Verso l'implementazione di un sistema di riconoscimento di allusioni al lessico dantesco nelle testimonianze del Lager: il caso d'uso in Voci dall'Inferno

Carla Congiu[1], Angelo Mario Del Grosso[2], Marina Riccucci[3]
[1] Università di Pisa, Italia- c.congiu1@studenti.unipi.it
[2] CNR-ILC, Italia – angelomario.delgrosso@cnr.it
[3] Università di Pisa, Italia- marina.riccucci@unipi.it

ABSTRACT (ITALIANO)

Voci dall'Inferno è un progetto di ricerca dell'Università di Pisa, sviluppato con il supporto dell'Istituto di Linguistica Computazionale "A. Zampolli". L'iniziativa ha due principali obiettivi scientifici: a) digitalizzare il primo corpus di testimonianze non letterarie di deportati sopravvissuti ai campi di concentramento e b) identificare al suo interno la presenza di citazioni e/o allusioni al lessico di Dante (Del Grosso et al.,2024). Al fine di raggiungere questo secondo obiettivo è stato sviluppato un prototipo di applicazione web denominata *Voci dall'Inferno Verse Similarity Search*. Il sistema è progettato per individuare citazioni e allusioni al lessico dantesco mediante approcci computazionali alla ricerca di frasi presenti nelle testimonianze e il confronto di essi con i versi presenti nella *Divina Commedia* di Dante Alighieri. L'applicazione, realizzata in Python, utilizza tecnologie avanzate come Weaviate, una piattaforma open-source per la ricerca vettoriale, e Streamlit, un framework per lo sviluppo di applicazioni web. Basandosi su metriche di *Sentence Similarity*, l'applicazione sfrutta modelli di machine learning per trasformare i testi in rappresentazioni di *embeddings* e in seguito misurarne la similarità. Attualmente l'applicazione non è ancora disponibile per l'uso da parte del pubblico, ciononostante l'infrastruttura di ricerca CLARIN-IT (H2IOSC) è stata contattata per ospitare l'applicazione garantendone accesso e sostenibilità. Una demo sarà predisposta per la conferenza qualora il contributo venisse accettato.
Parole chiave: Sentence Similarity; Sentence Transformers; vector database; embeddings; Voci dall'Inferno

ABSTRACT (ENGLISH)

Toward the implementation of a system for recognizing allusions to Dante's lexicon in Lager testimonies: the Voci dall'Inferno use case. Voci dall'Inferno is a research project by the University of Pisa, developed with the support of the Istituto di Linguistica Computazionale "A. Zampolli". The initiative has two main scientific objectives: a) to digitize the first corpus of non-literary testimonies from concentration camp, and b) to identify the presence of citations and/or allusions to Dante's lexicon within them (Del Grosso et al., 2024). To achieve this second objective, a prototype web application called *Voci dall'Inferno Verse Similarity Search* was developed. The system is designed to detect citations and allusions to Dante's vocabulary through computational approaches by searching for expression within the testimonies and comparing them with verses from Dante's Commedia. The application, built in Python, leverages advanced technologies such as *Weaviate*, an open-source vector search platform, and *Streamlit*, a framework for web application development. Adopting sentence similarity metrics, the application uses machine learning models to transform texts into embedding representations and subsequently measure their similarity. Currently, the application is not yet publicly available. However, the CLARIN-IT research infrastructure (within H2IOSC PNRR project) has been contacted to host the application, ensuring accessibility and sustainability. A demo will be prepared for the conference if the contribution will be accepted.
Keywords: Voci dall'Inferno; Sentence Similarity; Sentence Transformers; vector database; embeddings

1. INTRODUZIONE

Voci dall'Inferno è un progetto di ricerca dell'Università di Pisa, curato da Marina Riccucci (Università di Pisa) dal 2016 con il supporto di Angelo Mario Del Grosso (CNR-ILC, Pisa).
Nel corso degli ultimi anni il gruppo di lavoro ha avviato la digitalizzazione e la codifica, mediante lo standard XML-TEI (Burnard 2014), del primo corpus di testimonianze non letterarie provenienti dai campi di concentramento e ha sviluppato una applicazione web mediante la piattaforma eXist-db per la consultazione e la fruizione del repertorio (Del Grosso et al, 2024b). Il corpus ad oggi consta di 23 testimonianze appartenenti a 18 testimoni (vedi Figura 1) suddivisi in 17 deportati nei Lager, di cui 12 ebrei e 5 internati militari italiani, e 1 perseguitato, vale a dire il partigiano ebreo Emanuele Artom (Del Grosso et al., 2024). Di queste testimonianze oltre al recupero e alla digitalizzazione delle fonti scritte –

quasi tutte conservate presso il centro di Documentazione Ebraica Contemporanea di Milano oppure in fondi privati di familiari ed eredi dei testimoni (ad esempio i diari e le lettere di Nicola Ricci, di Luigi Giuntini, di Emanuele Artom, di Alberto Pacini e Romana Feld) vi sono molteplici testimonianze rilasciate in forma orale spesso conservate in microcassette e supporti non digitali (vedi Figura 2). Attualmente più di 18 ore di parlato sono presenti nell'archivio del progetto - per la maggior parte provenienti dal fondo inedito Segre-Pavoncello (13 testimonianze del fondo Segre-Pavoncello sono state digitalizzate su un totale di circa 25 sopravvissuti).

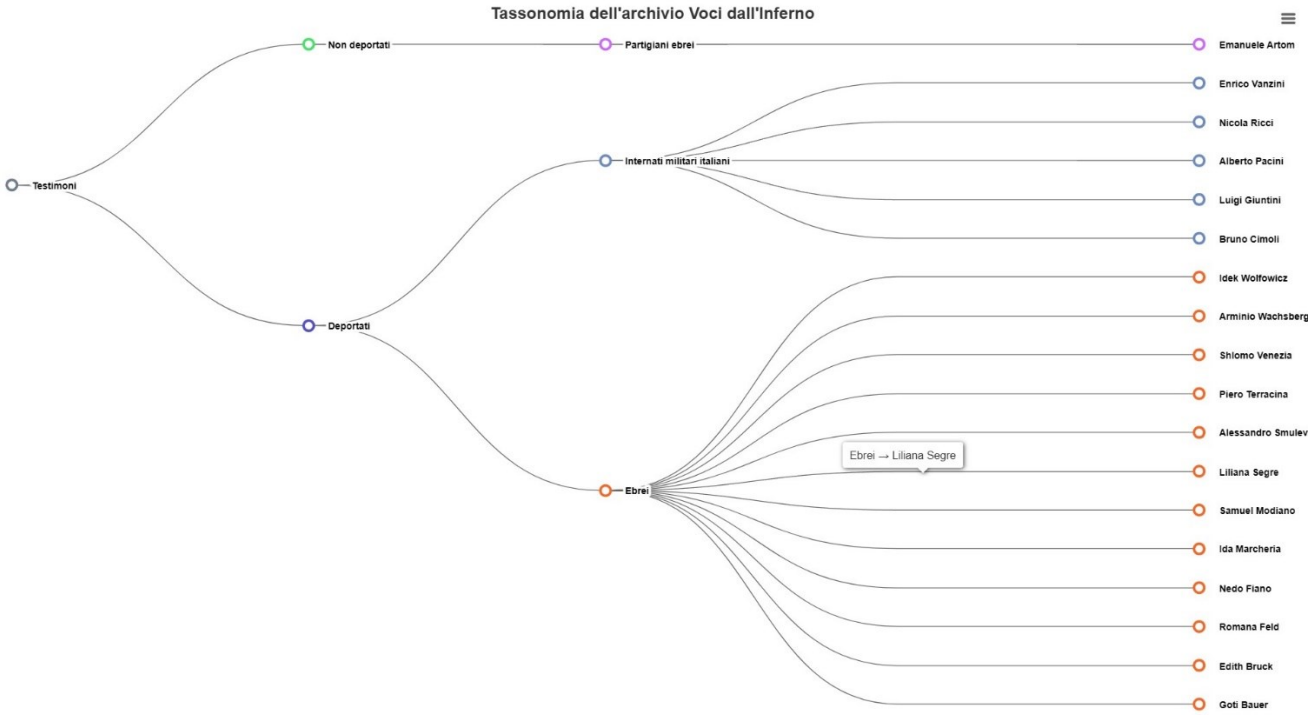

Figura 1. Tassonomia dell'archivio Voci dall'Inferno

Da una prima analisi generale del corpus è emerso che molti testimoni si servono del vocabolario dantesco per riportare le proprie esperienze "ineffabili" (Calderini, 2020). Per studiare le tessere dantesche si sono svolte differenti analisi e una classificazione puntuale in citazioni implicite, citazioni esplicite, allusioni e vocaboli danteschi. I dati che sono emersi dallo studio dell'archivio digitale dimostrano che la maggior parte delle citazioni sono riconducibili alla prima cantica. Nonostante ciò, anche passi del Purgatorio e del Paradiso sono stati individuati, seppur in misura minore (Del Grosso et al., 2024). Ad esempio, nel diario di Giuntini è citato il verso 12 del Canto I della seconda cantica per descrivere l'azzurro cielo estivo privo di nuvole:

> Stamani, guardando il cielo, potrei finalmente descriverlo come padre Dante: «Dolce colore d'oriental zaffiro» tanto è azzurro e senza una nuvola.

Lo studio delle testimonianze all'interno del progetto Voci dall'Inferno mira dunque ad analizzare il lessico e le allusioni dantesche presenti nel corpus, focalizzandosi più di altro sull'adozione del linguaggio dell'*Inferno* della *Divina Commedia* da parte dei sopravvissuti per descrivere le esperienze nei Lager. A supporto di questo obiettivo è stata sviluppata l'applicazione web *Voci dall'Inferno Verse Similarity Search*, finalizzata a identificare citazioni e allusioni al lessico dantesco mediante la ricerca di frasi contenute nelle testimonianze e il confronto con i versi della *Divina Commedia*.
L'applicazione che si introduce in questo contributo prende le mosse dal progetto *Latin Vulgate Verse Similarity Search*, sviluppato da William Mattingly per automatizzare l'identificazione e l'estrazione di potenziali citazioni bibliche in testi latini medievali (Mattingly, 2024).

Figura 2. Donut Chart con la distribuzione delle tipologie di testimonianze all'interno del corpus Voci dall'Inferno

2. METODOLOGIA

La realizzazione dell'applicazione *Voci dall'Inferno Verse Similarity Search* è stata sviluppata usando il linguaggio di programmazione Python e ha richiesto l'utilizzo di diversi strumenti tecnologicamente avanzati (vedi Figura 3). Nello specifico, sono state impiegate due tecnologie: la prima è *Weaviate* (Weaviate, 2025), una piattaforma di vector search open-source che, sfruttando modelli di Machine Learning allo stato dell'arte, trasforma dati non strutturati in un database vettoriale ricercabile sul quale è possibile eseguire interrogazioni e analisi semantiche. Il secondo strumento utilizzato è *Streamlit* (Streamlit, 2025), un framework open-source dell'ecosistema Python che facilita la creazione e lo sviluppo di applicazioni web personalizzate.

L'approccio adottato per l'individuazione dei tasselli danteschi nel corpus di testimonianze fa leva sull'uso di metriche per il calcolo della "Sentence Similarity" (Reimens 2019), ovvero il processo di valutazione della similarità tra due testi costituiti da poche parole che abbiano una coerenza sintattica/semantica. Ciò viene realizzato utilizzando un modello matematico di sentence similarity, che permette di trasformare i dati da testo piano in punti di uno spazio vettoriale misurabili, detti *embeddings* (Cao, 2024). Questo permette di misurare la similarità tra frammenti di testo, calcolando la loro vicinanza "semantica".

Per lo specifico sviluppo del caso d'uso presentato in questo contributo è stato adottato il modello di sentence similarity denominato SentenceTransformers *all-mpnet-base*, pubblicamente accessibile per mezzo della piattaforma *HuggingFace* (Huggingface, 2025).

Figura 3. - Tecnologie e Workflow dell'applicazione

Per il corretto funzionamento dell'applicazione è stato necessario prevedere diverse fasi di sviluppo. Dopo aver raccolto le terzine delle cantiche della Divina Commedia in una struttura tabellare, mediante il modello di SentenceTransformers sono stati calcolati i corrispondenti embeddings. L'utilizzo di Weaviate permette la configurazione di una collezione per archiviare le informazioni strutturate relative alla Divina Commedia e la creazione di un database per categorizzare e archiviare i dati relativi a cantiche, canti, versi e terzine. In questo modo è possibile archiviare gli embeddings ed eseguire ricerche semantiche sui dati. Dopo aver chiamato la funzione che consente il calcolo del punteggio di similarità tra la query ricercata e gli embeddings delle terzine il sistema restituisce i passi più simili ordinati secondo un punteggio di similarità (*similarity score*). Per esempio, data una sequenza di parole quali "vuolsi così colà" le prime due terzine restituite dal sistema sono; 1) "Non impedir lo suo fatale andare: Vuolsi così colà dove si puote ciò che si vuole, e più non dimandare" [Inf. V 22-24]; 2) "E 'l duca lui: Caron, non ti crucciare: vuolsi così colà dove si puote ciò che si vuole, e più non dimandare" [Inf. III 94-96]. Rispettivamente con similarity score 0.63 e 0.62.

3. L'APPLICAZIONE

L'applicazione "Voci dall'Inferno Verse Similarity Search" ha previsto lo sviluppo di uno strato software di front-end dedicato all'interazione web con l'utente studioso. A tal fine è stato adottato il framework Streamlit che dispone l'applicazione sul browser. Essa si presenta con una barra di ricerca nella parte alta della pagina web in cui inserire la query da analizzare e fonte testuale dell'interrogazione alla Commedia dantesca vettorializzata in Weaviate; in aggiunta, l'interfaccia mostra due componenti per la scelta delle sezioni che permettono l'eventuale selezione all'interno della Commedia rispettivamente di cantica e canto; in ultimo un componente slider dedicato alla personalizzazione del numero di risultati da presentare a video dopo l'interrogazione; il tutto è avviato dal pulsante di inoltro della ricerca (vedi Figura 4).

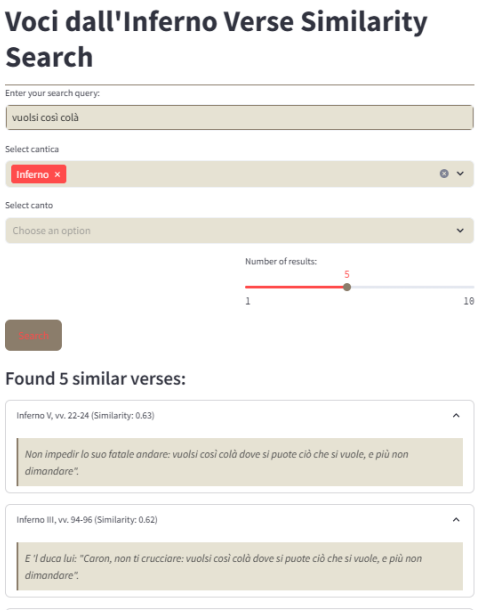

Figura 4. Screenshot dell'applicazione Voci dall'Inferno Verse Similarity Search

4. CONCLUSIONI

Il presente contributo ha illustrato l'idea progettuale e lo stato di sviluppo dell'applicazione "Voci dall'Inferno Verse Similarity Search" dedicata alla ricerca di citazioni alla Commedia di Dante (esplicite e implicite), di espressioni e vocabolario dantesco nonché di allusioni e riferimenti alle cantiche del poeta fiorentino. Il lavoro si inserisce nel più ampio contesto del progetto di ricerca Voci dall'Inferno che ha già raccolto e analizzato il primo corpus italiano di testimonianze non letterarie scritte e orali di sopravvissuti ai Lager nazisti.

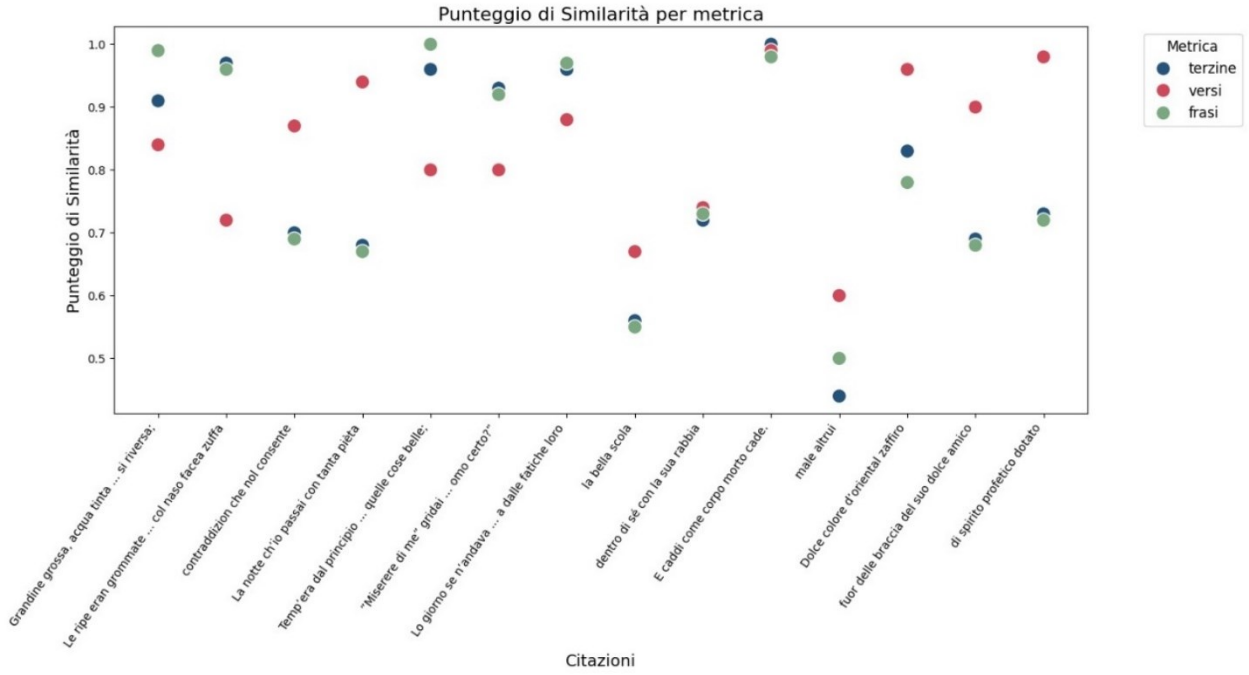

Figura 5. - Diagramma scatter plot che sintetizza i risultati della fase di valutazione dell'applicazione

L'applicazione è stata sottoposta a una fase di valutazione, mostrando un'elevata efficacia nell' individuazione dei versi danteschi in presenza di citazioni esplicite, ossia citazioni più o meno fedeli al testo originale (vedi Figura 5). Per quanto riguarda, invece, le citazioni implicite, ovvero i riferimenti meno immediati e diretti alla Commedia, l'applicazione attualmente sviluppata richiede ulteriori interventi di perfezionamento. In questa direzione, è già in corso una prima fase di *fine-tuning* del modello preposto alla generazione degli *embeddings,* con l'obiettivo di migliorare il reperimento automatico dei versi danteschi sulla base di esempi opportunamente codificati presenti nel corpus Voci dall'Inferno.
Ad esempio, con il modello attuale di *embeddings*, l'applicazione non è stata in grado di riconoscere l'allusione ai versi 109-114 del canto XIII attraverso la seguente espressione: "stare sempre agitato come selvaggina che può essere colta di sorpresa". Il sistema computazionale propone invece, con un indice di similarità pari a 0.69, il passo "Non han sì aspri sterpi né sì folti quelle fiere selvagge che 'n odio hanno tra Cecina e Corneto i luoghi còlti" [Inf. XIII 7-9].
Inoltre, la similarità tra testi viene calcolata considerando tre granularità testuali: 1) le terzine, 2) i singoli versi e 3) le frasi. Questa scelta consente di valutare in modo puntuale la granularità delle citazioni rilevate. La figura 5 mostra anche questa differenza di granularità. In particolare i valori di similarità sono rappresentati tramite piccoli cerchi colorati di colore blu per le terzine, di colore rosso per i versi e di colore verde per le frasi. Dal confronto emerge come i valori di similarità varino in funzione del livello di granularità considerato. Le citazioni più estese, anche meno precise, tendono ad avere valori di similarità più alti quando la ricerca avviene a livello di frase. Al contrario, le citazioni più brevi e aderenti al testo originale ottengono valori di similarità maggiori quando la ricerca avviene a livello di verso.
In conclusione, è attualmente in corso un'interlocuzione con l'infrastruttura CLARIN-IT, nell'ambito del progetto infrastrutturale H2IOSC, con l'obiettivo di individuare la soluzione di hosting più adeguata. Tale attività è finalizzata a garantire all'applicazione accessibilità e fruibilità per gli studiosi di Dante.

RINGRAZIAMENTI

Il Progetto Voci dall'Inferno si è avvalso del supporto dell'infrastruttura di ricerca CLARIN-IT, che ospiterà i dati e il software del progetto per la long-term preservation. Il CLARIN Knowledge Centre DiPText-KC ha messo a disposizione del progetto il proprio gruppo di esperti per consulenze scientifiche e operative (https://diptext-kc.clarin-it.it/helpdesk). Inoltre, si ringrazia la dott.ssa Elvira Mercatanti per i dati forniti e per la revisione del contributo.

BIBLIOGRAFIA

Burnard, L. (2014). What is the Text Encoding Initiative? : How to add intelligent markup to digital resources. OpenEdition Press. https://books.openedition.org/oep/426

Calderini, S., & Riccucci, M. (2020). L'ineffabilità della nefandezza: Dante «per dire» il Lager: un sondaggio preliminare nelle testimonianze non letterarie. *Italianistica*, *1*. https://doi.org/10.19272/202001301011

Cao, H. (2024). *Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark*. https://arxiv.org/abs/2406.01607

Del Grosso, A. M., Riccucci, M., & Mercatanti, E. (2024). The Impact of Digital Editing on the Study of Holocaust Survivors' Testimonies in the context of Voci dall'Inferno Project. In I. Anuradha, M. Wynne, F. Frontini, & A. Plum (A cura di), Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024 (pp. 1–9). ELRA and ICCL. https://aclanthology.org/2024.htres-1.1/

Del Grosso, A. M., Riccucci, M. & Mercatanti, E. (2024). Voci Dall'Inferno: Dante per Dire Il Lager - Digitalizzare e Studiare Le Testimonianze. In ME.TE. Digitali: Mediterraneo in Rete Tra Testi e Contesti. Proceedings of AIUCD 2024.

Hugging Face – The AI community building the future. (2025, gennaio 18). https://huggingface.co/

Mattingly, W. (2024). *Wjbmattingly/vulgata-spacy* [Python]. https://github.com/wjbmattingly/vulgata-spacy (Opera originale pubblicata 2022)

Microsoft/mpnet-base Hugging Face. (2025). Recuperato 20 gennaio 2025, da https://huggingface.co/microsoft/mpnet-base

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084

Streamlit. (2025). Recuperato 20 gennaio 2025, da https://latin-vulgate.streamlit.app/

Streamlit: A faster way to build and share data apps. (2021, gennaio 14). https://streamlit.io/

Text Encoding Initiative. (2025). Recuperato 20 gennaio 2025, da https://tei-c.org/

The AI-native database developers love | Weaviate. (2025). Recuperato 20 gennaio 2025, da https://weaviate.io/