Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study

Eva Sassolini[1], Anas Fahad Khan[1], Marco Biffi[2,3], Monica Monachini[1], Simonetta Montemagni[1]

[1] Istituto di Linguistica Computazionale "A. Zampolli" - CNR (Pisa, Italy)
[2] Accademia della Crusca (Firenze, Italy)
[3] Università degli Studi di Firenze (Italy)
E-mail: {eva.sassolini, fahad.khan, monica.monachini, simonetta.montemagni}@ilc.cnr.it, marco.biffi@unifi.it

Abstract

The paper describes ongoing work on the digitization of an authoritative historical Italian dictionary, namely *Il Grande Dizionario della Lingua Italiana* (GDLI), with a specific view to creating the prerequisites for advanced human-oriented querying. After discussing the general approach taken to extract and structure the GDLI contents, in the paper we report the encouraging results of a case study carried out against two volumes which have been selected for the different conversion issues raised. Dictionary content extraction and structuring is being carried out through an iterative process based on hand coded patterns: starting from the recognition of the entry headword, a series of truth conditions are tested which allow the building and progressive structuring, in successive steps, of the whole lexical entry. We also started to design the representation of extracted and structured entries in a standard format, encoded in TEI. An outline of an example entry is also provided and illustrated in order to show what the end result will look like.

Keywords: historical dictionaries; automatic acquisition; TEI representation

1. Introduction

The digitization of historical dictionaries represents a growing convergence between lexicographers, computational linguists and digital humanists.

Research in the area dates back to the origins of computational lexicography, and has proceeded along two main lines. Since the 1980s, pioneering studies have been carried into the transformation of Machine Readable Dictionaries (MRDs) into Computational Lexicons, mainly for use in machine-oriented applications. This strategy was proposed as a way to tackle the so-called "lexical bottleneck" caused by the lack of large-scale lexical resources, indispensable for the success of realistic applications in the field of Natural Language Processing (NLP), involving e.g. syntactic parsing, word sense disambiguation, speech synthesis, information extraction, etc. Such information was acquired by exploiting the lexical entry structure of dictionaries as well as through the automatic analysis of natural language definitions: a large literature exists on this

subject, from Amsler (1981) to Calzolari (1984), Boguraev and Briscoe (1989), Montemagni and Vanderwende (1992), to mention only a few. By the mid-1990s this line of research started to go into decline as it was concluded that MRDs could not be usefully exploited for NLP applications, especially when compared with other knowledge sources such as corpora (Ide & Veronis, 1993).

Together with the acquisition of lexical knowledge from MRDs, another important issue to be tackled concerns the identification of the optimal structure, organization and representation of the resulting computational lexicons. Since the 1990s, research has started to focus on the definition of lexical representation standards, which eventually led to the definition of i) the "Lexical Markup Framework" (LMF; Francopoulo, 2013), a framework for publishing computational lexicons that today is also an ISO standard (ISO-24613:2008), and ii) Ontolex-Lemon[1] which is a *de facto* standard for publishing lexicons as linked data. In addition, the Text Encoding Initiative (TEI)[2] is now very popular for representing digital editions of lexicographic resources in XML.

Although these lines of research were focused on the development of computational lexicons mainly designed for use within Natural Language Processing applications, methods and techniques developed for extracting, structuring and representing machine-oriented dictionaries still have a potential role to play in lexicographic tasks for dictionary publishers and lexicographers, i.e. for the design and construction of human-oriented resources. As pointed out by Granger (2012), the line between machine- vs human-oriented lexical resources is progressively narrowing, thus making the synergy between these two areas of research ever more interesting.

Over the last few years, e-lexicography research has moved towards the design and construction of human-oriented online dictionaries which allow for efficient access by multiple users and which can also be easily integrated with other lexical resources and corpora (Krek, 2019). In Italy, the *Accademia della Crusca*,[3] an institution regarded as the pre-eminent authority in the study of Italian language, is moving in this direction thanks to its work on the design and construction of a dictionary of the post-Unification[4] Italian language.

The current paper reports on preliminary results of a collaboration between the *Accademia della Crusca* and the *Istituto di Linguistica Computazionale* of the Italian National Research Council (ILC-CNR) with the aim of extracting the contents of the

[1] https://www.w3.org/2016/05/ontolex/

[2] https://tei-c.org/guidelines/P5/

[3] http://www.accademiadellacrusca.it/en/pagina-d-entrata

[4] The process of Italian unification took place in the 19th century; it began in 1815 with the Congress of Vienna and was completed in 1871 when Rome became the capital of the Kingdom of Italy: during this period the different states of the Italian peninsula were unified into the single state of the Kingdom of Italy.

Grande Dizionario della Lingua Italiana ('Great Dictionary of Italian Language', henceforth GDLI) in order to convert them into structured digital data for human use and to integrate them with other language resources, both dictionaries and corpora. This collaboration is being carried out within the framework of a national project strategic for the *Accademia della Crusca* and which aims at the construction of a *Dynamic Vocabulary of Modern Italian* ('Vocabolario dinamico dell'italiano moderno', in short VoDIM)[5], within which GDLI plays a central role. A prototype digital version of GDLI, recently released by *Accademia della Crusca*, represents the starting point of the case study presented in this paper.

This case study presents itself as a challenging test bed at different levels, in particular: the extraction and structuring of the contents of the dictionary, starting from methods and techniques developed over the years for acquiring lexical knowledge from digital dictionaries; the design of a lexical representation model for the extracted and structured entries of such a complex historical digital dictionary in a standard format, encoded in TEI, with a specific view to enabling interoperability, comparability and further ease of exploitation. In what follows, the results achieved so far are presented, together with the current directions of research. After a short introduction to the GDLI dictionary and its main features (Section 2), Section 3 illustrates the general strategy adopted for extracting and structuring the dictionary contents from the OCRed version of the dictionary, the challenges to be tackled, the solutions adopted and a preliminary evaluation of results achieved so far. The final section of the paper (4) discusses the issues which are being addressed to convert the extracted contents in a standardized lexical representation format and shows how the end result will look.

2. The dictionary

The *Grande Dizionario della Lingua Italiana,* conceived by Salvatore Battaglia and released periodically in successive volumes between 1961 and 2002, is the most important historical dictionary of Italian ever published and covers the entire chronological period of the language, from its origins in the XIII century to the present day. The dictionary was published under the aegis of UTET *Grandi Opere* and maintains the legacy of a great publishing tradition: the UTET publishing house is, in fact, the oldest in Italy, having been founded in 1791. GDLI consists of 22,700 pages divided into 21 volumes, containing 183,594 entries. Word usage is documented through

[5] The main goal of the VoDIM project is the construction of a vocabulary of post-unitary Italian that gathers together the national linguistic heritage of the official language of the State from 1861 to the present day. It was funded through two Research Projects of National Relevance (PRIN), in 2012 ('Corpus di riferimento per un Nuovo vocabolario dell'italiano moderno e contemporaneo'), and in 2015 ('Vocabolario dinamico dell'italiano post-unitario'). Numerous Italian universities and research centres are involved in the project: Piemonte Orientale, Milano, Genova, Firenze, Viterbo, Napoli, Catania, ITTIG-CNR (first phase only) and Università di Torino (second phase only). The *Accademia della Crusca* has collaborated in both projects as an external partner, the post-unitary Italian dictionary being one of its strategic activities.

14,061 citations by 6,077 authors: authors and works are indicated in the lexical entry with abbreviations, which are gathered in a separate volume with the index to authors and quotations (*Indice degli autori citati*). The dictionary also includes update volumes, published in 2004 and 2009, which document most recent and innovative uses of language.

The dictionary offers valuable information on the first attestations of words, on their variants (ranging e.g. from formal to diachronic or diatopic kinds), on the authors who quote them, and on their etymologies. The potential advantages of the digitization of such a monumental dictionary have always been clear to scholars who would have liked the same search functionalities for GDLI as those offered by the electronic version of the five editions of the *Vocabolario degli Accademici della Crusca* (1612, 1623, 1691, 1729-1738, 1863-1923) which can be accessed from the web site *Lessicografia della Crusca in Rete.*[6] The digitization and structuring of the GDLI text, by explicitly marking "macro-contexts" (e.g. lemmas, definitions, examples) as well as "micro-contexts" (e.g. foreign words, proverbs, idioms, etc.), would allow for more refined and in-depth search functionalities, permitting scholars to navigate through a rich and representative diachronic corpus of the Italian language (Biffi, 2018). This becomes even more crucial if we consider that from a careful analysis of the rich historical corpus of citations of GDLI it turned out that there are words occurring in it which were not selected as lemma entries.

Taking this idea as a starting point, the *Accademia della Crusca* signed an agreement with UTET *Grandi Opere* in September 2017 which led to the latter making the electronic version of the dictionary available for digitization and online publication. In May 2019, a prototype digital version of GDLI was released via the *Accademia della Crusca* "Digital Shelves"[7], which can be accessed and queried with basic full text functionalities. This version was acquired through optical character recognition (OCR) carried out with the FineReader application operating against the dictionary PDF files made available by UTET. Up till recently the process of text correction was limited to correcting page boundaries to avoid the erroneous splitting of words and entries. However, the manual correction of the text is now proceeding, including the correction of words in Greek. In parallel, a semi-automatic approach to text correction and structuring is being developed: the case study presented in this paper presents the general approach and the first steps taken in this direction so far. The OCR output used for the GDLI digital prototype represents the starting point of this case study.

[6] www.lessicografia.it

[7] http://www.gdli.it/

3. Extraction and structuring of dictionary contents

3.1 General approach

The process of extracting and structuring dictionary contents and converting them into TEI XML has been organized into several iterative steps, each with the function of progressively refining and organizing the dictionary structure previously identified. The iterative approach we follow consists of a series of successive refinement phases which, starting from the identification of the lemma vs the body of the lexical entry, aim to further refine this segmentation by recognizing, around this nucleus, the other fields/parts of the lexical entry. Each field requires specific strategies to identify its distinguishing features. Constraints are set incrementally, leading to an increasingly granular recognition of distinct sections/fields of the entry structure.

The final aim of the work is to structure the entire dictionary entry, but the problems due to the non-standard format do not currently allow us to make a precise estimation as to how long it will take to reach the goal. This is a long process, full of unknowns, in terms of both extraction strategies and the quality of the results. We have made a long-term work plan, that consists of milestones to be achieved progressively: 1) recognition of the headword; 2) identification of all fields of the main lemma; 3) number of main senses; 4) number of nested senses; 5) fields of every main sense; 6) fields of each nested sense 7) mapping to the standardized TEI format. To optimize the time required to complete the overall work, we decided to work on several objectives in parallel. In the case of milestone 7) it is in fact a matter of defining a final structure and format that can be implemented parallel to the extraction work. In this paper we describe the extraction work foreseen in 1) and 2) above (this section) and the mapping in TEI (Section 4).

This iterative approach to entry structure recognition was also designed to reduce the number of unavoidable errors, thanks to the semi-automatic correction of extracted and structured contents to be used as input for the further processing stages. For this reason, in parallel with the content parsing strategies, we have defined methods to facilitate manual data review and correction. At the present time we have not defined a final protocol for the treatment of cases like this, but we wanted to propose our approach as a case study for similar situations anyway, that is in situations where it is not possible to use consolidated or experimental tools and or procedures already known in the literature, and the data has a significant amount of errors. In fact, in these cases we cannot define only the extraction procedures, but at the same time we have to implement strategies to support the correction and an efficient system of revision and subsequent realignment of the extracted data.

3.2 Input data

The richly detailed resource described above poses numerous challenges for the extraction and structuring of dictionary contents which are carried out against an

OCRed version of the dictionary. As pointed out in Section 2, OCR was carried out with the conventional FineReader application operating against the PDF files made available by the publisher. Although desirable, due to time and resource constraints it was not possible to improve OCR accuracy through pre- and/or post-processing techniques on the output of a single or multiple OCR engines, as currently proposed in the literature on novel approaches for OCR accuracy enhancing.

The original text in paper format shows some stylistic features and layout choices that make OCR extremely complicated, and we had to deal with the problems which resulted. The published edition which was used adopted a subdivision of the page into 3 columns, used a non-white paper colour, as well as a very small typographic font and an equally small interline one. With a work covering a time span of 40 years, it was unavoidable that there have been changes and adjustments (even minor) which have been introduced over time to the structuring of entries and the reference corpus of GDLI. Although the basic entry structure remained constant through time there have been slight changes in its internal organization, even just at the level of layout, as exemplified in Figure 1 which reports OCRed text samples from different volumes. For this reason, this case study has been carried out on two different GDLI volumes (namely, I and XII), which were selected for the different challenges and parsing problems posed by the OCR results.

All these features made the acquisition via OCR subject to errors of various types, which prevented the possibility of using already available state of the art automatic parsing tools.

Vol.	OCR text	
01	Ammannimento, sm. Allestimento. *Fra Giordano [Crusca]:* Facevano per la guerra gli ammannimenti necessari. *Soderini,* I-215: Così fatti e simili ed altri deono essere gli ammannimenti che s'hanno ad avere in preparamento per potere a dilungo fabbricare. *Salvini,* 30-2-158: Le ore gloriosamente spendete nell'ammannimento delle nuove voci, e nella correzione, che molto importa,, del Vocabolario di nostra lingua. = Deriv. da *ammannire.*	Arricavo, sm. Marin. Estremità di un cavo fissata all'oggetto che deve essere alzato o trasportato; dormiente. Arriccare, tr. e rifl. Ant. Arricchire. *Rugieri d'Amici,* 1-20*:* Amor m'à sì ariccato / in tutto 'l meo volere, / e dato m'à a tenere / più ricca gioia mai non fue visato. *Iacopone,* 18-20: O taupino, a cui aduni? A arriccar li toi garzuni? / Da ch'èi morto, i gran boccuni se fo del tuo guadagnato. *Idem,* 26-34: Frate, non m'esser sì avaro, / ca molto caro me costi per volerte arriccare. = Deriv. da *ricco* (v.).
03	**Capocamerière,** sm. (plur. *capicamerièri).* In un albergo, in un ristorante, il primo cameriere, quello da cui dipendono tutti gli altri. *Moretti,* 17-296: Vuoi sapere che cosa si fa laggiù? Si fa : il cameriere e il capo-cameriere, il fornitore di viveri ai piroscafi italiani, il padrone di trattoria. = Comp. da *capo* e *cameriere* (v.).	Certatóre, sm. Ant. e letter. Combattente. - Anche al figur. *Alberti,* 267: Mai mi lascio stare in ozio, fugo il sonno, né giacio se non vinto dalla strachezza, che sozza cosa mi pare senza ripugnare cadere e giacere vinto, o come molti prima aversi vinti che certatori. = Deriv. da *certare.*
09	**Iniare[1], intr. con la particella pronom. (*m'inìo*). Ant. Diventare simile, identificarsi.** *Dante, Par.,* 33-44: Indi a l'etterno lume s'addrizzaro, / nel qual non si dee creder che s'inii / per creatura l'occhio tanto chiaro. *Ottimo,* III-729: [4] Nel qual non si de' creder ec. ' : cioè, si come più volte è detto, occhio creato non può iniarsi al fondo della divinitade. [4] Inii ' si è verbo informativo, ed è tanto a dire, come diventare simile di quella cosa ch'è considerata. = Denom. dal pronom. *io* col pref. *in-* con valore illativo.	**Ludro[1],** agg. e sm. Dial. Mascalzone, birbante, imbroglione, canaglia; persona avida, ingorda, insaziabile.
11	Nauseante **(part. pres. di** *nauseare),* **agg. Che provoca nausea, disgusto, voltastomaco; disgustoso, nauseabondo.** *O. Targioni Pozzetti,* I-213: Linneo li aveva divisi e classati [gli odori] in... ambrosiaci..., fragranti..., tetri o virosi..., nauseanti. *Massaia,* Vili-163: Riacquistate, con l'aiuto di quella putrida è nauseante acqua, alquanto le forze, si continuo il cammino per l'arido deserto. *Tarchetti,* 6-II-639: Permetti, bevo un bicchiere di decotto di gramigna, serrandomi prima delicatamente la punta del naso tra il pollice e l'indice della mano sinistra. Dio, che roba nauseante, è un beverone da cavallo. *Svevo,* 3- 569: Restai tranquillo a quel posto fumando quelle sigarette nauseanti. *Stuparich,* 1-333: Avverto intorno un puzzo di pesce rancido, nauseante.	Motosilurante, sf. Marin. Milit. Piccola unità navale da guerra, leggera e assai veloce, munita di motore a propulsione endotermica, per lo più a scoppio o Diesel o, anche, a turbina, armata con siluri e con qualche cannone di piccolo calibro a cadenza di tiro assai elevata, impiegata per attacchi di sorpresa in acque ristrette. *Migliorìni* [s. v.]: 'Motosilurante', nome masch. o femm.: leggiera imbarcazione a motore per il lancio dei siluri. Lo stesso che ' mas '. = Comp. da *motore*] e *silurante* (v.).

Figure 1: OCRed text samples from different volumes in Word format.

The input of the extraction and structuring work is represented by more than 23,000 pages of dictionary text, provided in a (non-standard) Word format and organized into 21 volumes preserving the same subdivision as the GDLI paper format. Since the resulting Word files are very heavy and difficult to manage, we tried to convert these to other formats (XML and TXT). It turned out that for the lemma extraction we had substantially the same problems as with the Word format, but errors in other parts of the structure made the extraction procedure more complex. Although lighter to handle and more readable, the TXT format extracted from the Word format left out important information pertaining to format and style, which is often crucial in the discrimination between, e.g. a lemma and a simple paragraph beginning (see below).

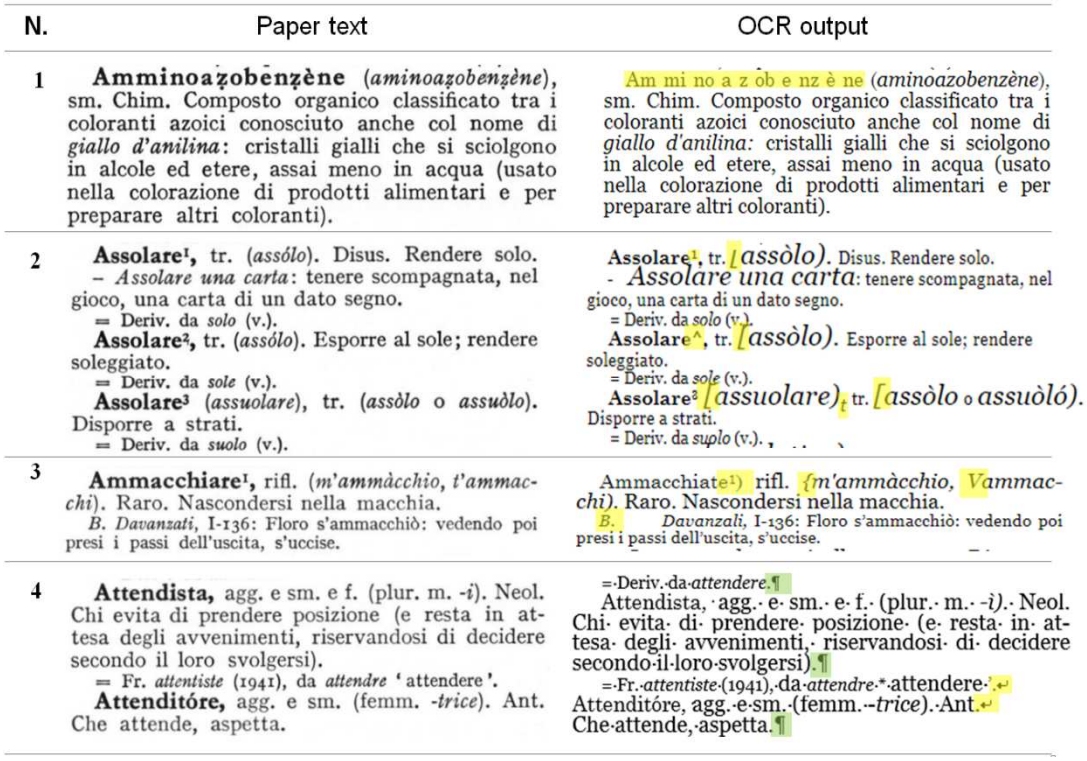
3.3 Segmentation strategy

The first phase of the work concerned the segmentation of the Word format (".doc") file of each individual volume into portions of no more than 50-60 pages, each of which was saved in a separate file, and analysed in succession by the parsing program. The entire process required the use of numerous software libraries capable of parsing the Word format and identifying the peculiarities of the structural and formatting characteristics of the text. The segmentation procedure was performed manually to avoid the inappropriate cutting up of individual entries across pages. At this stage and with unavoidably noisy input, a fully automatic system would have not produced a sufficiently good result when applied to dictionary texts in which lexical entries are typically organized in relatively long enumerations of nested senses each of which also contains related quotations.

The second step consisted in the segmentation of individual pages recognized at the previous step into lexical entries, whose boundaries were explicitly marked. For each identified lexical entry, the headword (or lemma) and a text area corresponding to the body of the entire entry is recognized. The segmentation procedure proceeds with the identification of the other entry fields, according to similar methods used for the headword.

These further steps include the iterative segmentation of the body of the lexical entry into different blocks with grammatical information (including the indication of possible variants, e.g. orthographic, diatopic, diachronic, etc.), main senses, sense attestations and examples, other numbered sub-senses with examples (if any), and etymology. Each main sense block is in its turn articulated into different sections within which quotations play a central role: to quote Beltrami and Fornara (2004), "the veritable fulcrum of the dictionary is the massive presence of text quotations from authors". These quotations cover a wide variety of language use, from everyday and literary language, dialectical and regional languages, to technical and scientific language, specialized languages, neologisms and foreign words.

The results of this further segmentation, which are currently being analysed in detail, are strongly influenced by the success of the lemma extraction phase. However, the type of recognition errors generated by the extraction system has also been analysed on each individual structural feature of the dictionary: lemma, spelling variants, grammatical category, usage codes, definition, etymology, main senses and additional senses (nested). Each of the fields shows errors of various types, ranging from errors in the segmentation of paragraphs, to those in the rendering of punctuation marks, to spelling errors, to the failure to identify the structural elements that define the different sections of the dictionary entry (bullet points, indentation, font size etc.). Figure 2 exemplifies some OCR errors negatively impacting on the further recognition process.



N.	Paper text	OCR output
1	**Amminoazobenzène** (*aminoazobenzène*), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di *giallo d'anilina*: cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).	Am mi no a z ob e nz è ne (*aminoazobenzène*), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di *giallo d'anilina:* cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).
2	**Assolare**[1], tr. (*assólo*). Disus. Rendere solo. – *Assolare una carta*: tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da *solo* (v.). **Assolare**[2], tr. (*assólo*). Esporre al sole; rendere soleggiato. = Deriv. da *sole* (v.). **Assolare**[3] (*assuolare*), tr. (*assòlo* o *assuòlo*). Disporre a strati. = Deriv. da *suolo* (v.).	**Assolare**[1], tr. (*assòlo*). Disus. Rendere solo. - *Assolare una carta*: tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da *solo* (v.). **Assolare**[2], tr. (*assòlo*). Esporre al sole; rendere soleggiato. = Deriv. da *sole* (v.). **Assolare**[3] (*assuolare*), tr. (*assòlo* o *assuòlo*). Disporre a strati. = Deriv. da *suolo* (v.).
3	**Ammacchiare**[1], rifl. (*m'ammàcchio, t'ammacchi*). Raro. Nascondersi nella macchia. *B. Davanzati*, I-136: Floro s'ammacchiò: vedendo poi presi i passi dell'uscita, s'uccise.	Ammacchiare[1]) rifl. (*m'ammàcchio, Vammacchi).* Raro. Nascondersi nella macchia. *B. Davanzali,* I-136: Floro s'ammacchiò: vedendo poi presi i passi dell'uscita, s'uccise.
4	**Attendista,** agg. e sm. e f. (plur. m. *-i*). Neol. Chi evita di prendere posizione (e resta in attesa degli avvenimenti, riservandosi di decidere secondo il loro svolgersi). = Fr. *attentiste* (1941), da *attendre* 'attendere'. **Attenditóre,** agg. e sm. (femm. *-trice*). Ant. Che attende, aspetta.	=·Deriv.·da·*attendere*.¶ Attendista, ·agg.· e· sm.· e· f.· (plur.· m.· *-i).*· Neol. Chi· evita· di· prendere· posizione· (e· resta· in· attesa· degli· avvenimenti,· riservandosi· di· decidere secondo·il·loro·svolgersi).¶ =·Fr.·*attentiste*·(1941),·da·*attendre*·*·attendere.·↵ Attenditóre, agg.·e·sm.·(femm.·-*trice*).·Ant.↵ Che·attende,·aspetta.¶

Figure 2: Examples of blocking OCR errors.

These errors often block the correct segmentation of the internal structure of the entry, especially for what concerns the extraction of senses and sub-senses. The frequent co-presence of more than one error within the same entry makes the recognition of the internal structure a more challenging problem.

3.4 Main error types

The main types of errors concern the OCR format, and they impose an unavoidable conditioning on the quality of the extraction phase. Other errors, introduced by the parsing phase, could be added to these. A bad interpretation of the structure of the entry during the OCR process will obviously mislead the system, invalidating the extraction both of the lemma and other fields. We tried to organize the variety of anomalous phenomena encountered so far into six main error types, listed below:

1. “omission”, occurring when parts of the lexical entry (including substrings of characters) have been omitted;

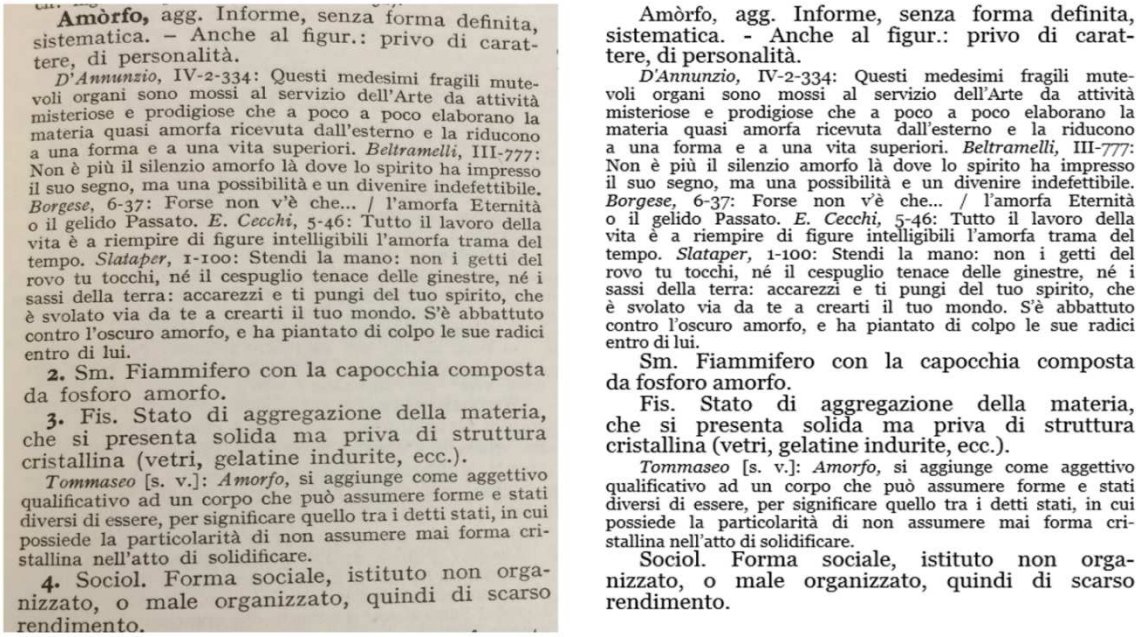
2. “illegal merger”, occurring when different fields within a lexical entry or two lexical entries are wrongly merged (see example n. 4 in Figure 2);

3. “illegal disjunction”, corresponding to wrongly segmented words: e.g. ‘Ab borrire e deriv.’ for ‘Aborrire e deriv.’; ‘A c cespugli are’ for ‘Accespugliare’; ‘Acetilèni co’ for ‘Acetilenico’; ‘Acòre e a còro’ for ‘Acòre e acòro’; etc.;

4. “incorrect graphemes”, corresponding to wrongly interpreted sequences of graphemes of the same length: e.g. ‘sl’ for ‘sì’, ‘ero’ for ‘cro’, ‘cto’ for ‘chi’; ‘ln’ for ‘ìn’ or ‘lì’ or ‘li’, etc.;

5. “exchange of graphemes”, corresponding to wrongly interpreted sequences of graphemes of different length (i.e. expansion or contraction): e.g. ‘lite’ for ‘nte’; ‘til’ for ‘rell’; ‘fif’, ‘flf’ or ‘tif’ for ‘ff’; ‘dd’ for ‘cìcl’; ‘g’ for ‘ci’, etc.;

6. “missing bullet points”, which are mainly concerned with the recognition of senses as exemplified in Figure 3, where the OCRed text on the right lacks sense numbers.



Amòrfo, agg. Informe, senza forma definita, sistematica. – Anche al figur.: privo di carattere, di personalità.
D'Annunzio, IV-2-334: Questi medesimi fragili mutevoli organi sono mossi al servizio dell'Arte da attività misteriose e prodigiose che a poco a poco elaborano la materia quasi amorfa ricevuta dall'esterno e la riducono a una forma e a una vita superiori. *Beltramelli,* III-777: Non è più il silenzio amorfo là dove lo spirito ha impresso il suo segno, ma una possibilità e un divenire indefettibile. *Borgese,* 6-37: Forse non v'è che... / l'amorfa Eternità o il gelido Passato. *E. Cecchi,* 5-46: Tutto il lavoro della vita è a riempire di figure intelligibili l'amorfa trama del tempo. *Slataper,* 1-100: Stendi la mano: non i getti del rovo tu tocchi, né il cespuglio tenace delle ginestre, né i sassi della terra: accarezzi e ti pungi del tuo spirito, che è svolato via da te a crearti il tuo mondo. S'è abbattuto contro l'oscuro amorfo, e ha piantato di colpo le sue radici entro di lui.
2. Sm. Fiammifero con la capocchia composta da fosforo amorfo.
3. Fis. Stato di aggregazione della materia, che si presenta solida ma priva di struttura cristallina (vetri, gelatine indurite, ecc.).
Tommaseo [s. v.]: *Amorfo,* si aggiunge come aggettivo qualificativo ad un corpo che può assumere forme e stati diversi di essere, per significare quello tra i detti stati, in cui possiede la particolarità di non assumere mai forma cristallina nell'atto di solidificare.
4. Sociol. Forma sociale, istituto non organizzato, o male organizzato, quindi di scarso rendimento.

Amòrfo, agg. Informe, senza forma definita, sistematica. - Anche al figur.: privo di carattere, di personalità.
D'Annunzio, IV-2-334: Questi medesimi fragili mutevoli organi sono mossi al servizio dell'Arte da attività misteriose e prodigiose che a poco a poco elaborano la materia quasi amorfa ricevuta dall'esterno e la riducono a una forma e a una vita superiori. *Beltramelli,* III-777: Non è più il silenzio amorfo là dove lo spirito ha impresso il suo segno, ma una possibilità e un divenire indefettibile. *Borgese,* 6-37: Forse non v'è che... / l'amorfa Eternità o il gelido Passato. *E. Cecchi,* 5-46: Tutto il lavoro della vita è a riempire di figure intelligibili l'amorfa trama del tempo. *Slataper,* 1-100: Stendi la mano: non i getti del rovo tu tocchi, né il cespuglio tenace delle ginestre, né i sassi della terra: accarezzi e ti pungi del tuo spirito, che è svolato via da te a crearti il tuo mondo. S'è abbattuto contro l'oscuro amorfo, e ha piantato di colpo le sue radici entro di lui.
Sm. Fiammifero con la capocchia composta da fosforo amorfo.
Fis. Stato di aggregazione della materia, che si presenta solida ma priva di struttura cristallina (vetri, gelatine indurite, ecc.).
Tommaseo [s. v.]: *Amorfo,* si aggiunge come aggettivo qualificativo ad un corpo che può assumere forme e stati diversi di essere, per significare quello tra i detti stati, in cui possiede la particolarità di non assumere mai forma cristallina nell'atto di solidificare.
Sociol. Forma sociale, istituto non organizzato, o male organizzato, quindi di scarso rendimento.

Figure 3: Bullet points in printed vs Word formats.

In Figure 4 below is a graph showing the percentage distribution of the six error types in the two volumes of the dictionary which were selected for this case study:

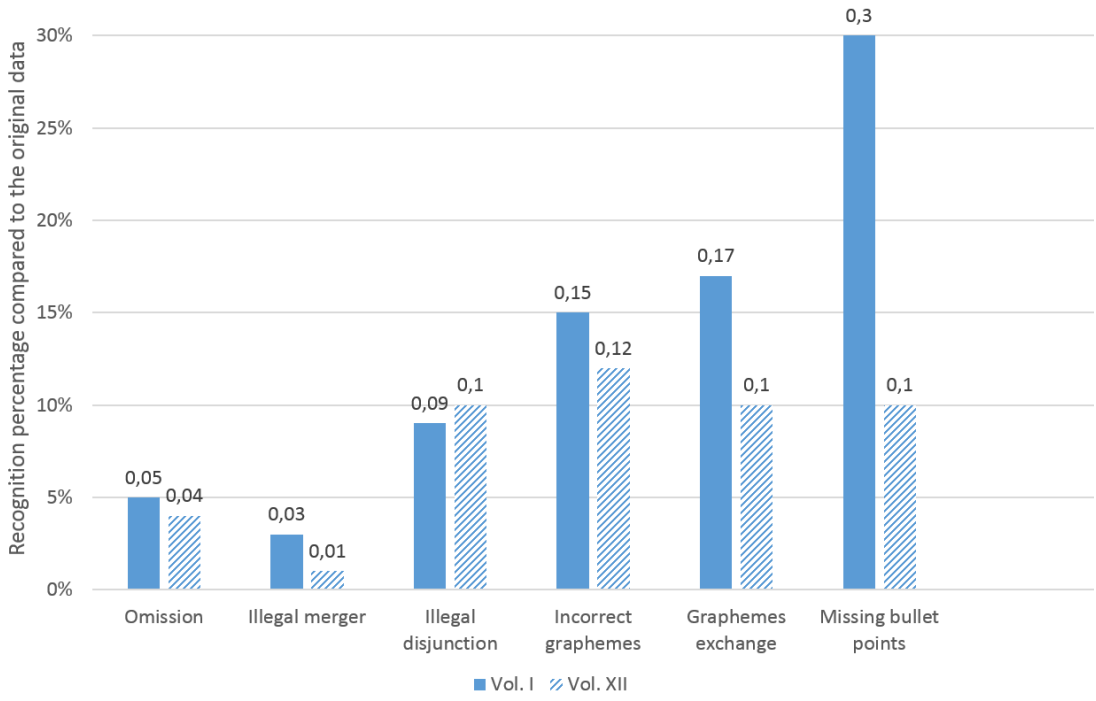


Figure 4: Distribution of error types between volumes I vs XII

It can be noted that the distribution of some types of errors differs significantly between volumes, suggesting a discrepancy of quality of OCR across them: this is the case, for instance, with "missing bullet points" and "exchange of graphemes". It is possible that the long phase of preparation of the work influenced the differences between the volumes: the quality of the print, the colours of the paper and the ink, etc. As we have already said, we have noticed differences across volumes which already visually explain the differences in the performance of OCR procedures. It is likely that the conservation status of the volumes from which the OCR was made also comes into play, and it is not certain that all the volumes were in the same condition.

3.5 Lemma extraction

The approach to the extraction and structuring of GDLI contents is illustrated here with respect to the first segmentation step, mainly aimed at lemma extraction. Due to the complexities sketched above, we decided, at least initially, to follow an approach based on pattern matching. The patterns we work with cover a wide range of characteristics ranging from the layout of the page to structural information relating to the different parts of the lexical entry. They also relate to linguistic aspects regarding the format and spelling of the lemma as well as lexicographic ordering, with respect to the lemmas that precede and follow an entry. The patterns have been defined manually

and start from the recognition of the lexical entry and its headword (lemma). The extraction phase is then determined by the identification of the characteristics listed above and the testing of different truth conditions that, placed in combination with each other, confirm, to a reasonable approximation, the beginning of the entry of the dictionary and its end.

The recognition phase of the lemma is integrated with strategies supporting the correction of incompletely or erroneously extracted lemmas.

Whenever the lemma cannot be recognized with certainty, a check on the number of conditions satisfied is activated: a lower number of verified conditions causes the positive matching of entries that are often erroneous. Based on experiments, two different thresholds have been defined: cases that verify 2/3 of the conditions for the correct recognition of the headword are reported as requiring a manual verification; those that reach 3/4 of the conditions, already acquired as headwords, are suggested for manual control, although with a lower priority assigned. These cases are recorded within a report file which is generated together with the outcome of the parsing phase. In this report file, the "candidate" lemma is written, followed by page indication and listing of conditions which have not been verified.

Even when the lemma is correctly segmented, there may be spelling errors. We analysed these cases in order to find a suitable reporting method. Starting from a cost/benefit evaluation, we studied different techniques to identify and report this type of error. One technique consists of applying lexicographic sorting criteria to the lists of lemmas extracted automatically. The comparison of the natural sequence of the headwords found in the pages, with the same lexicographically ordered list, brings out the differences in the cases of spelling errors. We have decided to turn this evidence into a correction support report. In particular, parallel to the parsing, the extraction system, for each volume analysed, produces a file containing the list of all the headwords extracted, ordered lexicographically and followed by the page number where each headword was found. In this way the misalignment between the page sequence and the ordering of the headwords is evident and provides concrete help to the manual correction phase. Another technique to test the correctness of the acquired lemma consists of looking up the acquired candidate lemma string in other reference lexical resources, historical dictionaries (for example, the *Tesoro della Lingua Italiana delle Origini* or TLIO) [8] as well as wide coverage contemporary dictionaries including historical lexical variants. Those entries for which no corresponding lemma has been found are reported for manual checking.

[8] http://tlio.ovi.cnr.it/TLIO/

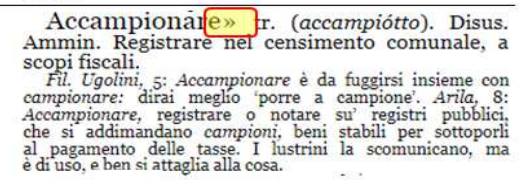
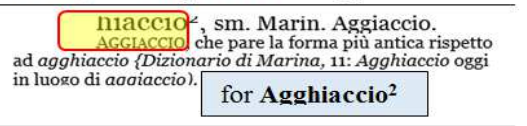
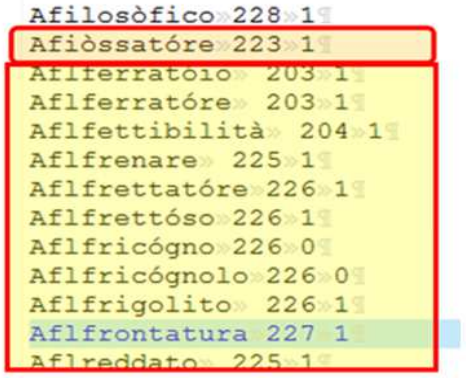
Error types	Fields	Adopted solutions	Examples
Orthographic (type n.1, 2, 3)	lexical entry	Ref. in the lemmas report	
Punctuation	lexical entry	Amendment strategies embedded in the parser	
Omissions	lexical entry	NA	
Lemma not found at the paragraph beginning	etym.	Event reported in the error report	(see Tab. 2. n.4)
Incorrect sequence of characters	lexical entry	Ref. in the lemmas report	

Table 1: Typical errors in lemma recognition

3.5.1 Specific error types

As far as lemma recognition is concerned, the largest number of errors found is distributed among error types 3), 4) and 5) listed in Section 3.4, namely "illegal disjunction", "incorrect graphemes" and "exchange of graphemes". Since these three error types have a greater impact on content extraction and structuring, it is on them that we have focused our strategies of manual correction support. Table 1 shows how these error types impact on the recognition of the lemma and the related strategies adopted to support the manual correction.

As for the "omission" type, besides manual correction, we have not found a solution at the moment. There are also possible errors when a string of characters corresponding to the true lemma is incorrectly interpreted by OCR, such that it overlaps with a previously recognized lemma.

3.5.2 Preliminary results

At the end of the acquisition experiments carried out against volumes I and XII, the results obtained for what concerns lemma extraction are promising, with an over 94% success rate, as shown in the pie chart in Figure 5. Lemmas are correctly extracted and identified in 75% of the cases; 15% of correctly acquired lemmas contain an OCR error, and 6% of them contain spelling errors (originating, for example, in the overlap with lemmas already extracted). This result, however, cannot be seen as exhaustive, because the amount of entries analysed, set against the total number contained in the GDLI, is around 10%.

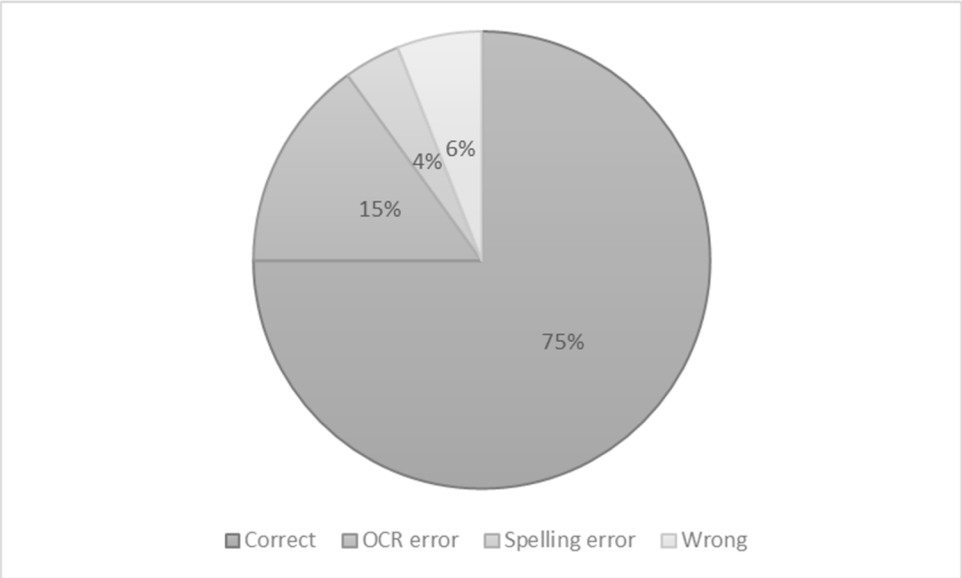


Figure 5: Lemma acquisition and identification results

4. TEI Mapping

4.1 Introduction

Although as regards the current state of progress of the work described in this paper we are still not in a position to discuss the technical details of the final conversion of the original source files into a standardized format for lexical resources such as TEI, we can show what it is we are aiming for and what the end result will look like. As pointed out above, we decided to work on both extraction and representation objectives in parallel: the reasons underlying this choice range from the optimization of the time required to complete the overall work to the fact that the adopted lexical representation model can influence, at least to some extent, the structuring of extracted contents.

In the following subsections we will describe the importance of using a specialized standard to encode the information in a resource such as the GDLI, as well as explain why we chose the TEI guidelines, and we will present an example entry from the GDLI and describe what a TEI encoding of the entry looks like.

4.2 Background on standards for lexical resources

The importance of the role of standards in the modelling, creation, and publication of computational lexical resources has gained increasing recognition in recent years. This is thanks not only to more general initiatives relating to the FAIR data principles (Wilkinson et al., 2016) but also to a growing appreciation of the critical worth of well-made lexical resources to much work in Computational Linguistics and Digital Humanities. There are several reasons why standards play such an important role in the specific case of lexical resources. For one thing the existence of lexical standards facilitates the harmonization of the different linguistic and metadata categories used in such resources, and is an important prerequisite to ensuring the interoperability of lexical datasets. Standards also allow resources to be re-used more easily and in various different contexts and tasks, and this is especially important in NLP where one single resource, such as WordNet, can be used in numerous different kinds of task. It is also more likely that, at least for the most popular and well known standards, there already exists software for creating, maintaining and publishing resources that adhere to the standards in question. Finally, in many cases these standards represent a community endorsed solution to those problems that are likely to arise when encoding various different types of lexical information.

When it comes to encoding lexical and, more specifically, lexicographic resources, there are a number of different relevant standards which should be taken into consideration, and in some cases a choice needs to be made between two or more competing standards encoding the same kinds of information. In our case it was important to choose a standard that was as widely used as possible and made use of common formats but that was also sufficiently expressive for our modelling needs. We wanted to annotate both those aspects of the resource pertaining to the source dictionary's status as a printed text, as well as to its conceptual, bibliographic and linguistic content: that is, we wanted a model that would allow us to annotate things like bibliographic citations, quotes, as well as lexical entries, senses, and etymologies. For these reasons and others we decided to choose the Text Encoding Initiative (TEI) guidelines, and especially the chapter on encoding dictionaries, as our main standard in encoding the GDLI.

In the next subsection we will look at two GDLI entries encoded in TEI to show what the end result will look like.

4.3 Example entry

In order to show what the end result of the process described in this paper will look like, as well as to highlight some of the most typical features of GDLI lexical entries and how the TEI guidelines allow us to encode these features, we present an example entry from the GDLI. The entry in question concerns the adjective *padronale*, which has the primary sense of 'pertaining to or deriving from the condition of being a boss

or master' and derives from the noun padrone 'boss, master'. The entry for *padronale* has four different senses, each of which is further subdivided into more specific sub-senses and each of which is provided with a list of citations from the corpus of historical Italian texts referred to by the GDLI. For reasons of space we will only discuss the first sense, which we show as Figure 6 (the page containing the full entry can be found here: http://www.gdli.it/JPG/GDLI12/00000348.jpg).

Padronale (*patronale*), agg. Che si riferisce o che deriva dalla condizione di padrone, dalle sue prerogative giuridiche di proprietario o dalla sua autorità nei confronti dei dipendenti, dei familiari o di altre persone; che denota, talora in modo ostentato, tale condizione; commesso, esercitato da un padrone; da padrone.

De Luca, 1-1-50: Quell'entrate e robbe, le quali abbiano annessa qualche giurisdizione o preminenza padronale, come per esempio sono li molini e forni. *Foscolo*, XVIII-253: Permetterò a Pietro d'incamminarsi tanto che dura la buona stagione; e se non altro sono consolato ch'egli non si dorrà mai giustamente di me, perché l'ho sempre trattato con volto padronale, ma con cuore fraterno. *Franzoj*, 13: I facili incrociamenti, le violenze patronali, le vergogne sifilitiche... ne hanno deturpato [degli arabi] il tipo fisico. *D'Annunzio*, IV-2-205: Don Giovanni Ussorio, presente sempre, aveva delle arie padronali. *Borgese*, 1-64: Egli passeggiava velocemente facendo cantare gli sproni..., avviato verso un'indignazione metà fredda e metà calda, donde desumeva chi sa che autorità maritale o padronale sulla donna di cui presentiva l'avvicinarsi. *Piovene*, 6-153: Era gentile per diplomazia padronale e per naturale indulgenza della persona superiore con la gente bassa.

– Per simil. Spavaldo, privo di ritegno.

Baldini, I-6526: Uscendo il treno dai monti in corsa verso il mare, il fischio della locomotiva righerebbe l'aria con quella padronale allegria della quale qui si sente propriamente la mancanza.

– Che spetta al proprietario di un podere, dominicale.

Tommaseo [s. v.]: 'Parte padronale': quella che in Toscano e altrove 'domenicale', la parte della rendita appartiene al padrone del fondo, a distinguerla da quel che ne viene al colono. *Einaudi*, 2-280: È vero il vecchio adagio del mezzadro il quale: « signor padrone – dice – venga a dividere la sua metà », ed il quarto padronale non basta a pagare le imposte.

Figure 6: Sense 1 of the *padronale* GDLI entry.

Here the nesting structure of the first sense is implicit in the sense that the sub-senses are not given identifiers (the other sub-senses of the entry are given numbers) but can be identified by the tab space and the dash. The first sense has a main sense (that starts after the part of speech information), and two more specific sub-senses.

The entry (seen at the top level with sense nodes unexpanded) is shown in Figure 7.

```
<entry>
    <form type="lemma">
       <orth>Padronale</orth>
    </form>
    <form type="variant">
       <orth>patronale</orth>
    </form>
    <gramGrp>
       <pos>agg.</pos>
    </gramGrp>
       <etym>= Deriv. da <mentioned xml:lang="it">padrone</mentioned>.</etym>
       <sense level="1" n="1"> [47 lines]
       <sense level="1" n="2"> [32 lines]
       <sense level="1" n="3"> [5 lines]
       <sense level="1" n="4"> [13 lines]
       <sense level="1" n="5"> [5 lines]
       <sense level="1" n="6"> [5 lines]
</entry>
```

Figure 7: TEI representation of the *padronale* GDLI entry with sense nodes unexpanded.

Here we have annotated the fact that the entry has the lemma *Padronale* using the TEI <form> element, specifying its type attribute as “lemma”, as well as the alternative form *patronale*. We have also annotated its part of speech using the <gramGrp> and <pos> elements, and represented the fact that the word is derived from another word using the <etym> element. Next we represent the fact that the entry has six senses (at the first level of nesting) using the <sense> element and the attributes @level and @n.

In Figure 8, we show the structure of the first sense and its two sub-senses (with the <cit> node unexpanded). All three senses have their definitions marked out using the <def> element, with each citation annotated using the <cit> element.

```
<sense level="1" n="1">
   <def>Che si riferisce o che deriva dalla condizione di padrone,
     dalle sue prerogative giuridiche di proprietario o dalla sua autorità nei confronti dei dipendenti,
     dei familiari o di altre persone; che denota, talora in modo ostentato, tale condizione;
     commesso, esercitato da un padrone; da padrone.</def>
   <cit> [2 lines]
   <cit> [3 lines]
   <cit> [3 lines]
   <cit> [2 lines]
   <cit> [4 lines]
   <cit> [3 lines]
   <sense level="2" n="1">
      <def>Per simil. Spavaldo, privo di ritegno.</def>
      <cit> [2 lines]
   </sense>
   <sense level="2" n="2">
      <def>Che spetta al proprietario di un podere, dominicale.</def>
      <cit> [2 lines]
      <cit> [2 lines]
   </sense>
</sense>
```

Figure 8: TEI representation of sense 1 of the *padronale* GDLI entry.

Finally, in Figure 9, we expand the first two citations of the first sense.

```
<sense level="1" n="1">
   <def>Che si riferisce o che deriva dalla condizione di padrone,
     dalle sue prerogative giuridiche di proprietario o dalla sua autorità nei confronti dei dipendenti,
     dei familiari o di altre persone; che denota, talora in modo ostentato, tale condizione;
     commesso, esercitato da un padrone; da padrone.</def>
   <cit>
      <bibl>De Luca, 1-1-50:</bibl><quote> Quell'entrate e robbe,
         le quali abbiano annessa qualche giurisdizione o preminenza padronale,
         come per esempio sono li molini e forni.</quote>
   </cit>
   <cit>
      <bibl>Foscolo, XVIII-253:</bibl>
      <quote>Permetterò a Pietro d'incamminarsi tanto che dura la buona stagione;
         e se non altro sono consolato ch'egli non si dorrà mai giustamente di me,
         perché l'ho sempre trattato con volto padronale, ma con cuore fraterno.</quote>
</cit>
   <cit> [3 lines]
   <cit> [2 lines]
   <cit> [4 lines]
   <cit> [3 lines]
```

Figure 9: TEI representation of citations in the *padronale* GDLI entry.

The first citation is from Giovanni Battista De Luca, the noted 17th century jurist and cardinal, and the second citation is taken from the works of Ugo Foscolo, the well-known 19th century Italian poet and political exile. In future work we are planning to add links to virtual authority files for the authors cited in the GDLI in the TEI-XML encoding itself.

From this brief description of the (manual) encoding of a single entry we hope it is clear how important such a conversion of the original resource is for rendering the linguistic, historical and cultural information inside the dictionary more machine actionable and more amenable to querying by human users.

5. Conclusion

In this paper, we presented the preliminary and encouraging results of a case study carried out to define the strategy to be adopted to extract and structure the contents of the most important historical dictionary of Italian, *Il Grande Dizionario della Lingua Italiana*, with a specific view to creating the prerequisites for advanced human-oriented querying, which allows for multiple and efficient access, can be integrated with other lexical resources and corpora, can be customized to meet specific user needs, etc. Dictionary content extraction and structuring is being carried out through an iterative process based on hand coded patterns: starting from the recognition of the entry headword, a series of truth conditions are tested which allow the building and progressive structuring, in successive steps, of the whole lexical entry. We also started to design the representation of extracted and structured entries in a standard format, encoded in TEI. After discussing the general approach taken, in the paper we focused on the early stages of the conversion of the dictionary contents into structured digital

data, with particular attention to supporting the semi-automatic correction of errors mainly originating in the OCRed parsed text.

The complex situation of the digitized version of the GDLI dictionary described in the previous sections, characterized by slightly different entry formatting and/or structuring conventions across volumes and the presence of OCR errors, led us to opt, at least for this first explorative phase, for a pattern-based approach. We are aware of the limits of this approach, i.e. the costly manual elaboration of complex patterns based on observing the organisation of the lexical information in dictionary entries, but at this stage this turned out to be the only viable approach. We are currently evaluating whether, once an appropriate quantity of dictionary entries from consistent GDLI portions has been reconstructed and corrected, a machine learning approach, such as that used by GROBID-Dictionaries (Khemakhem et al., 2017), could be usefully exploited for completing this work. The iterative approach to extraction and structuring of GDLI lexical entries proposed here creates the prerequisites for the creation of cascading extraction models which represent one of the main features of the GROBID-Dictionaries strategy for structuring digitized dictionaries.

For what concerns the GDLI representation, we are planning to evaluate whether and to what extent the representation model which is being developed within the European ELEXIS project ("European Lexicographic Infrastructure", Krek et al., 2018) aiming to establish a pan-European infrastructure for lexicography could effectively be used to represent such a complex historical digital dictionary, with a specific view to enabling efficient access to high quality lexicographic data.

6. Acknowledgments

The authors have been partly supported by the EU H2020 programme under grant agreement 731015 (ELEXIS – European Lexicographic Infrastructure).

7. References

Amsler, M. A. (1981). A taxonomy for English nouns and verbs. In *Proceedings of the 19th Annual Meeting of the ACL*, pp. 133-138.

Beltrami, P. G. & Fornara, S. (2004). Italian historical dictionaries: from the Accademia della Crusca to the web. *International Journal of Lexicography*, 17(4), pp. 357-384.

Biffi, M. (2018). Strumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano post-unitario. In D. Fioredistella et al. (eds.) *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data (JADT '18)*, Roma, Universitalia, vol. 1, pp. 99-107.

Boguraev, B. & Briscoe T. (eds.) (1989). *Computational Lexicography for Natural Language Processing*. Longman.

Calzolari, N. (1984). Detecting Patterns in a Lexical Database. In *Proceedings of the*

10th International Conference on Computational Linguistics, Stanford, California, pp. 170-173.

Francopoulo, G. (ed.) (2013). *LMF Lexical Markup Framework*. John Wiley & Sons *Grande Dizionario della lingua italiana*, Opera Diretta da Salvatore Battaglia, Torino, UTET, 1961-2002.

Granger, S. (2012). Electronic lexicography: From challenge to opportunity. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*, Oxford University Press, pp.1-11.

Ide, N. & Veronis, J. (1993). Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? *Knowledge Bases & Knowledge Structures*, 93, Tokyo.

Khemakhem, M., Foppiano, L. & 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.

Krek, S. (2019). Natural Language Processing and Automatic Knowledge Extraction for Lexicography. *International Journal of Lexicography*, 32(2), pp. 115-118.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B., Tiberius, C. & 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, pp. 881–891.

Montemagni, S. & Vanderwende, L. (1992). Structural Patterns versus String Patterns for Extracting Semantic Information from Dictionaries. In *Proceedings of COLING-1992*, Nantes, France, pp. 546-552.

TEI Consortium, Eds. "9. Dictionaries." TEI P5: Guidelines for Electronic Text Encoding and Interchange. [3.5.0]. [Last modified 29th January 2019]. TEI Consortium. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSFLT (17 June 2019)

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. doi:10.1038/sdata.2016.18.