Versione PROVVISORIA del contributo presentato al Convegno Annuale

La digitalizzazione del *GDLI*: un approccio linguistico per la corretta acquisizione del testo?

Eva Sassolini[1], Marco Biffi[2,3], Francesca De Blasi[1], Elisa Guadagnini[1], Simonetta Montemagni[1]
[1] Istituto di Linguistica Computazionale "A. Zampolli" (CNR – Pisa), Italia – {eva.sassolini, elisa.guadagnini, francesca.deblasi, simonetta.montemagni(«»)ilc.cnr.it
[2] Accademia della Crusca, Firenze, Italia
[3] Università degli Studi di Firenze, Italia – marco.biffi(«»)unifi.it

ABSTRACT
In questo articolo sono discussi metodi e strategie in via di elaborazione per la correzione (propedeutica alla successiva strutturazione) dei contenuti del *Grande dizionario della lingua italiana* (*GDLI*) fondato da Salvatore Battaglia, estratti da un formato digitale non standard. La presenza, in questo formato, di errori distribuiti di vario tipo ha condizionato la scelta dell'approccio all'estrazione e messo in luce tutte le difficoltà dell'operazione. Le sperimentazioni fatte sino a oggi portano a privilegiare una strategia di correzione multilivello, che procede scomponendo in sezioni distinte l'individuazione e la correzione degli errori, in modo da rendere gestibili interventi complessi di correzione semi-automatica, altrimenti improponibili, e consentire un loro raffinamento progressivo. Parallelamente alla definizione di regole di riconoscimento di struttura e formato, stiamo analizzando metodi e procedure in grado di migliorare la qualità dell'input e specializzare i moduli di estrazione per i singoli campi della voce a partire dal "lemma". Le finalità del lavoro sono duplici: l'estrazione e strutturazione dei contenuti e la produzione di un formato standard di rappresentazione dei dati. Si tratta di un percorso difficile perché il formato dei dati rende l'uso di strumenti reperibili in letteratura non applicabile. Solamente al termine del lavoro potremo capire se esistono le condizioni per trasformare l'approccio adottato in un protocollo di intervento replicabile.

PAROLE CHIAVE
Dizionari digitali, risorse linguistiche, estrazione dell'informazione, riconoscimento di errori, correzione del testo post OCR.

1. INTRODUZIONE
Il *Grande dizionario della lingua italiana* (*GDLI*) è uno strumento lessicografico monumentale, che descrive dettagliatamente la semantica e gli usi delle parole ricorrendo in particolare a una ricchissima esemplificazione tratta dai testi italiani (soprattutto letterari) dalle Origini al XX secolo[1]. I potenziali vantaggi della digitalizzazione e strutturazione del *GDLI* sono chiari agli studiosi: il vocabolario costituisce un ricchissimo corpus diacronico dell'italiano e potrebbe supportare analisi raffinate, che si gioverebbero moltissimo delle possibilità offerte da una consultazione anche non "lineare" (che è l'unica possibile su carta). L'urgenza dell'intervento è quindi nota ma, non avendo potuto migliorare la qualità dell'Optical Character Recognition (OCR) avvalendoci di tecniche di pre- e/o post-elaborazione dell'output, come viene attualmente proposto in letteratura [1], l'input del lavoro di estrazione e strutturazione è oggi costituito da oltre 23.000 pagine di testo in formato digitale non standard, prodotte da procedure di OCR in cui l'originale cartaceo non è stato sottoposto a nessun tipo di collazione, parziale o totale. L'acquisizione risultante, a causa delle caratteristiche particolarmente complesse della stampa[2], risulta quindi imperfetta tanto a livello puramente testuale (corretto riconoscimento dei caratteri) quanto riguardo alla segmentazione del testo (corretta individuazione delle voci e delle parti della voce). La situazione è ulteriormente complicata dal fatto che il *GDLI* unisce alla forte problematicità formale l'estrema variabilità sostanziale, dovuta al fatto che la lingua del testo accetta – data la

[1] La pubblicazione del *GDLI* è iniziata nel 1961 e finita nel 2002, presso la casa editrice torinese UTET, per cura di Salvatore Battaglia (per i primi sei volumi) e poi di Giorgio Barberi Squarotti. Ai 21 volumi che compongono l'opera si sono aggiunti due supplementi, pubblicati rispettivamente nel 2004 e nel 2009 (sotto la direzione di Edoardo Sanguineti).

[2] Come descritto in [1] [2], pongono grosse difficoltà a una lettura OCR il formato di stampa (i volumi di grande dimensione con testo distribuito su tre colonne; il corpo dei caratteri estremamente piccolo per le parti di definizione ed etimologia, e ancora minore per gli esempi citati, dati l'uno di seguito all'altro separati dal solo "punto fermo"; la presenza di grandi porzioni di testo che si distribuiscono di seguito con pochissimi a capo), la qualità non uniforme della carta, la presenza di caratteri speciali, i tratti particolarmente sottili dello stile corsivo, ecc.

massa degli esempi citati e la loro appartenenza a una diacronia lunghissima – una profonda variabilità sia grafico-fonetica sia morfosintattica. In questo contesto, il progetto di digitalizzazione strutturata è iniziato con molte incognite e una lunga strada davanti. L'individuazione e la correzione degli errori OCR di tipo ortografico è un problema noto e studiato nella comunità NLP (Natural Language Processing) sin dagli anni Settanta. Una rassegna dei primi lavori su questo problema è descritta da Kukich (1992) [3]. In passato la maggior parte dei tradizionali sistemi d'indagine sugli errori OCR si concentrava sulla costruzione di complesse "matrici di confusione" di caratteri (coppie) per rilevare la presenza di parole inesistenti; più recentemente l'utilizzo di informazioni sul contesto linguistico in cui l'errore appare migliora sperimentalmente la precisione (Evershed e Fitch, 2014) [4]. Un esempio di questo tipo è quello proposto da Bassil *et al*. (2012) per l'inglese, utilizzando modelli di n-grammi [5] o da Kissos *et al*. (2016), in cui si valuta l'impatto relativo di diverse fonti di informazione, combinando caratteristiche di modelli linguistici, informazioni sul processo OCR e informazioni sul contesto del documento [6]. In un ambito più affine si collocano esperienze che hanno affrontato output OCR di lessici storici trovando metodi per rilevare e rappresentare automaticamente il contesto semantico di un testo (Wick *et al*. 2013) [7], sul modello già adottato da Zhang e Chang (2003) che utilizza una combinazione lineare di modelli linguistici per correggere gli errori [8]. Anche riguardo l'estrazione della struttura abbiamo valutato altre esperienze che cercano come mappare le informazioni tipografiche ottenute sull'output di testo riconosciuto dall'OCR (Reul *et al*. 2019) [9]. Partendo da queste esperienze possiamo vedere il nostro lavoro come un altro esperimento in un complesso tentativo di correzione/strutturazione automatica dei risultati di OCR con errori su dati di rilevanza storico-linguistica. È importante rilevare che, se è vero che molti grandi vocabolari delle lingue occidentali redatti prima del XXI secolo sono oggi disponibili in una versione elettronica, va però sottolineato che essi sono stati acquisiti – o per lo meno integralmente corretti – manualmente, grazie alla ricopiatura/revisione del testo su supporto digitale da parte di operatori umani[3]. La correzione manuale pone a sua volta una serie di problemi, poiché i correttori fatalmente sbagliano, non correggendo errori già presenti nel testo come anche introducendone di nuovi (che, a quel punto, saranno aleatori e dunque assai difficilmente reperibili). Inoltre, prevedere anche soltanto una rilettura integrale di opere vaste come un grande vocabolario comporta tempi lunghi e un non trascurabile impegno economico.

2. L'APPROCCIO LINGUISTICO ALLA CORREZIONE

Con la finalità di arrivare alla digitalizzazione del *GDLI* con un input di questo tipo, la domanda che ci si pone è: è possibile impiegare intelligenza linguistica per mettere a punto una strategia di correzione efficace, che cerchi di ridurre al minimo la necessità di interventi manuali? Questa sfida riguarda l'estrazione e strutturazione dei contenuti e la loro conversione in un formato standard di rappresentazione, che rispetto alla digitalizzazione rappresenta una base di partenza necessaria. La definizione di un approccio e di norme (buone pratiche) di intervento che possano essere riutilizzati in casi analoghi è un obiettivo di lungo termine di cui ora si possono solo intravedere le basi. L'idea è nata due anni fa e oggi l'approccio adottato può già avvalersi di un insieme di regole e procedure software messe a punto nelle prime sperimentazioni, in grado di produrre una segmentazione progressiva del testo e in cui l'identificazione delle caratteristiche distintive è tradotta in vincoli di corretta attribuzione e impostata in modo incrementale all'interno del processo di parsing del testo digitale [1][2]. Queste prime sperimentazioni sono state fondamentali per individuare le strategie più adatte ad affrontare un lavoro così complesso, iniziando con una prospettiva "a grana grossa", di macro-interventi a tutto testo, gli errori sono stati suddivisi in "bloccanti" e "non bloccanti"[4]. Successivamente l'analisi dei risultati ha mostrato come questa gerarchia non fosse funzionale al lavoro di recupero puntuale di alcune porzioni strutturate di testo e si è manifestata l'esigenza di implementare anche micro-interventi. A fronte dell'estrema complessità del compito, si è deciso di tentare la strada di una strategia di correzione multilivello, che sfrutti il più possibile criteri di tipo linguistico.

In generale, un testo fortemente strutturato quale è un dizionario presenta "campi testuali" ben distinti, che hanno al loro interno delle omogeneità sfruttabili per la correzione: così, per esempio, il campo "lemma" è caratterizzato da una bassa variabilità morfologica (e per questo motivo si può mettere in relazione con una serie di risorse intra- ed extra-testuali,

[3] Per citare soltanto alcuni casi notevoli, è stato così per le prime quattro impressioni del *Vocabolario* della Crusca (http://www.lessicografia.it/), per il "Littré électronique" e il *Trésor de la langue française informatisé* (http://atilf.atilf.fr/), per l'*Oxford English Dictionary*.

[4] I primi impediscono la definizione dei confini della voce, inficiando in modo grave il processo di estrazione successivo; i secondi producono un'entrata con errori di vario tipo.

per cui vedi *infra* §4). D'altra parte, l'acquisizione OCR produce errori che sono trasversali a tutti i campi, che ricorrono "a tutto testo" con le medesime caratteristiche e possono essere ricondotti a due tipologie fondamentali (vedi tab. 1).
In questa fase la priorità è l'estrazione del lemmario corretto, per questo ci siamo concentrati su due livelli di correzione: la correzione degli errori "a tutto testo", con la finalità di ottenere una ripulitura generale del testo acquisito da OCR; la messa a punto di strategie specifiche di correzione del campo "lemma". Gli interventi correttivi si collocano in fasi e tempi diversi del processo di estrazione: il formato word del testo per gli errori a tutto testo; i testi dei campi della voce in fase di parsing, valutando la combinazione di diverse caratteristiche (per esempio nel caso del lemma si valuta la presenza di particolari suffissi in rapporto alla categoria grammaticale); i dati strutturati una volta estratti. In quest'ultimo caso l'approccio alla correzione può essere anche molto complesso e utilizzare più risorse linguistiche come lessici e liste di forme per le ipotesi di correzione.

1) cattiva segmentazione	1.1) introduzione di spazi bianchi dove non ci sono
	1.2) univerbazione di parole separate
2) grafie scorrette	2.1) omissione di singoli caratteri o sequenze
	2.2) cattivo riconoscimento di singoli caratteri o di sequenze di caratteri
	2.3) cattivo riconoscimento di caratteri che produce l'inserimento di caratteri aggiuntivi

Tabella 1: tipologie di errori

3. STRATEGIE DI CORREZIONE DEGLI ERRORI "A TUTTO TESTO"

Delle due tipologie di errore "a tutto testo", tralasciamo per il momento il tipo 1), la "cattiva segmentazione", che richiede strategie di correzione diversificate a seconda del campo della voce (vd. *infra* §4 per la correzione di questo errore nel campo "lemma"). Concentriamoci quindi sugli errori di tipo 2), vale a dire le grafie scorrette: che si tratti della fattispecie 2.1), 2.2) o 2.3) questi errori possono essere individuati (ed eventualmente corretti automaticamente) a diversi "livelli di correzione", a seconda dell'area di testo in cui si situa l'intervento correttivo richiesto:

LIVELLO DI CORREZIONE	AREA DI TESTO INTERESSATA DALL'ERRORE	DESCRIZIONE DELL'ERRORE	STRATEGIA DI INDIVIDUAZIONE/CORREZIONE DELL'ERRORE
livello 0	singolo carattere	presenza di un carattere non alfabetico (o non appartenente all'alfabeto italiano)	● analisi degli errori della "sezione *gold*" (sostituzioni 1:1)
livello 1	sequenza di 2 o più caratteri (ogni carattere di per sé è ammissibile)	sequenza di caratteri non ammessa in lingua italiana	● analisi degli errori della "sezione *gold*" ● "criterio fonotattico"
livello 2	parola (caratteri e sequenze di caratteri di per sé sono ammissibili)	parola errata	● analisi degli errori della "sezione *gold*" ● confronto con altri *corpora*

Un primo strumento che abbiamo utilizzato per la correzione di errori grafici è, come si vede dalla tabella, la "sezione *gold*". Essa consiste in uno specimen di *GDLI*, pari a 30 colonne complessive (scelte aleatoriamente), corretto a mano[5] e impiegato per reperire le corrispondenze univoche tra un riconoscimento OCR errato e una correzione, rispetto a uno o

[5] La correzione è stata fatta da Cecilia Palatresi e Silvia Dardi, collaboratrici dell'Accademia della Crusca.

più caratteri: la "sezione *gold*", quindi, non ha rilevanza statistica e non è servita per l'analisi degli errori, ma soltanto per una ripulitura del dettato rispetto agli errori univoci e ricorrenti. Dato che alcuni di questi errori interessano marche d'uso, grammaticali e semantiche, la loro correzione è funzionale tanto per la successiva strutturazione del testo, quanto per l'implementazione delle possibilità di ricerca trasversale del vocabolario. Sono esempi di questo tipo di errore la forma "aw." per l'abbreviazione "avv." nel *GDLI* a stampa; lo stesso vale per "medie." per "medic.", "pass," per "pass." (nell'abbreviazione "part. pass."), ecc. La sezione *gold* ha permesso anche la correzione di altri errori ricorsivi, quali il nesso "-z?-" per "-zz-", ecc.

Per l'individuazione degli errori di livello 1) abbiamo elaborato anche il "criterio fonotattico", che permette al sistema di isolare come erronee le parole che contengono i nessi consonantici (di 2 e 3 caratteri) non ammessi in italiano (cfr. [10] [11]). In questo caso la correzione non è automatica, ma sono segnalate alla correzione manuale singole parole certamente errate: l'applicazione del "criterio fonotattico" individua come sbagliata, per esempio, la forma "impcrvietà" (poiché il nesso triconsonantico -pcr- non è ammissibile in lingua italiana), che dovrà essere corretta in "impervietà".

Per la gestione degli errori di livello 2), che sono evidentemente i più complicati da individuare, è di grande ausilio il ricorso a risorse esterne, quali le banche dati testuali esistenti e i vocabolari digitalizzati (per un'applicazione specificamente al campo lemma cfr. §4).

4. STRATEGIE DI CORREZIONE DEGLI ERRORI NEL CAMPO "LEMMA"

Accanto alle strategie di correzione già descritte, abbiamo messo a punto una serie di procedure mirate per il campo "lemma". Va premesso che si è deciso di riprodurre il lemmario del *GDLI* cartaceo rispetto alla sequenza dei caratteri, facendo astrazione dai diacritici: le entrate presentano l'indicazione dell'apertura/chiusura delle "e" e "o" toniche e l'indicazione della natura sorda/sonora delle affricate e sibilanti. Posto che il cattivo riconoscimento del diacritico è un errore ricorsivo dell'OCR, che il diacritico utilizzato per le consonanti (un puntino sottoscritto) rappresenta un carattere non standard, e che la presenza dei diacritici ostacola il confronto con altri lemmari, si è deciso di non gestire questo ordine di informazioni[6].

Tornando, quindi, ai livelli di correzione degli errori (tab. 2), posto che gli errori di livello 0 sono gestiti interamente dalle strategie di correzione "a tutto testo", abbiamo messo a punto delle strategie specifiche per i livelli di errore 1 e 2, per la cui gestione possiamo quindi contare sui risultati del *parsing*; abbiamo inoltre pensato a dei criteri per individuare e correggere gli errori di cattiva segmentazione (tab. 1).

Per gli errori di livello 1, abbiamo ideato una strategia di correzione che si è rivelata efficace per gli errori di sequenza a inizio di parola. Poiché i lemmi devono presentarsi in sequenza alfabetica, il mancato rispetto dell'ordinamento alfabetico è un sicuro segnale di errore. Uno strumento di correzione valido consiste nel confronto del lemma con l'intervallo alfabetico segnalato dalle "testatine" (vale a dire le entrate della prima voce della prima colonna e dell'ultima voce della terza colonna, segnalati nell'intestazione di ogni pagina del *GDLI*): grazie a questo confronto è possibile formulare delle ipotesi di correzione della parte iniziale del lemma – è così possibile per esempio correggere l'errato "bfdicotomia" in "broncotomia".

Per gli errori, sempre di livello 1, di sequenza a fine di parola, abbiamo pensato di testare un "criterio morfologico": incrociando le terminazioni ammissibili in italiano per le diverse categorie grammaticali, sarà possibile individuare errori di terminazione e, in alcuni casi, formulare ipotesi di correzione (ad es., per correggere "chetonemla" in "chetonemia").

Per gli errori di livello 2 e gli errori di tipo 1) (cattiva segmentazione), un ulteriore strumento di correzione del lemmario è la collazione con i lemmari di altri vocabolari italiani: stiamo testando in particolare il confronto con il lemmario del Tommaseo-Bellini (di cui il *GDLI* costituisce l'esplicito "aggiornamento" e completamento). "Normalizzando" anche le entrate del T-B (cioè eliminando gli accenti) e considerando le sole entrate monorematiche, i risultati della collazione sono promettenti: il confronto consente la correzione dei lemmi del *GDLI* in cui sono scorretti singoli caratteri, ma anche dei lemmi in cui sono stati omessi singoli caratteri o che sono segmentati scorrettamente.

5. CONCLUSIONI

Il progetto complessivo di digitalizzazione e strutturazione è ancora in fase di realizzazione delle risorse e delle procedure, il livello "case study" non può dirsi concluso. L'inquadramento e la definizione della direzione del lavoro hanno avuto una lunga gestazione ma sono ora maturi per essere affrontati in modo sistematico. Questa consapevolezza

[6] Ma sarà naturalmente accentata la vocale finale degli ossitoni, conformemente all'ortografia italiana.

ha permesso, nel giugno 2020, di aggiungere al gruppo di lavoro il progetto *TrAVaSI*[7], che, con l'apporto di diverse competenze specifiche, offrirà al progetto di estrazione un respiro più ampio, sfruttando l'interdisciplinarietà degli ambiti di competenza per specializzare i criteri di estrazione per tutti i fenomeni che lo richiedono. Allo stato attuale l'indagine preliminare sulla qualità del testo OCR di partenza – che ricordiamo non essere descrittiva ma finalizzata unicamente alla realizzazione di procedure d'intervento correttivo quanto più efficaci possibile – può dirsi conclusa. Così anche la fase di lavori preparatori volti ad avviare la correzione assistita del lemmario: sulla base delle evidenze emerse nello studio della sezione *gold* e dei criteri di individuazione degli errori sopra descritti, è stata estratta la lista di lemmi ed è stato stilato un preciso protocollo di correzione automatica e manuale che si avvale di un opportuno strumento software di supporto dedicato.

BIBLIOGRAFIA

[1] Sassolini E., Biffi M., Strategie e metodi per il recupero di dizionari storici. Conferenza AIUCD 2020, 15-17 gennaio 2020, Università Cattolica del Sacro Cuore. Milano. ISBN 978-88-942535-4-2. DOI 10.6092/unibo/amsacta/6316. In: Quaderni di Umanistica Digitale. (2020): 235-239.

[2] Sassolini, E., Khan, A. F., Biffi, M., Monachini, M., Montemagni, S., Converting and structuring a Digital Historical Dictionary of Italian: a case study. 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. eds. (2019): 603-621.

[3] Kukich, K. Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR). (1992). 24(4):377-439.

[4] Evershed, J., Fitch, K. Correcting noisy OCR: Context beats confusion. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. ACM. (2014): 45-51.

[5] Bassil, Y., Alwani, M. OCR context-sensitive error correction based on google web 1t 5-gram data set. American Journal of Scientific Research. (2012).

[6] Kissos, I., Dershowitz, N. OCR error correction using character correction and feature-based word classification. In Proceedings of the 12th IAPR International Workshop on Document Analysis Systems (DAS2016). Santorini. Greece. (2016).

[7] Wick, M., Ross, M., Learned-Miller, Erik. Context-Sensitive Error Correction: Using Topic Models to Improve OCR. Document Analysis and Recognition, International Conference on. 2. (2007): 1168-1172.

[8] Zhang, D., Chang, S. A Bayesian framework for fusing multiple word knowledge models in videotext recognition. In IEEE Conference on Computer Vision and Pattern Recognition. (2003)

[9] Reul, C., et al. Automatic Semantic Text Tagging on Historical Lexica by Combining OCR and Typography Classification: A Case Study on Daniel Sander's Wörterbuch der Deutschen Sprache. DATeCH2019. (2019).

[10] Chiari, I. La fonotassi statistica dell'italiano e del tedesco: i nessi consonantici. In T. De Mauro e I. Chiari (a cura di), Parole e numeri: analisi quantitative dei fatti di lingua. Roma. Aracne. (2005): 67-84.

[11] Muljačić, Ž. Fonologia della lingua italiana. Bologna. il Mulino. (1972).

[7] Progetto POR FSE 2014 - 2020, Asse A Occupazione - Priorità di investimento A.2 – Obiettivo A.2.1 – Azione A.2.1.7. "ASSEGNI DI RICERCA IN AMBITO CULTURALE" (Bando per progetti congiunti di alta formazione attraverso l'attivazione di assegni di ricerca): il progetto, a cui partecipano ILC e l'Accademia della Crusca, mira a formare giovani ricercatori con professionalità e competenze specifiche.