Il cantiere dell'informatizzazione del *GDLI*: risultati, prospettive, sfide future

Il 2 settembre **2017**, l'**Accademia della Crusca** e la casa editrice **UTET** hanno firmato l'accordo che ha permesso l'avvio dei lavori per la digitalizzazione del più grande vocabolario dell'italiano mai realizzato, il **«Grande dizionario della lingua italiana»** (*GDLI*) ideato da Salvatore Battaglia. Dal 9 maggio **2019**, il *GDLI* è consultabile in rete dagli "Scaffali digitali" del sito dell'Accademia della Crusca o direttamente all'indirizzo <http://www.gdli.it/>.

Dal **2019**, l'**Accademia della Crusca** e il **Cnr-Istituto di Linguistica Com-putazionale** "Antonio Zampolli" lavorano per raffinare l'informatizzazione del *GDLI*, nell'ambito di accordi bilaterali di collaborazione scientifica e del progetto regionale toscano "Trattamento Automatico di Varietà Storiche di Italiano" (*TrAVaSI*, 2020-2022).

L'informatizzazione di un'opera monumentale come il *GDLI* è un'impresa intrinsecamente collettiva, che richiede il concorso di competenze ed esperienze diverse. Fanno parte del **gruppo di ricerca**: Marco Biffi (Accademia della Crusca, Università di Firenze), Sebastiana Cucurullo (CNR-ILC), Silvia Dardi (Accademia della Crusca), Francesca De Blasi (CNR-ILC), Manuel Favaro (CNR-ILC), Elisa Guadagnini (CNR-ILC), Anas Fahad Khan (CNR-ILC), Simonetta Montemagni (CNR-ILC), Cecilia Palatresi (Accademia della Crusca), Elena Pepponi (Università di Firenze), Paolo Picchi (CNR-ILC) ed Eva Sassolini (CNR-ILC).

Di seguito, i risultati della ricerca ottenuti fino a oggi (**febbraio 2024**).

Pubblicazioni

Eva Sassolini, Anas Fahad Khan, Marco Biffi, Monica Monachini e Simonetta Montemagni, *Converting and structuring a Digital Historical Dictionary of Italian: a case study*, in *Electronic lexicography in the 21st century*. Proceedings of the eLex 2019 conference (1-3 October 2019, Sintra), Brno, Lexical Computing, 2019, pp. 603-621.

Eva Sassolini e Marco Biffi, *Strategie e metodi per il recupero di dizionari storici*, in *La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*. Atti del IX Convegno Annuale

AIUCD, ed. Cristina Marras *et alii* [supplemento di «Quaderni di Umanistica Digitale»], 2020, pp. 235-239.

Manuel Favaro, Marco Biffi e Simonetta Montemagni, *Risorse e strumenti per le varietà storiche dell'italiano: il progetto* TrAVaSI, in *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, 2020, <http://ceur-ws.org/Vol-2769/paper_86.pdf>.

Eva Sassolini, Marco Biffi, Francesca De Blasi, Elisa Guadagnini e Simonetta Montemagni, *La digitalizzazione del* GDLI*: un approccio linguistico per la corretta acquisizione del testo?*, in *DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale*. Raccolta degli abstract estesi della 10ª conferenza nazionale (AIUCD 2021), Pisa, 2021, ed. Angelo Mario Del Grosso *et alii*, pp. 159-166, <https://aiucd2021.labcd.unipi.it/book-of-abstracts/>.

Francesca De Blasi e Manuel Favaro, Scheda del Progetto *Trattamento automatico di varietà storiche di italiano (TrAVaSI)*, in *Migrazione linguistica e trasmissione culturale nell'Italia medievale*, ed. Cosimo Burgassi *et alii*, CNR Edizioni [Collana PLURIMI – III], 2021, p. 92, <https://www.cnr.it/sites/default/files/public/media/attivita/editoria/collana_plurimi/PLURIMI_3_2021.pdf>.

Manuel Favaro, Marco Biffi e Simonetta Montemagni, *Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione*, in *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data (JADT22)*, ed. Michelangelo Misuraca *et alii*, Napoli, VADISTAT PressEditor, 2022, pp. 393-399, <https://www.researchgate.net/publication/361924391_Proceedings_of_the_16th_International_Conference_on_Statistical_Analysis_of_Textual_Data_JADT22>.

Marco Biffi, Francesca De Blasi, Manuel Favaro, Elisa Guadagnini, Simonetta Montemagni ed Eva Sassolini, *Parole in rete / reti di parole. Possibili impieghi didattici dei grandi vocabolari storici digitalizzati*, «Italiano a scuola», 4, 2022, pp. 143-188.

Marco Biffi e Elisa Guadagnini, *«Le citazioni riconducono il dizionario nell'ambito della letteratura e della vita»: un primo sguardo d'insieme sui citati del* GDLI, «Studi di Lessicografia Italiana», XXXIX, 2022, pp. 351-386.

Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi e Simonetta Montemagni, *Toward the Creation of a Diachronic Corpus for Italian: a Case Study on the* GDLI, in *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022, ed. R. Sprugnoli *et alii*, Association for Computational Linguistics (ACL), pp. 94-100, <http://www.lrec-conf.org/proceedings/lrec2022/workshops/LT4HALA/pdf/2022.lt4hala2022-1.13.pdf>.

Marco Biffi, Elisa Guadagnini, Simonetta Montemagni ed Eva Sassolini, *Il lemmario del «GDLI»: dati quantitativi e prime osservazioni*, «Studi di Lessicografia Italiana», XL, 2023, pp. 331-351.

Risorse linguistiche e lessicografiche

Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi e Simonetta Montemagni, *TrAVaSI_GDLI-quotation corpus*, Corpus con annotazione morfo-sintattica e lemmatizzazione rivista manualmente, che raccoglie un campione delle citazioni del *GDLI* dalla lingua delle origini fino ai giorni nostri, raggiungibile nel "repository" italiano dell'infrastruttura CLARIN-ERIC al seguente indirizzo: <http://hdl.handle.net/20.500.11752/ILC-984>.

Il «Grande dizionario della lingua italiana», voll. I-XXI. Versione strutturata in XML TEI, rilasciata insieme al lemmario completo e alla versione strutturata del volume con l'Indice degli autori citati. Versione 1.1, a cura di Eva Sassolini, Pisa, CNR-ILC, 2023.

Eventi divulgativi

Pisa Internet Festival 2021 *#phygital*: Marco Biffi, Francesca De Blasi, Manuel Favaro, Elisa Guadagnini, Simonetta Montemagni ed Eva Sassolini, *Parole in rete / Reti di parole. Sviluppi innovativi offerti dai dizionari digitali* (corso di formazione per docenti svolto il 27 novembre 2021), <https://drive.google.com/file/d/1_4NxJx-AC0zCRoE_aERmCMNqxRyvtCNR/view?usp=sharing>.

La primavera della ricerca – Una giornata dedicata alla scienza per i cento anni del CNR (Pisa, 12 maggio 2023): Elisa Guadagnini, Simonetta Montemagni ed Eva Sassolini, Laboratorio aperto *Viaggio nella storia della lingua italiana*, <https://laprimaveradellaricerca.cnr.it/labs/>.

Altri risultati della ricerca

Ai prodotti della ricerca già elencati deve essere aggiunto il contributo metodologico e tecnologico:

- definizione di modelli e strategie per la correzione post-OCR, con particolare attenzione alle sfide poste da testi tipograficamente complessi, al fine di permettere la strutturazione dei contenuti e la loro rappresentazione in un formato *standard* internazionalmente riconosciuto;
- definizione di modelli automatici per l'annotazione morfo-sintattica e la lemmatizzazione di varietà storiche della lingua italiana.

I modelli messi a punto in relazione al *GDLI* non si esauriscono in una dimensione puntuale, ma gettano le basi per l'estensione ad altre opere lessicografiche e a *corpora* testuali che raccolgano varietà di italiano latamente non *standard*.

Vale la pena sottolineare il fatto che le attività di ricerca hanno portato contributi originali, trasferimento e condivisione di conoscenze a diversi livelli: istituzionale, disciplinare e divulgativo. Esse hanno contribuito, inoltre, alla creazione di profili professionali innovativi e fortemente interdisciplinari, derivanti dall'integrazione tra cultura umanistica e conoscenze tecnologiche e applicative, che oggi rappresentano elementi irrinunciabili per un'adeguata valorizzazione del patrimonio culturale.

Prospettive di sviluppo

Cinque anni di lavoro hanno portato a creare i presupposti per una fruizione avanzata dei contenuti del *GDLI*: questo permetterà di fornire agli studiosi strumenti di indagine sempre più accurati ed efficaci, ma anche, in un'ottica cara a un'istituzione culturale di riferimento come l'Accademia della Crusca, di rendere uno strumento fondamentale per l'italiano pienamente e facilmente fruibile da parte del pubblico vasto.

Grazie alla strutturazione dei contenuti, il dizionario si libera dalla camicia di forza dell'ordinamento alfabetico delle voci e si trasforma in una rete di parole complessa e multidimensionale, dove l'accesso per lemma rappresenta solo una delle possibili modalità di esplorazione. È in questa direzione che sta proseguendo il lavoro.