Strategie e metodi per il recupero di dizionari storici

Eva Sassolini[1], Marco Biffi[2,3]
[1]Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa
[2]Accademia della Crusca, Firenze
[3]Università degli Studi di Firenze
eva.sassolini@ilc.cnr.it marco.biffi@unifi.it

Abstract

English. The article describes ongoing work on the digitization of an authoritative and historically important Italian dictionary, namely Il Grande Dizionario della Lingua Italiana (GDLI) of S. Battaglia, with a focus on the stages of the conversion of this text into structured digital data. We report on the preliminary results of a collaboration between the Accademia della Crusca and Istituto di Linguistica Computazionale "A. Zampolli", which aims to extract the contents of the GDLI to convert them into structured digital data for human use, and/or to be integrated with other language resources, both dictionaries and corpora. The extraction process is articulated on the one hand in the definition of data extraction procedures, on the other hand in the adoption of strategies aimed at supporting the correction of errors.

Italiano. L'articolo descrive un approccio sperimentale all'estrazione, da formato digitale non standard, della completa struttura delle entrate lessicali del Grande Dizionario storico della Lingua Italiana (GDLI) di S. Battaglia. Sono riportati i risultati preliminari di una collaborazione tra l'Accademia della Crusca e Istituto di Linguistica Computazionale "A. Zampolli" del CNR, che mira a convertire i contenuti testuali in dati digitali strutturati per offrirli alla consultazione e allo studio degli utenti e/o per la successiva integrazione con altre risorse linguistiche, sia dizionari che corpora. Il processo di estrazione si articola da un lato nella definizione di procedure di estrazione dei dati, dall'altro nell'adozione di strategie finalizzate al supporto alla correzione degli errori.

1 Introduzione

Il progetto, nato per strutturare l'intero elenco di voci del dizionario GDLI[1], ha richiesto un articolato procedimento di estrazione, data la complessità dei dati e la disponibilità di un formato digitale non standardizzato. Il testo digitale da cui siamo partiti era costituito da un formato Word parzialmente strutturato, ottenuto sottoponendo l'originale cartaceo a procedure di OCR[2], senza nessun tipo di collazione, parziale o totale. Il processo di acquisizione ha evidenziato caratteristiche stilistiche e scelte di layout derivate dall'originale che hanno reso l'OCR estremamente complicato. La versione edita presenta una suddivisione della pagina in 3 colonne, un colore della carta non sufficientemente bianco, nonché un carattere tipografico relativamente piccolo e una altrettanto minima interlinea. Per ragioni legate a tempo e costi dell'impresa non abbiamo potuto migliorare la qualità dell'OCR, almeno in questa fase del progetto, come viene attualmente proposto in letteratura nei nuovi approcci. Nel caso specifico, utilizzando tecniche di pre- e/o post-elaborazione dell'output eseguita attraverso l'uso di un singolo o più motori di OCR. Inizialmente abbiamo valutato l'utilizzo di sistemi di estrazione automatici, sia basati su regole che su tecniche di *machine learning* (Khemakhem et al. 2017) ma l'analisi dei dati ha escluso l'opzione. La complessità strutturale non è l'impedimento maggiore, più rilevante è il numero e la varietà degli errori, a cui si aggiunge la mancanza di un training corpus opportuno per l'addestramento. Tutto questo ci ha spinto verso un approccio sperimentale, basato su strategie di definizione di regole di estrazione dal formato Word. Abbiamo inoltre evidenziato una distribuzione non uniforme delle tipologie di errore nei vari volumi, probabilmente influenzata dalla lunga gestazione dell'opera editoriale complessiva (vedi Tab. 1).

[1] *Grande Dizionario della Lingua Italiana*, di Salvatore Battaglia (poi diretto da Giorgio Bàrberi Squarotti), Torino, UTET, 1961-2002, 21 voll.; con *Supplemento 2004*, diretto da Edoardo Sanguineti, Torino, UTET, 2004, e *Indice degli autori citati nei volumi I-XXI e nel Supplemento 2004*, a cura di Giovanni Ronco, Torino, UTET, 2004.

[2] La Optical Character Recognition è una tecnologia che permette di convertire un'immagine PDF o di altro tipo in testo digitale.

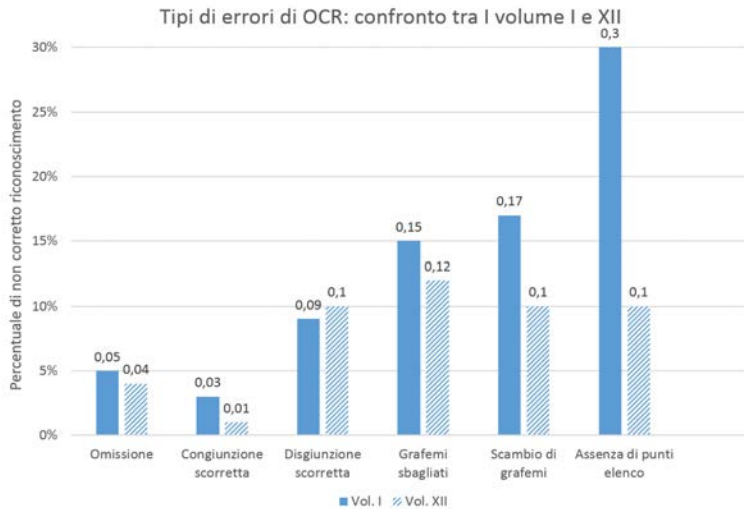

Tabella 1: confronto dei tipi di errore tra volumi diversi

In un lavoro realizzato in un arco temporale di 40 anni, era probabilmente inevitabile la presenza di cambiamenti e aggiustamenti (anche minori), introdotti nel tempo sia a livello delle voci che nel corpus di riferimento del GDLI, che hanno avuto influenze sulle procedure di OCR. Questo contesto ha reso difficile appoggiarsi ad esperienze di altri, pur se indirizzati come noi verso approcci non standard. In alcuni progetti simili si parla di "digitalizzazione attraverso una procedura primitiva" (Bausi, 2016), ma ci si appoggia poi principalmente alla ri-digitazione manuale da parte di studiosi ed esperti qualificati. Nel nostro caso, data la dimensione e complessità dei dati, è necessario limitare il ricorso alla correzione manuale, per le stesse ragioni di contenimento di tempi e costi indicate sopra.

2 L'approccio

Abbiamo impostato un piano di lavoro a lungo termine, che comprendesse 'tappe' da raggiungere progressivamente: 1) riconoscimento del lemma; 2) identificazione di tutti i campi del lemma principale; 3) numero di sensi principali; 4) numero di sensi annidati; 5) campi di ogni senso principale; 6) campi di ciascun senso annidato 7) *mapping* in formato TEI. L'approccio seguito consiste in fasi di riconoscimento successive che, partendo dall'identificazione del lemma dell'entrata lessicale, ne eseguono la segmentazione progressiva individuando, attorno a questo nucleo, gli altri campi dell'entrata. Potremmo definirlo un processo di parsing a più livelli. Ogni campo ha richiesto strategie specifiche per l'identificazione delle caratteristiche distintive, che, tradotte in vincoli di corretta attribuzione e impostati in modo incrementale, hanno portato ad un riconoscimento sempre più granulare della struttura dell'entrata. Oltre a definire procedure software di estrazione e codifica abbiamo implementato metodi di supporto alla correzione manuale e un sistema efficiente di revisione e riallineamento successivo dei dati estratti, per contenere il più possibile l'intervento manuale. L'articolo descrive l'approccio generale e le prime tappe del progetto; la conversione in un formato standard di rappresentazione, pur essendo un impegno rilevante e dall'impatto non trascurabile sul progetto, esula tuttavia dai nostri intenti.

2.1 L'analisi dei dati

Il GDLI è il principale dizionario storico dell'italiano pubblicato da UTET. I 21 volumi che lo compongono, terminati di pubblicare nel 2002, sono corredati da due supplementi integrativi, il primo del 2004 e l'altro del 2009, e da un *Indice degli autori citati*. Recentemente, è stato firmato uno storico accordo tra la UTET e l'Accademia della Crusca, che ha concesso a quest'ultima i diritti per un'edizione elettronica dell'opera, destinata alla consultazione gratuita. Dal maggio 2019 è quindi possibile consultare e interrogare il GDLI con un motore di ricerca per forma applicato al testo in formato Word sopra citato (www.gdli.it). Per quanto il testo elettronico presenti molte debolezze, l'approdo finale di ogni ricerca è la riproduzione in immagine dell'originale a cui si rimane del tutto fedeli, anche in questa edizione, consentendone una comoda lettura con l'ingrandimento a video, a differenza della versione cartacea in cui le dimensioni ridotte dei caratteri non permettono un facile accesso. Nella ricerca si possono certamente perdere alcuni risultati di forme "occultate" dagli errori di OCR ma, una volta arrivati alla pagina, il consultatore può attingere appieno a tutte le preziose informazioni del dizionario. Questa rappresenta soltanto una fase iniziale del progetto: il contributo scientifico dell'ILC si inserisce a fronte dell'esigenza di fornire un accesso più articolato alle informazioni. Grazie a

storiche esperienze nella lessicografia computazionale (Calzolari et al., 1987; Calzolari et al., 1993) stiamo stati coinvolti per implementare il complesso processo di estrazione e riconoscimento della struttura delle entrate. L'analisi dei dati ha evidenziato un input costituito da oltre 23.000 pagine di testo, rappresentate in un formato Word contenente diverse tipologie di errore. Come affermato nell'introduzione, il testo cartaceo originale presenta caratteristiche stilistiche e scelte di layout che hanno condotto il sistema di OCR verso inevitabili problemi di corretta interpretazione. Gli errori di riconoscimento sono stati analizzati su ogni singola caratteristica strutturale del dizionario: lemma, varianti ortografiche, categoria grammaticale, codici d'uso, definizione, etimologia, sensi principali e sensi aggiuntivi (annidati).

N.	Originale cartaceo	Testo OCR
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. (*assólo* o *assuólo*). Disporre a strati. = Deriv. da *suolo* (v.).	**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. (*assólo* o *assuólo*). 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.	Ammacchiate[1]) rifl. (*m'ammàcchio, Vammacchi*). Raro. Nascondersi nella macchia. *B. Davanzati*, 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.¶

Tabella 2: esempi di errori del sistema di OCR

Ciascuno dei campi presenta errori di vario tipo, che vanno dalla mancata segmentazione dei paragrafi, all'interpretazione errata della punteggiatura e dell'ortografia delle parole, al mancato rispetto delle diverse sezioni della voce del dizionario: punti elenco, rientro, dimensione del carattere ecc. (vedi Tab. 2). La presenza di errori ha assunto quindi un peso decisivo nel progetto e ha mostrato come le sole procedure automatiche, per quanto raffinate e puntuali, non sarebbero state sufficienti a produrre un risultato corretto.

2.2 Le fasi di lavoro

Siamo partiti da una sommaria classificazione dei problemi relativi all'inesattezza del dato distinguendo tra errori "bloccanti" e "non bloccanti", per poi procedere con i casi più specifici. La differenza sta nell'impatto dell'errore sulla procedura di *parsing* dei dati. Gli errori bloccanti sono costituiti prevalentemente dal mancato riconoscimento di un nuovo lemma. In questo caso, non potendo chiudere correttamente la voce precedente, si inficia il successivo processo di raffinamento, impedendo la definizione dei confini e campi dell'entrata (vedi Tab. 3).

Errore	Tipo	Correttivo	Esempi
Ortografico in lemma	Non bloccante	Riferimento con indicazione di vol. e pagina nel file di report	
Segni di puntegiatura con funzione di separatori tra campi	Non bloccante	Correzione automatica dei dati e rif. puntuale nel file di report	
Omissione	Bloccante	Non risolto	
Lemma non trovato ad inizio paragrafo	Bloccante	Evento comunque individuato e riportato nel file di report	(vedi Tab. 2. n.4)

Tabella 3: alcuni tipi di errore nel lemma

Un errore "non bloccante" interviene invece quando non è possibile separare il codice grammaticale, da quello d'uso, e/o dalle varianti ortografiche e/o queste dalla definizione. Questa tipologia di errori producono

un'entrata non corretta, ma sulla quale si possono impostare le successive fasi di raffinamento progressivo, procedendo in un certo senso a 'tappe' nella strutturazione della voce. Mentre per gli errori "bloccanti" non abbiamo trovato un'efficiente soluzione alternativa alla revisione manuale post-processing, per gli altri è possibile corredare il parser di meccanismi di annotazione puntuale. Segnalare quando mancano campi obbligatori o se il loro ordine non è rispettato, e riferendo puntualmente il caso in un file di report. In alcuni casi, quando è possibile impostare un'indagine più puntuale del dato, i file di report sono finalizzati al controllo delle soluzioni già inserite in fase di parsing, così da alleggerire il lavoro di revisione manuale.

3 Prospettive

Nelle fasi successive del progetto le risorse estratte hanno assunto una valenza autonoma, per esempio abbiamo prodotto un confronto tra i lemmi del GDLI e quelli del TLIO[3]: il primo dizionario storico di tutte le varietà dell'italiano antico fino al 1375. Stiamo pensando di allargare il confronto anche ad altri dizionari, primo fra tutti il Dizionario Macchina dell'Italiano (DMI) che è patrimonio di storiche linee di ricerca dell'ILC. Nel recente passato le ricerche nel settore si sono concentrate principalmente sullo sviluppo di lessici computazionali in applicazioni di elaborazione del linguaggio naturale, ma oggi i metodi e le tecniche sviluppati per estrarre, strutturare e rappresentare dizionari, possono avere un ruolo potenziale per la progettazione e costruzione di risorse orientate all'uomo, nelle attività lessicografiche dell'editoria, soprattutto digitale. I dizionari storici sono in grado di documentare l'evoluzione diacronica della lingua, mostrando la dimensione storica del lessico. I potenziali vantaggi della digitalizzazione e strutturazione di un dizionario monumentale come il GDLI risiedono anche nell'importanza delle citazioni che vi si possono consultare. Come sostenuto da Beltrami e Fornara (2004), il vero fulcro del dizionario è la presenza massiccia di citazioni di testo, che coprono un'ampia varietà di usi linguistici, dalla lingua quotidiana e letteraria, alle lingue regionali e/o specializzate/specialistiche, ai neologismi e alle parole straniere. Le citazioni offrono preziose informazioni sulle prime attestazioni delle parole, sulle loro varianti formali/diacroniche/diatopiche; sugli autori che le citano e sulle loro etimologie. Per questo motivo stiamo implementando procedure software che da un lato estraggano le varianti dalla struttura della voce e dall'altro, attraverso l'elaborazione delle informazioni estratte dal volume dell'indice degli autori citati, consentano di predisporre filtri su autore ed epoca/data per le rispettive citazioni.

4 Conclusioni

Il nostro impegno è finalizzato a rendere una delle maggiori risorse lessicografiche dell'italiano utilizzabile per il trattamento computazionale, ma l'analisi conclusiva dell'approccio adottato è ancora prematura, soprattutto per quanto riguarda l'estrazione dei sensi annidati. A progetto in corso un'analisi conclusiva del lavoro non è possibile, tuttavia ci sembra di comune utilità descrivere la nostra esperienza, come aiuto per pianificare progetti analoghi, per i quali mancano riferimenti certi in letteratura. Questi progetti, avendo un alto grado di complessità e di incognite, si sviluppano troppo spesso senza un'adeguata divulgazione, il che significa che spesso i ricercatori e gli studiosi devono in un certo senso "reinventare la ruota". L'intento di questo articolo è proporre il nostro approccio come *caso di studio* in contesti in cui non è possibile ricorrere a strumenti e/o procedure consolidate o sperimentali già note in letteratura e magari offrire spunti per discutere delle strategie specifiche che sono state utilizzate.

References

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

Khemakhem M., Herold A., Romary L. 2018. *Enhancing Usability for Automatically Structuring Digitised Dictionaries.* GLOBALEX workshop at LREC 2018. May 2018. Miyazaki. Japan.

Agostiniani L., Montemagni S., Paoli M., Picchi E. 2004. *Lessicografia dialettale e computer: questioni di rappresentazione e recupero dei dati*. Centro Interuniversitario di Studi Veneti, Venezia (Italia).

Beltrami P.G., Fornara S. 2004. *Italian historical dictionaries: from the Accademia della Crusca to the web.* In «International Journal of Lexicography». vol. 17. n. 4. pp. 357-384.

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

Grande Dizionario della lingua italiana. Opera diretta da Salvatore Battaglia. Torino. UTET. 1961-2002.

Calzolari N., Hagman J., Marinai E., Montemagni S., Spanu A., Zampolli A. 1993. *Encoding Lexicographic Definitions as Typed Feature Structures*. In: F. Beckmann, G. Heyeder (eds.). Theorie und Praxis des Lexikons. Beiträge zu einem Kolloquium über theoretische Lexicologie und praktische Lexikographie. Walter de Gruyter. Berlin. pp. 274-315.

Monachini M., Picchi E. 1993. *Computational lexicography: a query system for text corpora.*

Calzolari N., Picchi E. 1988. *Acquisition of semantic information from an on-line dictionary*. (1988). Proceedings.

Calzolari N., Picchi E., Zampolli A. 1987. *The Use of Computers in Lexicography and Lexicology*. (1985). Proceedings.

Calzolari N. 1984, "*Detecting Patterns in a Lexical Database".* Proceedings of the 10th International Conference on Computational Linguistics. Stanford. California. pp. 170-173.

TEI Consortium. eds. *"9. Dictionaries."* TEI P5: Guidelines for Electronic Text Encoding and Interchange. [3.5.0]. [Last updated on 16th July 2019]. TEI Consortium. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSFLT (30/10/19)