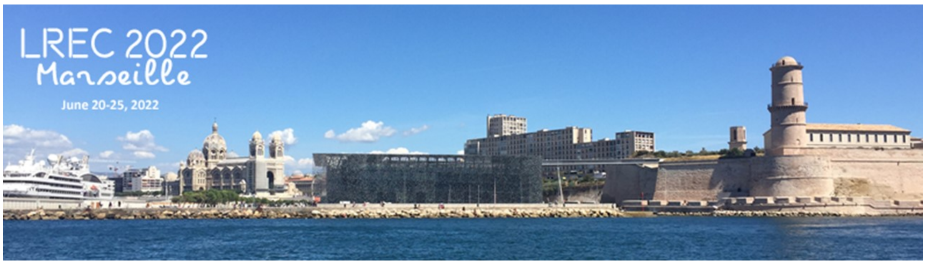


LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)

PROCEEDINGS

Editors: Rachele Sprugnoli and Marco Passarotti

Proceedings of the LREC 2022 Second Workshop on Language Technologies for Historical and Ancient Languages LT4HALA 2022

Edited by: Rachele Sprugnoli and Marco Passarotti

ISBN: 979-10-95546-78-8
EAN: 9791095546788

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

© European Language Resources Association (ELRA)

These workshop proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License

Preface

These proceedings include the papers accepted for presentation at the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022).[1] The workshop was held on June 25th 2022 in Marseille, France, co-located with the 13th Edition of the Language Resources and Evaluation Conference (LREC).[2]

The workshop wants to provide a venue to discuss research works on a wide range of topics concerning the building, analysis, exploitation and distribution of collections of digitized texts written in historical and ancient languages, with a specific focus on the development and application of Language Technologies (LTs) for such purposes.

The topics of the workshop are strictly bound to the peculiar characteristics of textual data for historical and ancient languages, which set them apart from modern languages, with a significant impact on LTs. Among the topics covered by the workshop are issues about the digitization process of textual sources, like handling spelling variation, and detecting and correcting OCR errors. Also concerned are questions about the automatic processing of various layers of metalinguistic annotation, which are made complex by the sparsity and inconsistency of texts that present considerable orthographic variation, are sometimes incomplete and belong to a large spectrum of literary genres. Such issues raise problems of adaptation of Natural Language Processing (NLP) tools to address diachronic/diatopic/diastratic variation in texts, which requires to be properly evaluated.

The various LTs tasks related to the topics of LT4HALA require a strict collaboration between scholars from different disciplinary areas. In such respect, the objective of the LT4HALA workshop series is to foster cross-fertilization between the Computational Linguistics community and the areas in the Humanities dealing with historical linguistic data, e.g. historians, philologists, linguists, archaeologists and literary scholars. Such a wide and diverse range of disciplines and scholars involved in the development and use of LTs for historical and ancient languages is mirrored by the large set of topics covered by the papers published in these proceedings, including the creation of annotated corpora and advanced computational lexical resources for historical languages, the development of models for performing various NLP tasks, the application of machine translation and linguistic analyses based on the empirical evidence provided by textual resources.

As large as the number of topics discussed in the papers is that of the either ancient/dead languages or the historical varieties of modern/living ones concerned. In total, the languages tackled in the proceedings are the following: Latin, Italian, Japanese, Chinese, Hungarian, French, Spanish, German, Portuguese, Dutch, Vedic Sanskrit, Ancient Greek (and Cypro-Greek), Ancient Hebrew, Maya, Umbrian and a set of languages of ancient Italy, namely Oscan, Faliscan, Celtic and Venetic.

In the call for papers, we invited to submit proposals of different types, such as experimental papers, reproduction papers, resource papers, position papers and survey papers. We asked both for long and short papers describing original and unpublished work. We defined as suitable long papers (up to 8 pages, plus references) those that describe substantial completed research and/or report on the development of new methodologies. Short papers (up to 4 pages, plus references) were instead more appropriate for reporting on works in progress or for describing a singular tool or project. We encouraged the authors of papers reporting experimental results to make their results reproducible and the entire process of analysis replicable, by distributing the data and the tools they used. Like for LREC, the submission process was single-blind. Each paper was reviewed but three independent reviewers from a program committee made of 24 scholars (12 women and 12 men) from 16 countries. In total, we received 24 submissions from 56 authors from institutions located in 10 countries: Italy (24 authors), Japan (7 authors), Switzerland (6 authors), Germany (5 authors), United States (4 authors), Belgium (3 authors), France (3 authors), Sweden (3 authors), Denmark (1 author), Spain (1 author). After the reviewing process, we accepted 18 submissions, leading to an acceptance rate of 75%.

[1] https://circse.github.io/LT4HALA/2022/

[2] https://lrec2022.lrec-conf.org/en/

LT4HALA 2022 was also the venue of the second edition of EvaLatin, the campaign devoted to the evaluation of NLP tools for Latin.[3] EvaLatin was started in 2020 (co-located with the first edition of LT4HALA) considering the important role played by textual data and linguistic metadata in the study of historical and ancient languages, with a special focus on Latin due to its prominence among such languages, both for the size and for the degree of diversity of its texts. Running evaluation campaigns in such a scenario is essential to understand the level of accuracy of the NLP tools used to build and analyze resources featuring texts that show those peculiar characteristics mentioned above. The second edition of EvaLatin focussed on three shared tasks (i.e. Lemmatization, PoS Tagging, Morphological Features Tagging), each featuring three sub-tasks (i.e. Classical, Cross-Genre, Cross-Time). These sub-tasks were designed to measure the impact of genre and diachrony on NLP tools performances, a relevant aspect to keep in mind when dealing with the diachronic and diatopic diversity of Latin texts, which are spread across a time span of two millennia all over Europe. Participants were provided with shared data in the CoNLL-U format and all the necessary evaluation scripts. They were required to submit a technical report for each task (with all the related sub-tasks) they took part in. The maximum length of the reports was 4 pages (plus references). In total, 2 technical reports of EvaLatin, corresponding to as many participants, are included in these proceedings. All reports received a light review by the organizers to check the correctness of the format, the exactness of the results and ranking reported, as well as the overall exposition. The proceedings also feature a paper detailing some specific aspects of the second edition of EvaLatin, like dataset, annotation criteria and results of the shared tasks.

Besides EvaLatin, LT4HALA 2022 hosted also the first edition of EvaHan, an evaluation campaign of NLP tools for the Ancient Chinese language, organized by a team of scholars directed by Bin Li (School of Chinese Language and Literature, Nanjing Normal University), which includes Yiguo Yuan (Nanjing Normal University), Minxuan Feng (Nanjing Normal University), Chao Xu (Nanjing Normal University) and Dongbo Wang (Nanjing Agricultural University).[4] EvaHan focussed on one joint task of Word Segmentation and PoS Tagging. Test data of Ancient Chinese, which is dated back around 1000BC-221BC, were provided in raw format, featuring only Chinese characters and punctuation. The participants were provided with two sets of test data, to evaluate the accuracy rates of the systems respectively on data excerpted from the same work (the Zuozhuan book) included in the training set, without overlapping, and on data from another, yet similar, text. A pretrained model consisting in word embeddings built over a large corpus of traditional Chinese was provided as well. In total, 9 technical reports of EvaHan, corresponding to as many participants, are included in these proceedings. Like for Evalatin, all reports received a light review by the organizers of EvaHan and the proceedings include a short paper with the details of the campaign.

We are grateful to the organizers of EvaHan, who contributed to extend the range of historical and ancient languages of the LT4HALA workshop and showed how some NLP-related issues concern ancient and historical languages per se, despite their typological differences.

Rachele Sprugnoli
Marco Passarotti

[3] https://circse.github.io/LT4HALA/2022/EvaLatin

[4] https://circse.github.io/LT4HALA/2022/EvaHan

Organizers:

Rachele Sprugnoli, Università degli Studi di Parma (Italy)
Marco Passarotti, Università Cattolica del Sacro Cuore (Italy)

Program Committee:

Marcel Bollmann, University of Copenhagen (Denmark)
Gerlof Bouma, University of Gothenburg (Sweden)
Harry Diakoff, Alpheios Project (USA)
Stefanie Dipper, Ruhr-Universität Bochum (Germany)
Hanne Eckhoff, Oxford University (UK)
Margherita Fantoli, University of Leuven (Belgium)
Heidi Jauhiainen, University of Helsinki (Finland)
Neven Jovanovic, University of Zagreb (Croatia)
Timo Korkiakangas, University of Helsinki (Finland)
Bin Li, Nanjing Normal University (P.R. China)
Eleonora Litta, Università Cattolica del Sacro Cuore (Italy)
Chao-Lin Liu, National Chengchi University (Taiwan)
Barbara McGillivray, Turing Institute (UK)
Beáta Megyesi, Uppsala University (Sweden)
Giulia Pedonese, Università Cattolica del Sacro Cuore (Italy)
Saskia Peels, University of Groningen (The Netherlands)
Matteo Pellegrini, Università Cattolica del Sacro Cuore (Italy)
Eva Pettersson, Uppsala University (Sweden)
Sophie Prévost, Laboratoire Lattice (France)
Philippe Roelli, University of Zurich (Switzerland)
Matteo Romanello, Université de Lausanne (Switzerland)
Halim Sayoud, USTHB University (Algeria)
Dongbo Wang, Nanjing Agricultural University (P.R. China)

EvaLatin 2022 Organizers:

Rachele Sprugnoli, Università degli Studi di Parma (Italy)
Margherita Fantoli, KU Leuven (Belgium)
Flavio M. Cecchini, Università Cattolica del Sacro Cuore, Milan (Italy)
Marco Passarotti, Università Cattolica del Sacro Cuore, Milan (Italy)

EvaHan 2022 Organizers:

Bin Li, Nanjing Normal University (P.R. China)
Yiguo Yuan, Nanjing Normal University (P.R. China)
Minxuan Feng, Nanjing Normal University (P.R. China)
Chao Xu, Nanjing Normal University (P.R. China)
Dongbo Wang, Nanjing Agricultural University (P.R. China)

Table of Contents

Conference Program

Saturday, June 25, 2022

Long and Short Papers

Identifying Cleartext in Historical Ciphers
Maria-Elena Gambardella, Beata Megyesi and Eva Pettersson

Detecting Diachronic Syntactic Developments in Presence of Bias Terms
Oliver Hellwig and Sven Sellmer

Accurate Dependency Parsing and Tagging of Latin
Sebastian Nehrdich and Oliver Hellwig

Annotating "Absolute" Preverbs in the Homeric and Vedic Treebanks
Luca Brigada Villa, Erica Biagetti and Chiara Zanchi

CHJ-WLSP: Annotation of 'Word List by Semantic Principles' Labels for the Corpus of Historical Japanese
Masayuki Asahara, Nao Ikegami, Tai Suzuki, Taro Ichimura, Asuko Kondo, Sachi Kato and Makoto Yamazaki

The IKUVINA Treebank
Mathieu Dehouck

Machine Translation of 16Th Century Letters from Latin to German
Lukas Fischer, Patricia Scheurer, Raphael Schwitter and Martin Volk

A Treebank-based Approach to the Supprema Constructio in Dante's Latin Works
Flavio Massimiliano Cecchini and Giulia Pedonese

From Inscriptions to Lexica and Back: A Platform for Editing and Linking the Languages of Ancient Italy
Valeria Quochi, Andrea Bellandi, Fahad Khan, Michele Mallia, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi and Cesare Zavattari

BERToldo, the Historical BERT for Italian
Alessio Palmero Aprosio, Stefano Menini and Sara Tonelli

In Search of the Flocks: How to Perform Onomasiological Queries in an Ancient Greek Corpus?
Alek Keersmaekers and Toon Van Hal

Contextual Unsupervised Clustering of Signs for Ancient Writing Systems
Michele Corazza, Fabio Tamburini, Miguel ValÃ©rio and Silvia Ferrara

Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations
Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi and Simonetta Montemagni

Automatic Translation Alignment for Ancient Greek and Latin
Tariq Yousef, Chiara Palladino, David J. Wright and Monica Berti

Handling Stress in Finite-State Morphological Analyzers for Ancient Greek and Ancient Hebrew
Daniel Swanson and Francis Tyers

From Inscription to Semi-automatic Annotation of Maya Hieroglyphic Texts
Cristina Vertan and Christian Prager

Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and Bert-based Models.
Sergio Torres Aguilar

Linguistic Annotation of Neo-Latin Mathematical Texts: A Pilot-Study to Improve the Automatic Parsing of the Archimedes Latinus
Margherita Fantoli and Miryam de Lhoneux

Saturday, June 25, 2022 (continued)

EvaHan Technical Reports

The First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff: Overview of the EvaHan 2022 Evaluation Campaign
Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang QU and Dongbo Wang

Automatic Word Segmentation and Part-of-Speech Tagging of Ancient Chinese Based on BERT Model
Yu Chang, Peng Zhu, Chaoping Wang and Chaofan Wang

Ancient Chinese Word Segmentation and Part-of-Speech Tagging Using Data Augmentation
Yanzhi Tian and Yuhang Guo

BERT 4EVER@EvaHan 2022: Ancient Chinese Word Segmentation and Part-of-Speech Tagging Based on Adversarial Learning and Continual Pre-training
Hailin Zhang, Ziyu Yang, Yingwen Fu and Ruoyao Ding

Construction of Segmentation and Part of Speech Annotation Model in Ancient Chinese
Longjie Jiang, Qinyu C. Chang, Huyin H. Xie and Zhuying Z. Xia

Simple Tagging System with RoBERTa for Ancient Chinese
Binghao Tang, Boda Lin and Si Li

The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS
Pengyu Wang and Zhichen Ren

Data Augmentation for Low-resource Word Segmentation and POS Tagging of Ancient Chinese Texts
Yutong Shen, Jiahuan Li, Shujian Huang, Yi Zhou, Xiaopeng Xie and Qinxin Zhao

A Joint Framework for Ancient Chinese WS and POS Tagging Based on Adversarial Ensemble Learning
Shuxun Yang

Glyph Features Matter: A Multimodal Solution for EvaHan in LT4HALA2022
Wei Xinyuan, Liu Weihao, Qing Zong , Zhang Shao Qing and Baotian Hu

Saturday, June 25, 2022 (continued)

EvaLatin Technical Reports

Overview of the EvaLatin 2022 Evaluation Campaign
Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli and Giovanni Moretti

An ELECTRA Model for Latin Token Tagging Tasks
Wouter Mercelis and Alek Keersmaekers

Transformer-based Part-of-Speech Tagging and Lemmatization for Latin
Krzysztof WrÃ³bel and Krzysztof Nowak

Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 94–100
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC 4.0

Towards the Creation of a Diachronic Corpus for Italian: a Case Study on the GDLI Quotations

Manuel Favaro[§], Elisa Guadagnini[§], Eva Sassolini[§], Marco Biffi[*^], Simonetta Montemagni[§]
[§]Istituto di Linguistica Computazionale “A. Zampolli” – CNR
[*]Università di Firenze
[^]Accademia della Crusca
manuel.favaro@ilc.cnr.it, elisa.guadagnini@ilc.cnr.it, eva.sassolini@ilc.cnr.it, marco.biffi@unifi.it, simonetta.montemagni@ilc.cnr.it

Abstract

In this paper we describe some experiments related to a corpus derived from an authoritative historical Italian dictionary, namely the *Grande dizionario della lingua italiana* (‘Great Dictionary of Italian Language’, in short GDLI). Thanks to the digitization and structuring of this dictionary, we have been able to set up the first nucleus of a diachronic annotated corpus that selects—according to specific criteria, and distinguishing between prose and poetry—some of the quotations that within the entries illustrate the different definitions and sub-definitions. In fact, the GDLI presents a huge collection of quotations covering the entire history of the Italian language and thus ranging from the Middle Ages to the present day. The corpus was enriched with linguistic annotation and used to train and evaluate NLP models for POS tagging and lemmatization, with promising results.

Keywords: Diachronic Corpus, Adaptation of Annotation Tools, Historical Dictionaries

1. Introduction

Over the past decades, the number and variety of historical corpora available for different languages has been progressively growing. They represent an invaluable asset in the era of Digital Humanities, given the increasing interest in applying quantitative and computational methods to diachronic linguistics and historical text analysis.

For Italian, diachronic corpora are still few. Among them, covering a large timespan going from the origin of the Italian language to the present day, it is worth mentioning the *MIDIA corpus* (Gaeta et al., 2013, D’Achille and Grossmann, 2017), from which the *CODIT* was developed (Micheli, 2022), the *Letteratura italiana Zanichelli* (LIZ, later reissued as BIZ), and *BibIt*[1]. Other corpora focus on specific periods, such as the *Corpus OVI dell’Italiano antico* (Squillacioti, 2021) for Old Italian, the epistolary corpus *CEOD*[2] for 19th c. Italian, the *DiaCORIS corpus* (Onelli et al., 2006) and the reference corpus built for the construction of a *Dynamic Vocabulary of Modern Italian* (*VoDIM*, Marazzini and Maconi, 2018) for post-unitarian Italian. Many of these corpora have been enriched with linguistic annotation (typically, POS tagging and lemmatization), carried out (semi-)automatically or manually, and can be queried through advanced search tools. Yet, they are not distributed as linguistically annotated corpora: they were conceived as reference resources to be queried by scholars for the analysis of linguistic phenomena over the covered period of time and across the different varieties of language use testified (e.g. textual genres). Unfortunately, this feature makes them of limited use for the application of NLP-based methods with a specific view to the adaptation of linguistic annotation tools for the processing of historical varieties of language and for computational analyses focusing e.g. on semantics or style.

To the best of our knowledge, only two linguistically annotated corpora testifying historical varieties of language are available for Italian: the *Voci della Grande Guerra* corpus (VGG, Lenci et al., 2020) containing texts related to different varieties (both textual genres and registers) of Italian at the time of the World War I; and the corpus of the politician Alcide De Gasperi’s public documents (Tonelli et al., 2019), a multi-genre corpus spanning 50 years of European history, written or transcribed between 1901 and 1954. VGG and Alcide corpora are available as multi-level annotated corpora, with both silver and gold annotations, which are compliant to internationally recognized representation standards. In Alcide, gold annotation was used to assess the accuracy of lemmatization, POS tagging and named entity annotation which was performed with tools trained on contemporary language. In VGG, gold annotation was also used to specialize the annotation tools to deal with the challenges posed by the linguistic varieties subsumed in the corpus (De Felice et al., 2018): retrained models were then used to annotate the rest of the corpus.

In the general picture depicted above, the aim of this paper is twofold. First, it illustrates the preliminary steps towards the creation of a linguistically annotated diachronic corpus for Italian, whose time span goes from old to contemporary Italian. Second, it reports the results of experiments aimed at assessing the accuracy of linguistic annotation (lemmatization and POS tagging) carried out with specialized annotation models against a diachronically representative sample of the corpus (gathering texts both in prose and poetry, going from the 13th to the 20th century).

For the composition of the corpus, we decided to use an interesting diachronic textual collection, represented by the set of quotations in a historical dictionary of Italian, namely the *Grande dizionario della lingua italiana* (‘Great

[1] http://www.bibliotecaitaliana.it/

[2] http://ceod.unistrasi.it/

Dictionary of Italian Language', in short GDLI). Since quotations are seen as the "bedrock" of any historical dictionary (Hawke, 2016), we believe that they can be usefully exploited to build a wide coverage diachronic corpus. Studies carried out on quotations databases (see e.g. Hoffman, 2004; Rohdenburg, 2013) demonstrate how they can be used as a valuable information source for different typologies of studies, including quantitative ones.

The challenges of the linguistic annotation of historical texts are well known (Piotrowski, 2012). For Italian, an exploratory study on a diachronic corpus with texts (both prose and poetry) from the 13th to the 19th century focusing on morphological and morpho-syntactic annotation (Pennacchiotti and Zanzotto, 2008) highlights the specific issues (mostly, graphical, phonological, and morphological variability) connected with the automatic processing of Italian historical texts. More recently, adaptation experiments have been carried out to improve the performance of the automatic analysis tools by using manually revised sub-corpora to retrain the automatic linguistic annotation tools, with promising results. This is the case of De Felice et al. (2018) for the VGG Corpus and of Favaro et al. (2020) for a subset of the VoDIM corpus.

The paper is organized as follows. Section 2 describes the GDLI source with a specific view to the huge collection of quotations. Section 3 illustrates the selection criteria and the corpus composition of the first nucleus of the diachronic corpus. Sections 4 and 5 report the results of the annotation experiments carried out on the corpus. The final section mainly highlights current directions of research.

2. The Corpus Source: GDLI

GDLI, edited by Salvatore Battaglia and later by Giorgio Barberi Squarotti, is the most important historical dictionary of the Italian language in existence. Published by UTET in 21 volumes between 1961 and 2002 (with the addition of two update volumes, published in 2004 and 2009, for a total number of over 23,000 pages), GDLI covers the entire history of the Italian language, from the Middle Ages to the present day. Born with the aim of updating the *Dizionario della lingua italiana* known as "Tommaseo-Bellini" (1861-1879), which in turn was a sort of update of the famous *Vocabolario degli Accademici della Crusca*, GDLI—like its predecessors—bases its lexicographic description of Italian words on quotations taken from mainly literary works and authors. Within the entries, each definition and sub-definition is accompanied by a rich (often very rich) set of quotations, which attempt to cover the widest possible chronological span. Like and more than its predecessors, GDLI draws its quotations from a very wide range of authors and works: within the confines of the Italian literary or paraliterary (treatises, letters, translations, a few statutes) written tradition, not only those who are part of the canon of the major authors, but also a huge number of minor and minimal authors and works enter among the quoted. Overall, the breadth of the range of authors and works cited is impressive: GDLI quotes 6,226 authors and 13,848 sources (cf. Biffi and Guadagnini 2022).

Each quotation tends to preserve the syntactic autonomy of the textual passage, or rather to restore its overall sense (often beyond sentence boundaries): the GDLI entries are in fact conceived as a sort of small anthology of authorial citations, aimed at representing the uses of that particular word in the history of Italian writing (and specifically of Italian literature). These characteristics of the quotation cutting methods, combined with the very high number of authors and works consulted, make the corpus of quotations of GDLI an extremely rich textual set that can potentially be exploited as a resource in its own right.

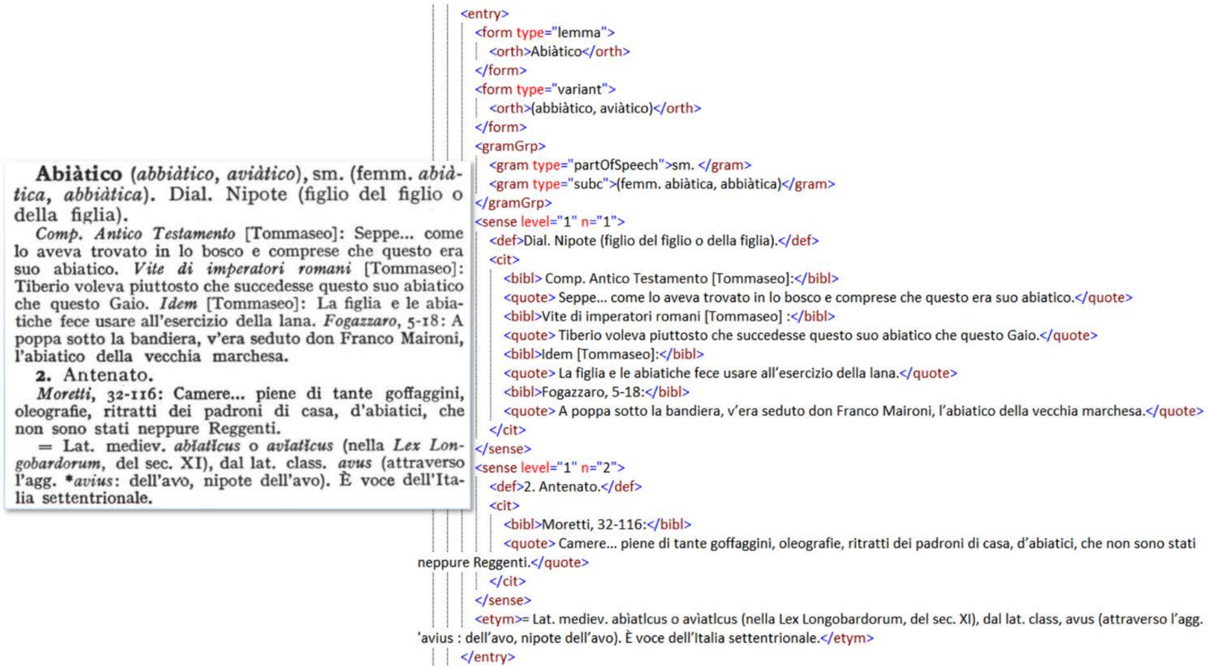
Given the peculiar history of Italian, which is in fact a written and literary language until the twentieth century, a corpus that collects all the quotations present in the GDLI entries (henceforth, referred to as GDLI Quotations Corpus, in short GDLI-QC) can be considered as a "representative" diachronic corpus of Italian (Biffi, 2018). Provided, of course, that by "representativeness" we mean the ability, offered by this corpus, to extrapolate data regarding the use of words within the boundaries of the Italian literary tradition, as it is documented by the texts that have come down to us (possibly through the medium of previous dictionaries) (Burgassi and Guadagnini, 2017, p. 11; Kabatek, 2013). It must be kept in mind, of course, that GDLI-QC is particularly appropriate for lexical research, while it is far less reliable for investigations on other linguistic planes—namely spelling and phonology. Indeed, it should be remembered that GDLI draws virtually all of its quotations from printed texts, which are not always modern critical editions: e.g., medieval or otherwise pre-normative texts may be quoted from nineteenth-century printings, where the spelling and sometimes morphological features happen thus to be sometimes modified and modernized.

In this paper, we illustrate a case study aimed at creating the GDLI-QC. With this in mind, we have created a first nucleus of a linguistically annotated corpus (divided into two sub-corpora: Annotated GDLI-QC-prose and Annotated GDLI-QC-poetry) that is somewhat representative of the overall corpus. For the time being, linguistic annotation focused on POS tagging and lemmatization.

3. GDLI-QC Construction and Composition

3.1 GDLI Quotation Extraction

GDLI quotations were automatically extracted from the TEI XML version of the dictionary, obtained through a semi-automatic conversion process aimed at structuring the dictionary contents from the OCRed version of the dictionary. The goal of semi-automatically reaching an articulated structuring of GDLI entries has been organized into several iterative steps, each with the function of progressively refining and organizing the dictionary structure previously identified. The general approach to the extraction and structuring of GDLI contents, described in Sassolini et al. (2019), Biffi et al. (2020) and Sassolini et al. (2021), adopts a strategy substantially based on pattern matching. The specific identification criteria cover a wide range of features ranging from the layout of the page to structural information relating to the different parts of the lexical entry. The goal is focused on the conversion of the dictionary contents into macroareas structured and mapped in the XML TEI standard format.



Abiàtico (*abbiàtico, aviàtico*), sm. (femm. *abiàtica, abbiàtica*). Dial. Nipote (figlio del figlio o della figlia).

Comp. Antico Testamento [Tommaseo]: Seppe... come lo aveva trovato in lo bosco e comprese che questo era suo abiatico. *Vite di imperatori romani* [Tommaseo]: Tiberio voleva piuttosto che succedesse questo suo abiatico che questo Gaio. *Idem* [Tommaseo]: La figlia e le abiatiche fece usare all'esercizio della lana. *Fogazzaro*, 5-18: A poppa sotto la bandiera, v'era seduto don Franco Maironi, l'abiatico della vecchia marchesa.

2. Antenato.

Moretti, 32-116: Camere... piene di tante goffaggini, oleografie, ritratti dei padroni di casa, d'abiatici, che non sono stati neppure Reggenti.

= Lat. mediev. *abiatĭcus* o *aviatĭcus* (nella *Lex Longobardorum*, del sec. XI), dal lat. class. *avus* (attraverso l'agg. **avius*: dell'avo, nipote dell'avo). È voce dell'Italia settentrionale.

```
<entry>
  <form type="lemma">
    <orth>Abiàtico</orth>
  </form>
  <form type="variant">
    <orth>(abbiàtico, aviàtico)</orth>
  </form>
  <gramGrp>
    <gram type="partOfSpeech">sm. </gram>
    <gram type="subc">(femm. abiàtica, abbiàtica)</gram>
  </gramGrp>
  <sense level="1" n="1">
    <def>Dial. Nipote (figlio del figlio o della figlia).</def>
    <cit>
      <bibl> Comp. Antico Testamento [Tommaseo]:</bibl>
      <quote> Seppe... come lo 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. abìatlcus o avìatlcus (nella Lex Longobardorum, del sec. XI), dal lat. class, avus (attraverso l'agg.
'avius : dell'avo, nipote dell'avo). È voce dell'Italia settentrionale.</etym>
</entry>
```

Figure 1: TEI representation of the *abiàtico* GDLI entry

Quotation extraction is part of this iterative process. In what follows we briefly exemplify the TEI XML conversion of the GDLI quotation macrofield, which includes author, reference and quotation text information. Figure 1 exemplifies the source GDLI entry and the automatically generated TEI XML counterpart for the lemma *abiàtico* 'grandchild'. It can be seen that, for each sense, the set of quotations is annotated using the <cit> element which in turn contains one or more pairs of <bibl>/<quote> elements, respectively encoding a loosely-structured bibliographic citation (whose sub-components are not further structured at the moment) and the quotation text. For this case study we used only volumes I and II of GDLI, for which the manual revision of entry segmentation was completed.

3.2 GDLI-QC Composition

We developed two sub-corpora selected to be representative of the whole GDLI-QC. The most cited authors in the dictionary were considered (cf. Biffi and Guadagnini, 2022), choosing those who would allow to cover the widest chronological span. These writers are milestones in Italian literature and history of Italian language, such as Dante, Boccaccio, Petrarca, Ariosto and Manzoni. Their different linguistic features, determined by diachronic and stylistic factors, are very valuable to test and possibly retrain linguistic annotation tools, that, as we already observed, are usually trained on contemporary language varieties (typically, newswire texts). Moreover, we chose authors and works representative of different text typologies: texts belong to several genres, such as chronicle, literary prose, poetry, treatises. This is a first experiment carried out with a view to the future structuring of GDLI-QC in balanced sub-corpora both in diachrony and based on text belonging to different genres.

Author	Century	Quotes	Tokens
Dante Alighieri (*Convivio*)	XIV	100	2839
Giovanni and Matteo Villani (*Nuova Cronica*)	XIV	100	2114
Giovanni Boccaccio (*Decameron*)	XIV	100	2681
Leon Battista Alberti	XV	100	1931
Baldassarre Castiglione	XVI	100	2307
Niccolò Machiavelli	XVI	100	2102
Giorgio Vasari	XVI	100	2549
Daniello Bartoli	XVII	100	2843
Giambattista Vico	XVII-XVIII	100	2149
Giacomo Leopardi	XVIII-XIX	100	2089
Alessandro Manzoni (*I promessi sposi* [1840])	XIX	100	2327
Ippolito Nievo	XIX	100	2363
Oscar Luigi Pirandello	XIX-XX	100	1982
Alberto Moravia	XX	100	2166
Vasco Pratolini	XX	100	2294
	tot.	1500	34736

Table 1: Annotated GDLI-QC_prose composition

As a result, the first nucleus of GDLI-QC are two balanced sub-corpora, concerning works written between 14th and 20th century: one collecting 1500 prose quotes (henceforth, Annotated GDLI-QC_prose) from 15 authors (100 each), see Table 1; one gathering 500 poetry quotes (henceforth, Annotated GDLI-QC_poetry) from 10 authors (50 each), see Table 2. Annotated GDLI-QC_prose size is about 35.000 tokens, whereas Annotated GDLI-QC_poetry is about 10.000.

Author	Century	Quotes	Tokens
Francesco Petrarca	XIV	50	1043
Matteo Maria Boiardo	XV	50	1109
Ludovico Ariosto	XVI	50	1115
Torquato Tasso	XVI	50	1152
Giovan Battista Marino	XVII	50	1111
Vittorio Alfieri	XVIII	50	1099
Ugo Foscolo	XVIII-XIX	50	947
Giosuè Carducci	XIX-XX	50	937
Giovanni Pascoli	XIX-XX	50	880
Eugenio Montale	XX	50	762
	tot.	500	10115

Table 2. Annotated GDLI-QC_poetry composition

4. Linguistic Annotation

Next step was corpus annotation. First, texts were preprocessed to reach a unified text segmentation. In fact, each quote of both sub-corpora was processed as an individual sentence; furthermore, we removed slashes, used to separate lines in poetry quotes, to focus on the underlying syntactic structure while disregarding the verse unity (which potentially pertains a distinct annotation layer). We could do that also because GDLI quotations are syntactically complete. This means that, already in the dictionary, poetry quotations are considered as “normal” sentences.

Both sub-corpora were automatically annotated through Stanza (Qi et al., 2020), a state-of-art fully neural pipeline for multilingual NLP trained on Universal Dependencies treebanks (UD, De Marneffe et al. 2021). Annotation concerned tokenization, POS tagging and lemmatization (sentence splitting was not needed here due to the overlapping with the quotation).

Automatic annotation was then manually revised and whenever needed corrected to create gold standard corpora. Regarding lemmatization, we chose a low-level lemmatization strategy; in fact, we kept the same graphical and phonological features for historical variants (e.g. *amministragione* vs *amministrazione*) and allotropes (e.g. *vizio* vs *vezzo),* which potentially cause errors in all models. The only exception regards variants with apocope (*cor* vs *core*, *fratel* vs *fratello* etc.), because this linguistic phenomenon, widespread in poetic language, is also common in contemporary Italian (*dir* vs *dire*, *buon* vs *buono* etc.). Normalization of lemma variants will be carried out as a post-processing step, in order to make it possible—in perspective—to query the corpus on different abstraction levels.

To improve the POS tagging and lemmatization accuracy on historical varieties of Italian, each gold Annotated GDLI-QC sub-corpus was split in two parts: 80% was used for retraining, and the remaining 20% for testing. ISDT, the biggest UD treebank for contemporary Italian (Bosco et al., 2013), was used in combination with corpora representative of the historical varieties of language to be analysed, in particular: for prose annotation, the VoDIM annotated corpus (Favaro et al., 2020) and the Annotated GDLI-QC_prose sub-corpus to be used for training; for poetry annotation, the Annotated GDLI-QC_poetry sub-corpus was also used for retraining. Tables 3 and 4 show the composition of the corpora used for retraining, for prose and poetry respectively.

Training corpus	Tokens
ISDT	260173
VoDIM	16250
Annotated GDLI-QC_prose	27711
tot.	304310

Table 3. Annotated GDLI-QC_prose training corpus

Training corpus	Tokens
ISDT	260173
VoDIM	16250
Annotated GDLI-QC_prose	27711
Annotated GDLI-QC_poetry	8090
tot.	312400

Table 4. Annotated GDLI-QC_poetry training corpus

5. Evaluation of POS Tagging and Lemmatization

Tables 5 and 6 show the accuracy scores respectively obtained for POS tagging and lemmatization, with the baseline and retrained models.

	UPOS	XPOS	UFeats
Baseline Model	96%	96%	96%
GDLI-QC prose retrained Model	97%	97%	96%
Baseline Model	92%	92%	92%
GDLI-QC poetry retrained Model	94%	94%	93%

Table 5. POS tagging accuracy

	Lemma
Baseline Model	94%
GDLI-QC prose retrained Model	97%
Baseline Model	90%
GDLI-QC poetry retrained Model	94%

Table 6. Lemmatization accuracy

As a baseline, we used the Stanza “combined” model, pre-trained with a combination of available Italian UD treebanks. The retrained models for prose and poetry were obtained by using the training corpora listed in Tables 3 and 4 above. To test the performances of the different models (baseline and retrained), we used a 5-fold cross validation. So, the results in the tables are an average of 5 training iterations and 5 test set evaluations.

Let us compare now the overall results achieved with the baseline and retrained models. Contrary to our expectations, baseline POS tagging models are still effective in relation to GDLI-QC_prose, even in the case of older diachronic varieties (see below). Indeed, the accuracy of the GDLI-QC prose retrained POS tagging model increases only by 1% for both Universal POS (UPOS) and language-specific POS (XPOS). No improvement is reported for what concerns Universal Features (UFeats), showing the same value in both baseline and retrained models. Regarding GDLI-QC_poetry, the accuracy distance between the baseline POS tagging model and the GDLI-QC poetry retrained model is bigger (+2% for UPOS and XPOS, +1% for UFeats). This distance further increases if we consider lemmatization results: both retrained lemmatizers show higher accuracy values (97% for prose and 94% for poetry). Although there is still room for improvement, we believe that the strategy adopted is already able to effectively face the language variability and complexity typical of historical varieties of language.

A last remark is in order here. Namely, model retraining doesn't require a large amount of data. Performances significantly increase through just a handful of tokens concerning specific historical varieties: for prose they represent 15% of the whole training corpus and for poetry 17%.

		Baseline Model		Retrained Model	
Author	Century	POS	Lemma	POS	Lemma
Dante	XIV	95%	91%	95%	95%
Villani	XIV	98%	96%	98%	98%
Boccaccio	XIV	94%	93%	95%	97%
Alberti	XV	91%	87%	93%	93%
Castiglione	XVI	98%	93%	98%	97%
Machiavelli	XVI	96%	93%	97%	96%
Vasari	XVI	98%	96%	98%	97%
Bartoli	XVII	97%	95%	97%	97%
Vico	XVII-XVIII	96%	96%	97%	98%
Leopardi	XVIII-XIX	98%	94%	99%	98%
Manzoni	XIX	97%	96%	98%	98%
Nievo	XIX	98%	96%	98%	99%
Pirandello	XIX-XX	98%	96%	97%	98%
Moravia	XX	98%	95%	98%	98%
Pratolini	XX	98%	97%	98%	97%

Table 7. Authors accuracy (GDLI-QC prose)

We also carried out an analysis of the annotation accuracy registered for single authors, detailed in Tables 7 and 8 (note that POS accuracy values refer here to the Universal POS, UPOS).

		Baseline Model		Retrained Model	
	Century	POS	lemma	POS	lemma
Petrarca	XIV	86%	86%	91%	95%
Boiardo	XV	92%	88%	97%	93%
Ariosto	XVI	93%	90%	94%	94%
Tasso	XVI	91%	91%	96%	95%
Marino	XVII	94%	90%	93%	94%
Alfieri	XVIII	91%	87%	91%	95%
Foscolo	XVIII-XIX	91%	89%	96%	96%
Carducci	XIX-XX	95%	92%	97%	96%
Pascoli	XIX-XX	96%	90%	95%	95%
Montale	XX	96%	92%	95%	95%

Table 8. Authors accuracy (GDLI-QC poetry)

In general, POS tagging and lemmatization results achieved with retrained models show a significant improvement with respect to the baseline. The biggest difference between baseline and retrained models is recorded for Alberti (prose), Petrarca and Alfieri (poetry). Only for Petrarca this distance could be explained in terms of diachronic factors: most part of the errors involves historical variants, such as functional words with apheresis (*'l* vs *il*), words with single consonant instead of double (*abassare* vs *abbassare*), verbal polimorphy (*fuor* vs *furono*), to mention only a few. These kinds of errors are fewer in the retrained model (5 vs 18 in the baseline model), but still significant since they represent 56% of the total number of errors (in the baseline model the error percentage was 67%). For Alberti and Alfieri, annotation difficulties are more likely concerned with other features of their language use. For example, Alberti adopted an Italian language graphically near to Latin, making even functional elementary words like conjunction *e* 'and' (in Alberti *et*) difficult to process. In particular, we observe that *et* and words with similar graphical features (*adricto* vs *addiritto*, old italian form for *diritto* or *dritto*; *adviato* vs *avviato* etc.) cover 30% of the errors in the baseline model, whereas this percentage drops to 12% in the retrained model.On the other hand, Alfieri uses in his verses a solemn style, full of classical poetic forms, both phonological—many words are contracted with apocope, e.g. *cor, figliuol*—and lexical (*alma, nascoso, prisco* etc.) variants, that correspond to 57% of errors in the baseline model, and drop to 27% in the retrained.

Besides the individual cases reported above, it is very interesting to note that retrained models reach very good results also in relation to Middle Ages authors, especially with prose quotations. For example, Villani's (14th century) accuracy scores are very close to values reported for 19th and 20th century authors.

Stylistic features also affect POS tagging performances, due to complex and archaic syntactic constructions occurring in these texts. Consider, for example, the following Alfieri's quotation from *Rime* (Maggini F. ed., Firenze, 1933, 83):

> *«Cede ei talor, ma ai tempi rei non serve; abbonito e temuto da chi regna, non men che dalle schiave alme proterve»* (Eng. 'Sometimes he surrenders, but in guilty times (it) doesn't serve; calmed down and feared by those who reign, not less than by insolent slave souls')

where the sequence *schiave alme proterve* represents a complex syntactic structure, used mostly in poetry as a figure of speech, formed by a noun nestled between two adjectives, one on its right, one on its left. So, because of the rare syntactic construction as well as the rare used poetic words (*alme* and *proterve*), only *schiave* 'slave' was properly tagged as an adjective by the models, whereas *alme* 'souls' and *proterve* 'indolent' were erroneously annotated as verbs (instead of noun and adjective, respectively), which also lead to lemmatization errors.

These preliminary results require further investigation; however, they clearly show that diachronic factors are not the only ones contributing to the distance between the investigated authors and contemporary Italian. Underlying this distance there could be stylistic factors, or the textual genre or the linguistic register the text belongs to (see Favaro et al., 2020). The used reference editions represent another variable that will need to be carefully evaluated and managed in subsequent developments.

6. Conclusions

We presented the first steps towards the creation of a linguistically annotated diachronic corpus for Italian, including both prose and poetry and covering a wide timespan (going from the 14th to the 20th century), which is compliant with respect to the current *de facto* representation standard of Universal Dependencies. We focused on the design, preprocessing and composition of the corpus and on the adaptation of annotation tools to reliably process diachronic varieties of language use. The encouraging results achieved so far suggest that it will soon be possible to linguistically annotate the whole GDLI-QC with a high degree of accuracy, which however can be further improved. Current directions of research include: experiments aimed at identifying the most appropriate model for processing texts of a given author or specific variety of language use; the definition of an incremental strategy for lemmatizing texts characterized by a high degree of variability.

7. Acknowledgements

This work was supported by the project "Trattamento Automatico di Varietà Storiche di Italiano" ('Automatic processing of historical varieties of Italian', TrAVaSI) funded by Regione Toscana (POR FSE 2014 - 2020) with the financial support of Accademia della Crusca.

8. Bibliographical References

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

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, in press.

Biffi, M. and Sassolini, E. (2020), Strategie e metodi per il recupero di dizionari storici, in Marras, C., Passarotti, M., Franzini, G. & 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, pp. 235-239.

Bollmann M. (2013), POS Tagging for Historical Texts with Sparse Training Data. In Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 11-18, Sofia, Bulgaria, August. Association for Computation Linguistics (ACL).

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.

D'Achille P. and Grossmann M. (2017), Per la storia della formazione delle parole in italiano. Un nuovo corpus in rete (MIDIA) e nuove prospettive di studio, Florence, Italy: Franco Cesati.

De Felice, I., Dell'Orletta, F., Venturi, F., Lenci, A. and Montemagni S. (2018), Italian in the Trenches: Linguistic Annotation and Analysis of Text of the Great War. In Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it, 10-12 dicembre 2018), pages 160-164, Torino, Italy, December. Associazione Italiana di Linguistica Computazionale (AILC).

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

Dereza O. (2018), Lemmatization for Ancient Languages: Rules or Neural Networks?, in Ustalov D., Filchenkov A., Pivovarova L. & Žižka J. (Eds), Artificial Intelligence and Natural Language 7th International Conference, AINL 2018 St. Petersburg, Russia, October 17–19, 2018 Proceedings, Cham, Switzerland: Springer, pp. 35-47.

Favaro M., Biffi M. and Montemagni S. (2020), Risorse e strumenti per le varietà storiche dell'italiano: il progetto TrAVaSI. In Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020), Online, Bologna, Italy. Associazione Italiana di Linguistica Computazionale (AILC).

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

Hämäläinen M., Partanen N. and Alnajjar K. (2021), Lemmatization of Historical Old Literary Finnish Texts in Modern Orthography. In Actes de la 28e Conférence sur le

Traitement Automatique des Langues Naturelles, pages 189-198, Lille, France, June. TANL-RECITAL.

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

Hupkes D. and Bod R. (2016), POS-tagging of Historical Dutch. In LREC 2016: Tenth International Conference on Language Resources and Evaluation (May 23-28), pages 77-82, Portorož, Slovenia, May. European Language Resource Association (ELRA).

Iacobini C., De Rosa A., Schirato G., Part-of-Speech tagging strategy for MIDIA: a diachronic corpus of the Italian language, in Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014, 9-11 December 2014, pages 213-218, Pisa, Italy, December. Associazione Italiana di Linguistica Computazione (AILC).

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

Lenci, A., Montemagni, S., Boschetti, F., De Felice, I., De Rossi, F., Dell’Orletta, F., Di Giorgio, M., Miliani, M., Passaro, L. C., Puddu, A., Venturi, G., and Labanca, N. (2020), Voices of the Great War: A Richly Annotated Corpus of Italian Texts on the First World War. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020, 11– 16 maggio 2020), pages 911-918, Marseille, France, May. European Language Resource Association (ELRA).

LIZ 4.0: letteratura italiana Zanichelli CD-ROM dei testi della letteratura italiana. Stoppelli, P., Picchi, E. (Eds.), Zanichelli, 2001

Marazzini, C. and Maconi L. (2018), Il Vocabolario dinamico dell’italiano moderno rispetto ai linguaggi settoriali. Proposta di voce lessicografica per il redigendo VoDIM, Italiano digitale, 7(4): 101-20.

Micheli, M.S. (2022), CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field, Corpus (23).

Onelli C., Proietti D., Seidenari C. and Tamburini F. (2006), The DiaCORIS project: a diachronic corpus of written Italian. In 5th Conference on Language Resources and Evaluation (LREC2006), pages 1212-1215, Genoa, Italy, May. European Language Resource Association (ELRA).

Pennacchiotti, M., Zanzotto, F.M. (2008). Natural Language Processing Across Time: An Empirical Investigation on Italian. In Proceedings of GoTAL - 6th International Conference on Natural Language Processing (25-27 August 2008), pages 371-382, Gothenburg, Sweden, August. Centre for Language Technology (CLT).

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

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, pp 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, pp. 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. & 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, pp. 159-166.

Squillacioti, P. (2021), I progetti digitali dell’OVI. “Griseldaonline”, 20, 2, pp. 197-203.

Tonelli, S., Sprugnoli, R., Moretti, G., & Kessler, F. B. "Prendo la Parola in Questo Consesso Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain". In Proceedings of CLiC-it 2019.

Yang. I., Eisenstein J. (2016), Part-of-Speech Tagging for Historical English. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (June 2016), pages 1318-1328, San Diego, California, June. Association for Computational Linguistics (ACL).