Dalla parola al dato

Metodi e problemi nella creazione di liste lessicali per l'estrazione automatica dei dati linguistici

Reading the Italian Novel at a Distance (1830–1930)

Floriana Carlotta Sciumbata
Università degli Studi di Trieste · fsciumbata@units.it

LIX Congresso internazionale di studi della Società di Linguistica Italiana «I dati nelle scienze del linguaggio» · Bologna, 10–12 settembre 2026

1 Introduzione e obiettivo

La creazione di liste lessicali per l'estrazione automatica dei dati è una fase cruciale, spesso poco visibile, del passaggio dal testo al dato: le risorse lessicali mediano fra il corpus e l'analisi quantitativa.

Il contributo riflette su questa fase a partire dall'esperienza sul corpus RIND (che adotta il *distant reading* di Moretti 2013) di narrativa italiana (1830–1930), su tre ambiti lessicali usati come indicatori di trasformazioni sociali e culturali: **mestieri e professioni**, **titoli legati al potere** (militari, religiosi, nobiliari) e **luoghi artificiali** (Sciumbata/Nadalutti/Tringali 2021; Sciumbata 2023).

L'obiettivo del poster è mettere in luce le difficoltà ricorrenti nella costruzione delle liste e le scelte metodologiche adottate per affrontarle.

2 Repertori lessicali e metodo di estrazione

Sono state costruite ed estratte **due liste lessicali**: **1.241 lemmi** per i luoghi artificiali (8 categorie), **2.947** per i ruoli sociali (professioni, gradi militari, gerarchie ecclesiastiche, titoli nobiliari). Per ciascun lemma sono raccolte le forme utili al recupero nel corpus (maschile/femminile, singolare/plurale, alterati, varianti grafiche storiche come gli agentivi in *-ajo*), aggiunte con il supporto di ChatGPT e sottoposte a verifica.

Le occorrenze sono individuate da uno script che trova le forme nei testi del corpus.

IL CORPUS RIND (ONDELLI/MAZZARISI 2025)

1.000
opere di narrativa (1830–1930), italiane e tradotte

500 / 268
opere italiane analizzate / autori

29,9 mln
parole nel subcorpus italiano

144 / 356
opere canoniche / non canoniche

3 Fonti lessicali e criticità

Dizionari ed enciclopedie

Utile il *Dizionario analogico* (Feroldi e Dal Pra 2011); altre fonti sono state escluse per la loro scarsa sistematicità. Anche quando le definizioni sono buone, **lo scraping automatico a partire da esse resta difficile**. Sapere.it, in particolare, dedica alle professioni una pagina strutturata per voci, utile come base di partenza; anche qui, però, serve un'estrazione ulteriore.

Wikidata

Base di dati multilingue, strutturata e interrogabile: estrazione più controllata e riproducibile (es. *muratore*, *insegnante di inglese*, *giudice di pace*). Richiede comunque selezione e verifica; **tende a concentrarsi su professioni e luoghi «alti» e contemporanei**, cariche di prestigio e istituzioni riconosciute, più che su mestieri comuni o storici, da cui un vaglio piuttosto oneroso.

Modelli linguistici (LLM)

ChatGPT e strumenti come NotebookLM aiutano in fase esplorativa, ma un elenco di 400 professioni richiesto a ChatGPT **non ha offerto vantaggi sulla lista già costruita**; per i luoghi propongono nomi propri (Torre Eiffel, Colosseo) invece di nomi comuni. Chi legge i testi resta insostituibile.

4 Problemi morfologici nell'estrazione

Una lista basata solo sui lemmi non basta: il genere, il numero, le alternanze grafiche e la polisemia fanno sì che la stessa parola compaia nei testi in forme che il lemma da solo non prevede, o che una stessa forma corrisponda a significati diversi. Le risposte possibili sono tre: **lavorare sulle forme invece che sui lemmi**, ricorrere a testi annotati per parte del discorso, scartare caso per caso le occorrenze ambigue. Ma **nessuna è risolutiva da sola**, come mostra la tabella qui sotto.

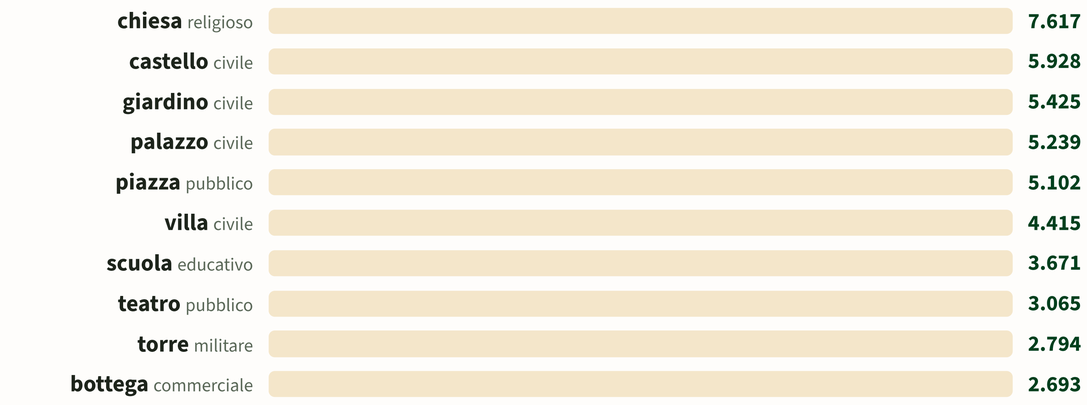
5 Problemi, esempi e soluzioni adottate

Problema	Esempio	Soluzione adottata
Variazione grafica storica	***lampionajo***	**Aggiunta** manuale della variante nella lista, oppure normalizzazione della grafia del corpus.
Categoria grammaticale non univoca	***generale***: aggettivo o grado militare	**Lavoro sulle forme** e verifica manuale delle concordanze; il POS tagging aiuterebbe, ma gli strumenti disponibili sono addestrati su testi contemporanei (quotidiani, Wikipedia, social media), non sulla prosa ottocentesca.
Polisemia e omografia	***caffè, ufficio, pensione*** (luogo o altro); ***chiesa, scuola, teatro*** (edificio, gruppo sociale o istituzione)	**Revisione** manuale delle concordanze; la disambiguazione automatica resta solo parziale.
Frequenza eccessiva / falso positivo	***forte***, ***centro***	**Esclusione** del lemma dalla lista a monte, non semplice revisione caso per caso.
Variazione di genere e alterazione	***operaio*** / ***operaia***; ***maestra*** / ***maestrina***	**Aggiunta** manuale delle forme femminili, epicene e alterate (es. diminutivi): non deducibili dal solo lemma maschile, e non sempre esaustive.

6 Casi di studio

I 10 lemmi di luogo più frequenti

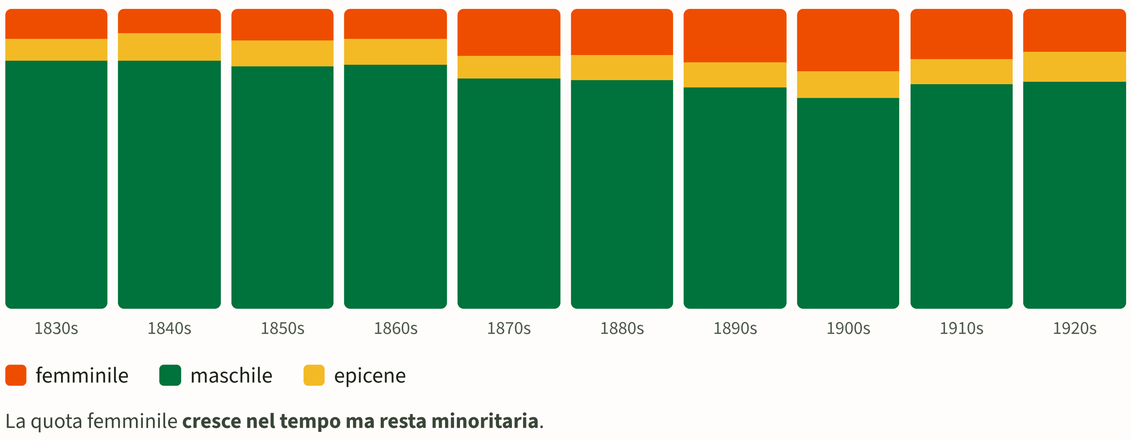
occorrenze nel subcorpus italiano (1830–1930) · escluse le forme sistematicamente ambigue *caffè*, *corte* e *chiese* (omografo del passato remoto di «chiedere»)



Insieme coprono circa il **37%** delle occorrenze totali: edifici religiosi e dimore storiche (*chiesa, castello, palazzo, villa*) convivono con spazi funzionali e urbani (*scuola, teatro, bottega*).

Professioni per genere grammaticale, per decennio

quota % di occorrenze femminili, maschili ed epicene sul totale delle professioni, 1830–1929



La quota femminile **cresce nel tempo ma resta minoritaria**.

7 Conclusioni

La costruzione di liste lessicali per l'estrazione automatica richiede l'integrazione di fonti diverse (dizionari, basi di dati strutturate, modelli linguistici) e un **controllo umano costante**, che nessuno strumento sostituisce.

Le difficoltà morfologiche e semantiche dell'italiano storico impongono scelte esplicite: lavorare sulle forme più che sui lemmi, filtrare a mano le occorrenze ambigue, documentare ogni esclusione. Un **confronto sistematico fra il sottocorpus italiano e quello tradotto**, non ancora condotto, potrebbe far emergere differenze nella distribuzione lessicale delle categorie qui discusse, fra cui, per esempio, quella di genere nei ruoli sociali.

8 Riferimenti bibliografici

Moretti, F. (2013). *Distant Reading*. Londra–New York: Verso.

Ondelli, S. / Mazzarisi, P. (2025). Il corpus di prosa letteraria del progetto RIND (1830–1930). In Rebora et al. (a cura di), *AIUCD 2025*, Verona: Università di Verona–AIUCD, 289–296.

Sciumbata, F.C. / Nadalutti, P. / Tringali, L. (2021). «Trovare lavoro» in un corpus di narrativa del XIX–XX secolo. *Rivista internazionale di tecnica della traduzione*, XXIII: 235–268.

Sciumbata, F.C. (2023). Titles Associated with Power in a Corpus of Fictional Prose in Italian (XIX–XXI Century). In Mišić Ilić et al. (a cura di), *Jezik, Književnost, Moć*, Niš: Filozofski fakultet, 279–297.

https://rind.units.it