. (2022). Parole in rete / reti di parole. Possibili impieghi didattici dei grandi vocabolari storici digitalizzati. *Italiano a scuola*, 4, pages 143-188.

Biffi, M., Guadagnini, E., Montemagni, S., and Sassolini, E. (2023), Il lemmario del «GDLI»: dati quantitativi e prime osservazioni, *Studi di Lessicografia Italiana*, vol. 40, pages 331-351.

Biffi, M., Guadagnini, E., Montemagni, S., and 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*, vol. XLII, pages 267-294.

Bosco, C., Montemagni, S., and Simi, M. (2013). Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the "7th Linguistic Annotation Workshop & Interoperability with Discourse"* (August 8-9, 2013), pages 61-69, Sofia, Bulgaria, August. Association for Computation Linguistics (ACL).

Burgassi C. and Guadagnini E. (2017). *La tradizione delle parole. Sondaggi di lessicologia storica*, Strasbourg, ÉLiPhi.

De Marneffe M. C., Manning C. D., Nivre J. and Zeman D. (2021), Universal Dependencies. *Computational Linguistics*, 47(2), pages 255-308.

Favaro M., Guadagnini E., Sassolini E., Biffi M. and Montemagni S., (2022). Toward the Creation of a Diachronic Corpus for Italian: a Case Study on the GDLI. In R. Sprugnoli et alii (Eds.), *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022, Association for Computational Linguistics (ACL), pages 94-100.

Granger, S. (2012). Electronic lexicography: From challenge to opportunity. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*, Oxford University Press, pages1-11.

Hartmann, R. (2005). Pure Or Hybrid? The Development Of Mixed Dictionary Genres. *Facta Universitatis*, Series: Linguistics and Literature, 3 (2), pages 193-208.

Hawke, A. (2016), Quotation Evidence and Definitions. In Durkin P. (Ed.), *The Oxford Handbook of Lexicography*, Oxford University Press, pages 176-202.

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

Kabatek J. (2013). ¿Es posible una lingüística histórica basada en un corpus representativo?, *Iberoromania*, 77, pages 8-28.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B., Tiberius, C. and Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, pages 881–891.

Khemakhem, M., Foppiano, L. and Romary, L. (2017). Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. In I. Kosem et al. (Eds.) *Proceedings of eLex 2017*, September 2017, Leiden, Netherlands. Brno, Lexical Computing.

Qi P., Zhang I., Zhang Y., Bolton J., Manning C. D. (2020). 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*, ACL 2020, July 5-10, pages 101-108, Online, July. Association for Computational Linguistics (ACL).

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, pages 136-157.

Sassolini, E., Fahad Khan, A., Biffi, M., Monachini, M. and Montemagni S. (2019). Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study. In Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (Eds.), *Electronic lexicography in the 21st century: Smart lexicography.*

Proceedings of the eLex 2019 conference (1-3 October 2019, Sintra, Portugal), Brno: Lexical Computing CZ, pages 603-621.

Sassolini, E., Biffi, M., De Blasi, F., Guadagnini, E., and Montemagni, S. (2021). La digitalizzazione del GDLI: un approccio linguistico per la corretta acquisizione del testo? In Boschetti, F., Del Grosso A. M. and Salvatori E. (Eds.), *AIUCD 2021 - DHs for society: e-quality, participation, rights and values in the Digital Age*. Book of extended abstracts of the 10th national conference, Pisa, Associazione per l'Informatica Umanistica e la Cultura Digitale, pages 159-166.

Sassolini, E. Cucurullo, S. and Biffi, Marco (2025). Exploring the GDLI: a Multidimensional Approach to Historical Lexicography. In *Proceedings of the 8th IEEE CiSt'25 Congress*, 6-8 October 2025, Marrakech.

Stein, A. (2020). Preserving Semantic Information from Old Dictionaries: Linking Senses of the *Altfranzösisches Wörterbuch* to WordNet. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, May 2020, pages 3063-3068.

Tasovac, T., Romary, L., and Salgado, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. ELEXIS - European Lexicographic Infrastructure. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.