Rivista Internazionale di Tecnica della Traduzione / International Journal of Translation 2021 (23), 235-268
DOI: 10.13137/2421-6763/33468 ISSN 1722-5906 (print) ISSN 2421-6763 (online)
https://www.openstarts.units.it/dspace/handle/10077/2503

"Trovare lavoro" in un corpus di narrativa del XIX-XX secolo. Procedure, aspetti e problemi di creazione, estrazione e rappresentazione dei dati

Floriana Carlotta Sciumbata
Università di Trieste
fsciumbata@units.it

Paolo Nadalutti
Gruppo Interdisciplinare di Analisi Testuale
paolo.nadalutti@gmail.com

Luca Tringali
tringalinvent@libero.it

Abstract

This paper illustrates methods, tools, and preliminary results of a study aimed to create a list of job titles and analyze them in a corpus of 100 Italian canonical and non-canonical fictional prose works published between 1825 and 1923. Job titles are interesting because they reflect socio-economic changes, as well as giving important information on how literary settings and genres changed. After a short introduction on job titles, we will discuss tools and methods used for the creation of a list of words, data extraction, and data representation, both from a linguistic and programming point of view. Some preliminary results will be shown and discussed with a statistical approach: data do not suggest significant patterns over time: whereas job titles appear more or less consistent from a chronological point of view. Finally, some advantages and limitations will be examined. The goal of this study is to develop a set of tools and methods that can be easily reproduced to build complex lexical lists, find their items, and represent data for corpora of any genre and size in a simple and effective way.

Keywords

Job titles, fictional prose, chronological analysis, lexical extraction, automatic text analysis

1. Introduzione

Il presente contributo[1] presenta metodologie, strumenti e risultati parziali, ma anche problemi e potenzialità, di uno studio[2] multidisciplinare realizzato con strumenti linguistici, informatici e statistici. La ricerca ha come obiettivo la messa a punto di un sistema per l'estrazione automatica di dati riguardanti i nomi di mestieri e professioni sul corpus di narrativa, che raccoglie testi pubblicati nel secolo che va dal 1825 al 1923.[3] I lavori analizzati sono quindi rappresentativi di un periodo caratterizzato da profondi mutamenti, sia dal punto di vista storico sia da quello letterario, e particolarmente interessante da osservare.

L'estrazione automatica dei nomi di professioni in un corpus non è però di facile esecuzione poiché non sono disponibili liste elettroniche da impiegare per applicazioni automatiche. Anche le caratteristiche morfologiche e sociolinguistiche dell'italiano pongono alcuni ostacoli che devono essere tenuti in considerazione per perfezionare l'estrazione dei dati.

Di seguito saranno illustrate le metodologie scelte per:

- individuare le fonti e creare le liste di nomi di mestieri (§ 3);
- estrarre i dati dal corpus (§ 4);
- rappresentare i risultati (§ 5).

Le questioni saranno osservate anche dal punto di vista informatico e della programmazione (§ 6). Saranno poi commentati alcuni dati ottenuti (§ 7), che sono da considerare come risultati preliminari, qui mostrati per testare le funzionalità del sistema che abbiamo messo a punto. Infine, saranno discussi aspetti, problemi e prospettive di ricerca future (§ 8).

Per raggiungere l'obiettivo di questo studio, abbiamo scelto di combinare più metodi e di creare alcuni strumenti *ad hoc*. In particolare, per automatizzare le procedure, abbiamo messo a punto *ExSTRA* (*Extracting social STatus Ranks Automatically*), cioè un insieme di diversi *script* nei linguaggi di programma-

1 Il lavoro di ricerca e i testi che lo illustrano sono il frutto di un approccio interdisciplinare che applica metodi linguistici, statistici e informatici e che ha visto la piena collaborazione dei tre autori sotto tutti i punti di vista. Ai soli fini dell'attribuzione, specifichiamo che Floriana Carlotta Sciumbata ha redatto i §§ 1, 2, 3, 4, 5 (con Paolo Nadalutti), 7, 7.1 e 8; Paolo Nadalutti, oltre che del § 5, si è occupato del § 7.2; Luca Tringali ha scritto il § 6.

2 Lo studio che descriveremo è stato avviato nell'ambito del progetto Distant Reading for European Literary History (COST Action CA16204), che ha finanziato una missione scientifica breve, intitolata "Tackling Research Questions from WG3 to WG2: analysing reported speech, exploring historical places and meeting social groups in the Italian and English ELTeC corpora", svolta da Floriana Carlotta Sciumbata presso l'Università di Fiume (Croazia) tra gennaio e marzo 2020.

3 Per la descrizione del corpus rimandiamo alla premessa a questo articolo e a quello di Sciandra, Trevisani e Tuzzi.

zione *Python* e *R*,[4] che rappresentano una vera e propria *pipeline*, una catena di montaggio che rende le operazioni facili e veloci, ma soprattutto facilmente replicabili in altre ricerche, che possono partire da liste o *corpora* diversi da quello qui utilizzato, e addirittura da lingue diverse dall'italiano.

2. I nomi di mestieri e professioni

I nomi di mestieri e professioni sono senza dubbio un dato interessante poiché sono la cartina al tornasole di cambiamenti avvenuti in ambito economico e sociale nel corso del tempo, che si riverberano anche sulla lingua. Una dimostrazione dell'importanza dell'argomento sono, ad esempio, le numerose discussioni sui nomi di professioni al femminile che animano il dibattito di studiosi, giornalisti e parlanti ormai da qualche decennio (Burr 1995; Cardinaletti e Giusti 1991; Gheno 2019; Robustelli 2012; Sabatini 1987). L'analisi di mestieri e professioni è rilevante anche per chi s'interessa di statistica e sociologia (Gallo e Scalisi 2013), ma anche di storia o di letteratura, poiché riflettono aspetti come l'evoluzione dell'ambientazione in letteratura (contrapposizione tra ambiente urbano e rurale) oppure l'andamento e la rappresentazione delle diverse classi sociali.

Dal punto di vista linguistico, i nomi di mestieri e professioni rientrano nella categoria dei nomi d'agente, etichetta utilizzata nel campo della morfologia per indicare una "persona che fa, ha fatto o farà, una o più volte, per occasione accidentale, per abitudine e/o per funzione un'azione connessa con l'elemento di base" (Lo Duca 1990: 31), dove per "elemento di base" si intende il nome o il verbo che costituiscono il lessema di base della parola derivata. I nomi di agente possono essere quindi creati a partire da nomi, verbi e altri elementi combinati con un'ampia varietà di suffissi. Perciò, costituiscono una categoria estremamente ricca e variegata che ricorre a un'ampia gamma di risorse morfologiche e lessicali, con apporti dalle lingue straniere e da quelle classiche; cambiamenti di denominazione nel corso del tempo; convivenza di definizioni uguali che fanno riferimento a lavori differenti; presenza di diverse varietà diatopiche, con geosinonimi e geomonimi (D'Achille e Grossmann 2017: 147). L'interesse per l'argomento è dimostrato anche dalla pubblicazione di numerosi repertori più o meno recenti (tra cui, a titolo esemplificativo, Barbieri 1874; Fissi 1983; Garzoni 1584; Medici 1967; Proietti 1991; riportati anche in D'Achille e Grossmann 2017).

4 Gli script, i *corpora* e i dati estratti si trovano all'indirizzo https://github.com/flometis/ExSTRA/ (consultato il 5/12/2021). Sempre a meri scopi di attribuzione, l'ideazione, la progettazione e la programmazione degli script in *Python* sono a cura di Luca Tringali e Floriana Carlotta Sciumbata. Lo script *R* per i grafici a barre è a cura di Luca Tringali, quello per la creazione interattiva dei grafici è a cura di Paolo Nadalutti.

3. Creazione della lista di nomi di professioni

Uno dei problemi in cui siamo incorsi durante lo studio è la creazione della lista da cui partire per l'estrazione dei nomi dal corpus. Nonostante la letteratura e i repertori esistenti, non ci sono liste elettroniche e sistematiche di nomi di professioni,[5] necessarie per l'estrazione automatica. Abbiamo quindi deciso di stilare una lista apposita per lo studio, la cui creazione ha richiesto la combinazione di diverse fonti estratte con approcci diversi.

In primo luogo, abbiamo consultato dizionari ed enciclopedie generici online per cercare di estrarre in automatico i lemmi rilevanti per la nostra ricerca. Dizionari ed enciclopedie hanno senza dubbio il vantaggio di contenere numerosi lemmi con le relative definizioni, ma si sono rivelati poco utili per la loro scarsa sistematicità e perché contengono dati non strutturati. Per esempio, nel dizionario De Mauro si trovano le seguenti definizioni:

- fornaio: nei forni, addetto alla lavorazione e cottura del pane | chi vende il pane; proprietario o gestore di un forno o di un negozio per la vendita di pane;
- insegnante: chi insegna, spec. per professione;
- medico: chi pratica la medicina avendo conseguito il titolo accademico e l'abilitazione all'esercizio della professione;
- muratore: chi è addetto alla costruzione di opere in muratura;
- oculista: medico specialista nella cura delle malattie degli occhi.

Sebbene alcune regolarità, come la specifica dell'aspetto professionale in alcuni casi e la presenza del pronome *chi* in altri, le definizioni non sono sufficientemente regolari per poter distinguere in automatico i lemmi pertinenti solo a partire dalle definizioni stesse: la creazione di regole per l'estrazione automatica sarebbe perciò estremamente dispendiosa, così come lo spoglio manuale delle voci. Negli altri dizionari sono riscontrabili problemi simili, come nel vocabolario Treccani, che presenta le seguenti definizioni per gli stessi lemmi:

- fornaio: operaio addetto alla lavorazione e cottura del pane; padrone di un forno per il pane; chi ha negozio per la vendita del pane (con o senza il forno annesso) e, talvolta, anche di altri generi alimentari;
- insegnante: chi si dedica all'insegnamento, chi esercita la professione d'insegnare;
- medico: chi professa la medicina;
- muratore: operaio addetto alla costruzione di opere murarie;
- oculista: medico specializzato nello studio e nella cura delle malattie dell'occhio.

Anche le enciclopedie sono caratterizzate da scarsa sistematicità, e le entrate riferite a mestieri e professioni sono più scarse rispetto ai dizionari. Per esempio, la ricerca nell'enciclopedia Treccani non restituisce voci per *fornaio* e *oculista*. La ricerca sulle enciclopedie è stata comunque fruttuosa: ci ha permesso di trovare una pagina dell'enciclopedia *Sapere* di DeAgostini che raccoglie proprio tutte le

5 Cfr. anche Gallo e Scalisi (2013: 43), che lamentano l'assenza di un repertorio nazionale delle professioni.

voci su mestieri e professioni,[6] da *abbadatore* a *zoologo*. La pagina è stata sottoposta a una procedura di *scraping*, cioè all'estrazione automatica dei suoi contenuti: il risultato è una tabella contenente 1.211 lemmi singoli, non sono quindi presenti locuzioni più complesse. Le entrate sono state modificate perché contenevano lettere accentate per indicare apertura e chiusura delle vocali toniche di ogni parola. Le lettere accentate non previste dalla grafia sono perciò state sostituite con gli equivalenti senza accenti poiché avrebbero impedito la corretta individuazione nel corpus, dove le parole sono normalmente scritte senza particolari indicazioni sulla pronuncia.

Un'altra fonte utilizzata per la realizzazione della lista è *WikiData*, cioè la banca dati alla base dei progetti collegati a *WikiMedia*, di cui fa parte anche l'enciclopedia libera e collaborativa *Wikipedia*. *WikiData* presenta diversi vantaggi: oltre ad avere una licenza di pubblico dominio che la rende liberamente utilizzabile, contiene dati strutturati, cioè organizzati per campi predefiniti, che facilitano l'estrazione automatica delle informazioni. Infine, *WikiData* è un database multilingue e ogni voce è collegata ai suoi corrispettivi in tutte le traduzioni disponibili. Tale caratteristica è particolarmente utile perché rende flessibile e replicabile anche in altre lingue il sistema utilizzato in questo studio, che, in prospettiva, si inserisce in un progetto internazionale.[7]

La banca dati contiene una sezione chiamata "Profession", usata per categorizzare le voci riferite a persone di *Wikipedia*. Anche le informazioni di *WikiData* sono state estratte in automatico e inserite in una tabella simile a quella già descritta per *Sapere*, che contiene 1.969 voci composte da lemmi (es. *custode, muratore*, *sarto*), locuzioni complesse (es. *insegnante di inglese, data scientist, musicista di strada*) ed espressioni polirematiche (es. *agente di commercio, giudice di pace*). La lista va da *abbé* a *zootecnico*.

Le due liste ottenute da *Sapere* e *WikiData* sono state combinate in un'unica tabella: dopo aver unificato le voci doppie, il risultato finale comprende 2.947 entrate, che riportano la forma base del lemma, della locuzione o della polirematica (maschile singolare nella maggior parte dei casi, femminile singolare in alcuni, come nel caso di *ballerina, massaia, mercantessa*); la fonte (*Sapere* o *WikiData*); l'URL alla pagina della voce; e infine, per alcune voci, una breve definizione data dalla fonte stessa. Le voci di *Sapere* contengono anche informazioni su genere e numero del sostantivo.

Nella lista finale si trovano sia nomi di mestieri più antichi e ormai desueti, spesso legati all'ambito del lavoro manuale (es. *buffatore, cisiario, lapicida*); sia riferi-

6 https://www.sapere.it/sapere/enciclopedia/storia-e-societ%C3%A0/economia-e-statistica/generale/mestieri-e-professioni.html (consultato il 30/11/2021).

7 *WikiData* contiene anche versioni in alcuni dialetti, tra cui siciliano, lombardo, veneto, romagnolo eccetera, potenzialmente interessanti per analizzare autori dialettali o tentare di individuare dialettismi. Bisogna tuttavia considerare diverse limitazioni, come la disomogeneità nella grafia dei lavori in dialetto, le diverse varietà locali, e ovviamente il numero più limitato di voci, probabilmente redatte da un gruppo molto ristretto di autori.

menti a mestieri più contemporanei (*data scientist, social media manager, influencer*): ai primi contribuisce soprattutto la lista di *Sapere*, ai secondi quella di *WikiData.* Sebbene per il corpus che abbiamo preso in analisi siano senza dubbio più utili le forme più arcaiche, la lista ottenuta è piuttosto ampia e versatile e può essere utilizzata su testi di tipologia diversa e prodotti in diversi periodi. In *WikiData* sono anche presenti alcuni forestierismi non molto comuni in italiano (es. *akyn, ama, muhtasib, opričniki, yobidashi*), oltre a diversi anglicismi (oltre a quelli già citati, altri esempi sono *business analyst, character designer, controller, general manager*).

Infine, una terza fonte che abbiamo preso in considerazione è la lista di mestieri e professioni usata dall'ISTAT per le sue rilevazioni, che contiene 800 voci. Oltre a quelle generiche come *avvocati, gelatai, pellettieri*, si trovano diverse locuzioni molto complesse che fanno riferimento a definizioni tecniche (*addetti alla preparazione, alla cottura e alla vendita di cibi in fast food, tavole calde, rosticcerie ed esercizi assimilati*; *piastrellisti e rivestimentisti in pietra e materiali assimilati*; *tagliatori di tessuti*), probabilmente poco utili nell'ambito di una ricerca su un corpus di narrativa o, in generale, su testi non tecnici. Inoltre, come si può notare, questa lista contiene le forme al plurale delle parole, diversamente dalle altre due fonti, e avrebbe dovuto essere rimaneggiata per il metodo di estrazione adottato che vedremo nella sezione successiva. Per questi motivi, abbiamo deciso di escludere la lista dell'ISTAT.

4. Estrazione dei dati

Una volta estratte le liste, gli elementi non sono stati individuati direttamente nel corpus perché ciò avrebbe determinato due problemi non trascurabili. Il primo si ricollega al fatto che le voci della lista sono forme base (maschile o femminile singolare): cercarle direttamente nei testi avrebbe inevitabilmente escluso tutte le forme al plurale. Il secondo problema riguarda i falsi positivi: si pensi alla parola "domestico", che può essere classificata come aggettivo, ma anche come sostantivo, e solo nel secondo caso corrisponde a un mestiere.

Per ovviare a entrambi i problemi, abbiamo deciso di trattare il corpus usando il software *UdPipe 1* (Straka *et al.* 2016), che converte i testi in formato CoNLL-U,[8] cioè li lemmatizza e aggiunge informazioni sulla parte del discorso (*PoS, part of speech*), le caratteristiche morfologiche e il ruolo sintattico dei singoli *token*.[9]

8 Cfr. https://universaldependencies.org/format.html (consultato il 10/12/2021).

9 Per il training di *UDPipe*, abbiamo provato a mettere insieme quattro *corpora* di *training* attualmente disponibili per l'italiano: *ISDT*, *PARTUT*, *POSTWITA* e *VIT* (cfr. https://universaldependencies.org/treebanks/it-comparison.html); necessari a "insegnare" al sistema a riconoscere le caratteristiche dei *token*, per tentare di ovviare alla scarsità di risorse. Occorre tuttavia tenere in considerazione che tutti i *corpora* di *training* sono basati su testi non letterari e in italiano contemporaneo: potrebbero perciò presentarsi diversi errori nella lemmatizzazione e nella classificazione delle caratteristiche linguistiche.

Abbiamo quindi elaborato uno *script* che, dopo aver convertito il testo in formato CoNLL-U, cerca i nomi di mestieri solo tra i *token* lemmatizzati e classificati come sostantivi, così da ottenere risultati più precisi e pertinenti.[10] Per avere un'idea sui risultati, riprendiamo l'esempio di *domestico*, presente nella nostra lista di mestieri. Le parole lemmatizzate come *domestico* occorrono 516 volte: 211 sono classificate dal *PoS tagger* come aggettivi, una come avverbio (dal *token domesticamente*), una erroneamente indicata come nome proprio e una non classificato, e infine 302 come sostantivi. Per il nostro conteggio di nomi di mestieri abbiamo preso in considerazione solo questi ultimi, riducendo quindi il numero di falsi positivi.

Lo *script* che abbiamo realizzato, che si occupa di individuare i lemmi nelle opere che compongono il corpus, restituisce sia una tabella delle occorrenze dei lemmi per singolo testo, sia una tabella riassuntiva con tutti i dati estratti. Per ogni lemma, oltre al numero di occorrenze, sono indicati anche i dati del testo (file, autore, titolo, sesso dell'autore, anno di pubblicazione, decennio, lista di provenienza). Tali indicazioni permettono di creare filtri complessi per l'estrazione dei dati: le occorrenze possono infatti essere calcolate per periodo, ma anche per singolo autore o in base al sesso. È inoltre possibile affiancare altre categorie, come il genere letterario, la provenienza geografica dell'autore e così via in un'ulteriore tabella, utile a incrociare diversi dati e quindi a ottenere informazioni complesse.

5. Rappresentazione

Per la rappresentazione dei dati, abbiamo creato due sistemi. Il primo è uno *script* in *R* che permette di ottenere grafici in modo automatico dalle tabelle con le occorrenze. Tali diagrammi a barre consentono in modo rapido e intuitivo di dare rappresentazione del numero di occorrenze nel corpus di ognuna delle forme grafiche individuate.

Un secondo metodo di rappresentazione è invece dinamico e interattivo, cioè l'utente può scegliere i dati da confrontare e il tipo di analisi da applicare. Il sistema è basato sull'analisi delle corrispondenze (Greenacre e Blasius 1994) e consente di cogliere in un unico grafico il sovra- o sotto-utilizzo delle forme grafiche all'interno dei singoli testi.[11]

10 Per questo studio, abbiamo limitato la ricerca solo alle entrate della lista di nomi di mestieri composte da lemmi singoli, ma è allo studio un filtro per individuare anche locuzioni complesse e forme polirematiche.

11 Rimandiamo al § 7.2 per un approfondimento sul funzionamento del sistema dinamico.

6. Dal punto di vista informatico

In questa sezione rivedremo dal punto di vista informatico alcune delle questioni già illustrate in precedenza, approfondendo in particolare alcuni aspetti della programmazione delle diverse funzioni di *ExSTRA*. Tali questioni risultano interessanti sia perché hanno richiesto l'applicazione di diversi approcci sia per le loro potenzialità anche in vista di applicazioni a materiale e ambiti diversi.

6.1. La pulizia dei testi

Dal punto di vista della programmazione, il progetto *ExSTRA* si è rivelato complessivamente piuttosto semplice: ottenuta la lista dei lemmi da cercare (il vocabolario) e la versione lemmatizzata dell'intero corpus, si tratta solo di eseguire un ciclo su tutti i lemmi del vocabolario per cercarli uno a uno all'interno dei testi. Come già accennato, per risolvere il problema della lemmatizzazione abbiamo deciso di ricorrere a *UdPipe*, uno strumento basato su un algoritmo di intelligenza artificiale. Questo perché scrivere un algoritmo che racchiuda tutte le possibili regole di lemmatizzazione di una lingua complicata come l'italiano avrebbe comportato un consumo di tempo notevole, per poi probabilmente terminare comunque con una percentuale di affidabilità non troppo lontana da quella di *UdPipe* (tra il 95% e il 97% con i principali modelli per l'italiano).[12] Tra i vari *tagger* presi in considerazione, quando abbiamo iniziato ad approcciarci al problema, *UdPipe* era l'unico che soddisfacesse facilmente i requisiti: avere modelli sufficienti per la lingua italiana, supportare un buon numero di lingue europee, essere sufficientemente leggero da poter funzionare anche su macchine dotate di poche risorse, e quanto più semplice possibile da integrare in *Python*.

L'unica condizione da rispettare, per utilizzare *UdPipe* o un qualsiasi altro *tagger*, consiste nel ripulire i testi. Molti dei testi, infatti, risultavano "sporchi", chiaramente scansionati da libri cartacei e tradotti in testo digitale da software OCR. Pochi erano i veri ebook, nati come testi digitali e dunque perfettamente puliti. La maggioranza dei testi aveva, quindi, gravi problemi di disallineamento tra colonne ed errori nell'impaginazione, per esempio: capitava spesso di ritrovare il numero di pagina all'interno del testo stesso, mentre è ovvio che per un'analisi in cui si è interessati solo al contenuto di un'opera la numerazione delle pagine e dei paragrafi è irrilevante o, peggio, controproducente. A causa di questi errori dell'OCR e della successiva rimozione degli invii a capo non desiderati (dovuti soltanto ai vincoli spaziali della pagina cartacea), poteva capitare che una parola venisse spezzata nel mezzo dal numero di una pagina o di un capitolo. Se per un umano è ovvio che "gua42rdare" sia la parola "guardare" interrotta dal numero di

12 Cfr. i dati riportati sul manuale del software: https://github.com/ufal/udpipe/blob/master/MANUAL (consultato il 10/12/2021).

pagina 42, per *UdPipe* si tratterebbe di un lemma non riconoscibile. Seppure alcuni errori possano essere rimasti, un controllo a campione ci permette di stabilire che la maggioranza di questi problemi è stata identificata con una serie di semplici espressioni regolari e corretta automaticamente. Abbiamo sviluppato un apposito *script* che lancia tutte le espressioni regolari richieste, scritte per pulire gli errori più comuni, in sequenza sui vari testi, in modo da rendere la procedura applicabile anche nel caso si voglia lavorare su un corpus differente.

6.2. Lo *scraping* di sapere.it

L'aspetto più interessante, dal punto di vista informatico, è legato alla creazione del vocabolario di *ExSTRA*, la lista dei mestieri. Abbiamo deciso di sviluppare una metodologia facilmente adattabile ad altri contesti, quindi il vocabolario è realizzato come una tabella che per ogni lemma contiene anche una serie di altre informazioni. Il lemma non è univoco: è infatti logico pensare che, avendo diverse fonti, molti lemmi siano presenti in ciascuna di esse. Inoltre, è possibile che uno stesso lemma appaia due volte perché dispone di significati differenti. Per questo motivo abbiamo optato per l'inserimento di un'altra colonna con la sorgente del lemma e l'eventuale descrizione: questa non è disponibile per tutti i lemmi, ma può aiutare a distinguere i lemmi duplicati durante un controllo manuale. Le due fonti di lemmi attualmente implementate sono l'enciclopedia *Sapere* e la base dati di *WikiData*. Abbiamo deciso di partire da una lista di lemmi estratta dal sito web dell'enciclopedia *Sapere* per avere una fonte autorevole: in questo modo chi desidera cercare soltanto lemmi "sicuri" può filtrare il vocabolario in base alla fonte dei lemmi e scegliere soltanto questi. Questa enciclopedia non è pensata per una ricerca programmata: non offre alcun tipo di API di ricerca e le informazioni devono essere estrapolate dalle pagine HTML. Abbiamo dunque dovuto sviluppare un algoritmo di *scraping*, cioè di estrazione automatica dei contenuti, per leggere il codice HTML di ciascuna pagina e estrarre le informazioni sui lemmi in base agli elementi del DOM.[13] Per fortuna, il sito sapere.it è almeno stato progettato per fornire *Rich Snippets*[14] ai motori di ricerca: ciò significa che è possibile identificare il lemma grazie al tag *title* e la descrizione grazie al tag meta chiamato *description*, secondo il fac simile

```
<title=”lemma”>
<meta content=”Testo della descrizione” name=”description”\>
```

13 Il DOM è l'insieme dei vari tag HTML che compongono la pagina.

14 I *Rich Snippets* sono un formato standard di tag riconosciuti da tutti i motori di ricerca per estrarre le informazioni principali da una pagina web. Sono in sostanza l'anteprima che il motore di ricerca mostra della pagina web. Cfr. https://backlinko.com/hub/seo/snippets (consultato il 10/12/2021).

Questo ha facilitato l'estrazione delle informazioni: è stato sufficiente scorrere tutta la homepage della categoria delle professioni (nella quale sono presenti i link ai vari lemmi censiti) per ottenere gli URL delle varie pagine nelle quali trovare gli *snippet* da estrarre. Naturalmente, lo svantaggio dello *scraping* è che, se il sito web venisse modificato, sarebbe necessario rifare da capo l'algoritmo. Tuttavia, un'enciclopedia tradizionale come Sapere è da considerarsi "immutabile", visto che gli aggiornamenti sono minimi e sarebbe più facile aggiungere manualmente eventuali nuovi lemmi, quindi il problema non sarebbe troppo difficile da risolvere.

6.3. *WikiData*: una fonte aperta

Per avere una fonte più flessibile, multilingue, e in continuo aggiornamento, abbiamo optato per l'utilizzo di *WikiData*. È necessario fare una precisazione: *WikiData* ha una propria redazione e un meccanismo di controllo e revisione più rigido rispetto a *Wikipedia*, quindi, seppure sia sempre possibile un errore, la si può considerare una fonte affidabile di informazioni, soprattutto per lingue molto diffuse e con molti autori.[15] Il vantaggio di *WikiData* è che, essendo una sorgente di dati strutturati, è possibile fare ricerche mirate e incrociare i dati per aumentare l'affidabilità dei risultati. Inoltre, è possibile migliorare l'affidabilità dei risultati controllandoli attraverso altre banche dati. Per esempio, le principali voci di *WikiData* hanno un riferimento all'ID del thesaurus della Biblioteca Nazionale Centrale di Firenze,[16] che potremmo usare in futuro per controllare e relazionare i lemmi.

Il database di *WikiData* può essere interrogato tramite *query* SPARQL, grazie a comode API HTTP.[17] Occorre tuttavia impostare in modo ottimale la *query* per evitare di far finire la richiesta in *timeout*. La logica di *WikiData* è facile da comprendere leggendo la scheda di una qualsiasi entità: tutti gli oggetti censiti in *WikiData* sono infatti chiamati *entità*, e ogni entità ha anche *proprietà*, tutte accessibili con codici univoci. Per esempio, l'entità https://www.wikidata.org/wiki/Q901 (*scienziato*) ha come proprietà P279 ("sottoclasse di") le altre entità *ricercatore* e *erudito* e come proprietà P31 ("istanza di") l'entità *professione*. Questo perché fare lo *scienziato* è una *professione* svolta dalle *persone erudite* o, se vogliamo vedere la questione da un'altra angolazione, uno scienziato è un tipo di ricercatore.

15 Per esempio, gli autori attivi per la lingua italiana sono mediamente 3.000, mentre per il siciliano sono 8 e per il veneto sono 15 (https://stats.wikimedia.org/#/it.wikipedia.org/contributing/active-editors/normal|line|2-year|(page__type)-content*non-content|monthly, consultato il 10/12/2021).

16 Per esempio https://thes.bncf.firenze.sbn.it/termine.php?id=3509 per il lemma "scienziato").

17 https://query.wikidata.org (consultato il 10/12/2021).

Il metodo più efficiente per ottenere le varie professioni nella lingua italiana consiste nell'eseguire una *query SELECT* prelevando tutti gli *item* che hanno la proprietà *istanza di* (P31) uguale al tipo di entità che si vuole ottenere (per esempio *Q28640*, che indica la *professione*). L'*item* è l'intero insieme di dati che *WikiData* possiede per quella entità, e nelle ricerche viene rappresentato dal suo *ID* univoco (i codici con lettera Q o P che abbiamo già citato). Nel caso di *ExSTRA*, tra tutti questi dati è in realtà utile principalmente la *label* (l'etichetta, cioè il lemma), e al massimo la descrizione. Queste due entità si possono ottenere come *itemLabel* e *itemDescription*. La lingua predefinita è l'inglese, ma è possibile selezionare una qualsiasi lingua tra quelle supportate da *WikiData*. La selezione della lingua è poco intuitiva, perché non basta specificare il parametro *language*: bisogna ricorrere al servizio *wikibase:language*, e dunque serve una sintassi articolata. Però nel complesso la *query* è abbastanza leggibile. Per esempio, per ottenere tutte le professioni:

```
SELECT ?item ?itemLabel ?itemDescription WHERE {
 SERVICE wikibase:label { bd:serviceParam wikibase:language "it". }
 ?item wdt:P31 wd:Q28640.
}
```

Le *label* sono interessanti perché, oltre a essere declinate in base alla lingua desiderata, possono anche essere declinate in base al genere (maschile o femminile). Attualmente nel vocabolario di *ExSTRA* sono presenti principalmente forme maschili, ma stiamo valutando l'ampliamento della lista anche con le forme femminili. Questo consentirebbe, per esempio, di capire con precisione quando compaia una forma femminile di una professione.

Non tutti i lemmi possiedono una forma femminile (proprietà *P2521*), quindi la lettura di questo dato deve essere opzionale, e la *query* verrebbe modificata in questo modo:

```
SELECT ?item ?itemLabel ?flabel ?itemDescription WHERE {
OPTIONAL { ?item wdt:P2521 ?flabel . BIND(lang(?flabel) as ?lang) . FILTER(?lang="it")}
 SERVICE wikibase:label { bd:serviceParam wikibase:language "it". }
 ?item wdt:P31 wd:Q28640.
}
```

Un altro tipo di risultato già fornito dall'attuale *query* (e presente nel vocabolario *ExSTRA*) sono gli *ngrams*, cioè sequenze di parole: *WikiData* offre intere locuzioni come *professore universitario*, utili per distinguere gli ambiti (un *professore* generico da uno che insegna specificamente in un'università). Al momento ci siamo limitati alla ricerca dei lemmi singoli, ma il meccanismo che utilizziamo può già essere modificato per la ricerca degli *ngrams*.

La lista di lemmi ottenuta va ripulita, ma *WikiData* offre già una buona base di partenza. È ovvio che alcune professioni potrebbero non avere la *label* in lingua

italiana, perché la lista che si ottiene rappresenta ogni singola professione recensita su una qualsiasi edizione internazionale di Wikipedia. Possono quindi essere presenti professioni totalmente sconosciute nel nostro Paese e mai tradotte in lingua italiana. Per esempio, il lemma giapponese *ama* (le pescatrici giapponesi) esiste anche per l'italiano. Invece, il lemma portoghese *repentista* (un particolare tipo di poeta) non possiede alcuna *label* in lingua italiana perché la parola non è nota nell'ambiente italiano.

6.4. Utilità delle *query* complesse

Per questo primo studio abbiamo scelto di cercare esclusivamente le "professioni", ma abbiamo già previsto nello *script* la possibilità di integrare per ricerche future la lista di lemmi tramite le entità "professioni storiche" (anche note come "professioni abbandonate", Q16335296), oppure i titoli nobiliari e le particelle onorifiche (Q355567 e Q355505).

Per ottenere ulteriori risultati, è già stato implementato anche un altro meccanismo: la lista che si ottiene tramite le istanze dell'entità "professione" è infatti una lista parziale di veri e propri mestieri considerati da *WikiData* come tali. È tuttavia possibile che esistano altri mestieri particolarmente nuovi o rari, oppure che non sono propriamente dei mestieri ma che in alcuni casi potrebbero essere considerati tali. Perciò, non dispongono di una propria pagina dedicata su *WikiData*, ma possono essere identificati perché qualche persona li ha esercitati nell'arco della Storia. Per identificarli, la soluzione scelta consiste nel prendere un intervallo di anni e scorrere l'elenco di tutti gli esseri umani nati in quell'intervallo, andando a vedere cosa venga dichiarato nel loro campo *professione* (P106, *occupation*). In buona parte si tratta di mestieri già identificati tra le entità *professione*, ma con questo meccanismo è possibile identificare qualche termine *borderline*. Per esempio, il lemma "coltellinaio" non è presente come istanza di una *professione*, probabilmente perché troppo specifico, ma potrebbe tornare utile in alcuni particolari *corpora*, ed è stato rintracciato grazie a questo meccanismo perché è indicata come professione per James Black,[18] coltellinaio statunitense nato il 1° maggio 1800. In questo specifico esempio, la professione è stata identificata anche nell'enciclopedia *Sapere*. Invece, il mestiere di coppiere (servitore dedicato a versare il vino durante i banchetti,[19] non è presente nemmeno nella lista di *Sapere*, e non viene ritrovato con la semplice ricerca delle professioni su *Wikipedia* perché è troppo specifico. Può però essere scoperto cercando i mestieri svolti dagli esseri umani nati tra il 1300 e il 1400 (segnalato come mestiere per Gaston of Foix-Béarn):[20]

18 https://www.wikidata.org/wiki/Q6129818 (consultato il 10/12/2021).

19 https://www.wikidata.org/wiki/Q1075791 (consultato il 10/12/2021).

20 https://www.wikidata.org/wiki/Q3758791 (consultato il 10/12/2021).

```
SELECT ?item ?itemLabel ?occupationLabel ?genderLabel ?citizenshipLabel ?
birth
WITH {
SELECT ?item ?gender ?occupation ?citizenship ?birth WHERE {
  ?item wdt:P31 wd:Q5; #P31 means we are looking for items that are
instances of the entity
 wdt:P21 ?gender;
 wdt:P106 ?occupation;
 wdt:P27 ?citizenship;
 wdt:P569 ?birth. hint:Prior hint:rangeSafe true.
  filter (?birth > "1300-00-00"^^xsd:dateTime && ?birth < "1400-00-
00"^^xsd:dateTime)
}
} AS %results
WHERE {
INCLUDE %results.
?item rdfs:label ?itemLabel. FILTER( LANG(?itemLabel)="it" )
?gender rdfs:label ?genderLabel. FILTER( LANG(?genderLabel)="en" )
?occupation rdfs:label ?occupationLabel. FILTER( LANG(?
occupationLabel)="it" )
?citizenship rdfs:label ?citizenshipLabel. FILTER( LANG(?
citizenshipLabel)="en" )
}
```

La soluzione scelta per evitare il *timeout* consiste nel paginare le ricerche in base all'anno di nascita, facendo quindi una diversa *query* per ogni anno di nascita delle persone. Tra l'altro, naturalmente le persone censite aumentano man mano che ci si avvicina all'età contemporanea quindi, se tra il 1300 e il 1400 si può tranquillamente fare una ricerca unica, ha senso limitare la singola *query* alle persone nate tra il 1800 e il 1801 con questo filtro per data:

```
 filter (?birth > "1800-00-00"^^xsd:dateTime && ?birth < "1801-00-
00"^^xsd:dateTime)
```

Un altro meccanismo che potrebbe tornare utile per restringere il campo consisterebbe nell'applicare filtri agli esseri umani ricercati in base alle loro proprietà.[21] Per esempio, se si cercano specificamente professioni svolte da donne può avere senso restringere il campo ai soli esseri umani di genere femminile. È un meccanismo utilizzabile anche per fare una ricerca inversa e trovare le persone che hanno svolto una determinata professione. Per esempio, la *query*

```
SELECT ?item ?itemLabel ?occupationLabel ?genderLabel ?citizenshipLabel ?
birth
WITH {
```

21 Si vedano "properties for this type": https://www.wikidata.org/wiki/Q5#P1963 (consultato il 10/12/2021).

```
SELECT ?item ?gender ?occupation ?citizenship ?birth WHERE {
 ?item wdt:P106 wd:Q36180;
 wdt:P21 ?gender;
 wdt:P106 ?occupation;
 wdt:P27 ?citizenship;
 wdt:P569 ?birth. hint:Prior hint:rangeSafe true.
  filter (?birth > "1800-00-00"^^xsd:dateTime && ?birth < "1801-00-
00"^^xsd:dateTime)
}
} AS %results
WHERE {
INCLUDE %results.
?item rdfs:label ?itemLabel. FILTER( LANG(?itemLabel)="it" )
?gender rdfs:label ?genderLabel. FILTER( LANG(?genderLabel)="en" )
?occupation rdfs:label ?occupationLabel. FILTER( LANG(?
occupationLabel)="it" )
?citizenship rdfs:label ?citizenshipLabel. FILTER( LANG(?
citizenshipLabel)="en" )
}
```

troverebbe tutte le persone che hanno esercitato la professione di scrittore[22] tra chi è nato nell'anno 1800. La condizione di essere umano (Q5) è omessa perché si dà per scontato che solo gli esseri umani esercitino professioni, e quindi non serve specificare che l'entità ricercata debba anche essere una persona. Questo consente di velocizzare la *query*.

Potremmo utilizzare le *query* avanzate anche per fare una statistica delle professioni censite entro un certo periodo storico, in modo da confrontare la distribuzione "reale" dei mestieri con la distribuzione di questi mestieri nello stesso corpus. Per esempio, con la *query*

```
SELECT ?occupationLabel (COUNT ( ?occupationLabel) AS ?count)
WITH {
SELECT ?item ?gender ?occupation ?citizenship ?birth WHERE {
  ?item wdt:P31 wd:Q5; #P31 means we are looking for items that are
 instances of the entity
 wdt:P21 ?gender;
 wdt:P106 ?occupation;
 wdt:P27 ?citizenship;
 wdt:P569 ?birth. hint:Prior hint:rangeSafe true.
  filter (?birth > "1820-00-00"^^xsd:dateTime && ?birth < "1821-00-
00"^^xsd:dateTime)
}
} AS %results
WHERE {
INCLUDE %results.
?item rdfs:label ?itemLabel. FILTER( LANG(?itemLabel)="it" )
?gender rdfs:label ?genderLabel. FILTER( LANG(?genderLabel)="en" )
```

22 https://www.wikidata.org/wiki/Q36180 (consultato il 10/12/2021).

```
?occupation rdfs:label ?occupationLabel. FILTER( LANG(?
occupationLabel)="it" )
?citizenship rdfs:label ?citizenshipLabel. FILTER( LANG(?
citizenshipLabel)="en" )
} GROUP BY ?occupationLabel
ORDER BY DESC(?count) ?occupationLabel
```

si scopre che i mestieri più esercitati da persone nate nel 1820 (e dunque presumibilmente in attività dopo il 1840) censite su *WikiData* erano *politico, pittore, avvocato, scrittore, giornalista, professore universitario, ufficiale, giudice, medico, poeta*. Mentre nel corpus utilizzato per questo studio vediamo che i mestieri più citati nelle opere pubblicate nel 1840 sono *re, giudice, servitore, medico, scrittore, contadino, barbiere, mercante, deputato, e poeta*. Possiamo riassumere e visualizzare i risultati con un semplice diagramma di Venn (Figura 1):

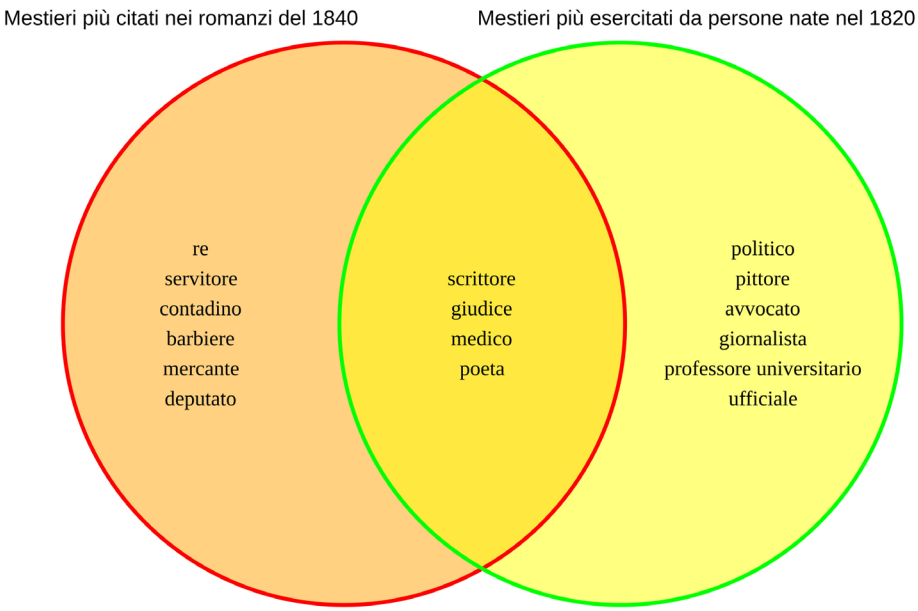

Figura 1 – Diagramma di Venn dei mestieri più citati nelle opere del corpus del 1840 e di quelli più esercitati da persone nate nel 1820 secondo WikiData

Chiaramente una ricerca su un singolo anno non ha rilevanza statistica, ma suggerisce che sarebbe possibile fare una classifica più ampia e capire se le opere di narrativa parlino di mestieri in voga nell'epoca in cui sono stati scritti o se tendano a parlare d'altro (e, soprattutto, dopo quanti anni dalla prima comparsa di un nuovo mestiere nella società si inizi a parlarne nelle opere letterarie). La singola *query* va eseguita su un solo anno al massimo, per evitare che sia troppo lunga: il *group by* viene eseguito dopo l'ottenimento dei risultati e non aiuta il *query planner* a trovare

una strada più rapida, quindi non riduce il tempo di esecuzione della *query*. Tuttavia, è possibile eseguire più *query* e sommare i risultati in un'unica tabella.

È chiaro che il limite di questo meccanismo di ricerca sia il fatto che le persone censite in un'enciclopedia siano solo persone in qualche modo "famose", quindi non si troveranno molti mestieri "umili", che però possono essere facilmente identificati con il meccanismo di ricerca descritto in precedenza e attualmente impiegato per il vocabolario di *ExSTRA*, cioè le istanze dell'entità *professione*.

7. Risultati parziali

In questa sezione vedremo a titolo esemplificativo alcuni risultati ottenuti grazie alle procedure che abbiamo illustrato in precedenza. Oltre a dare un'idea di come sono cambiati i mestieri nel periodo rappresentato dal corpus preso in analisi, i grafici che abbiamo realizzato permettono di testare la funzionalità e di evidenziare alcuni problemi dovuti a strumenti e metodi che abbiamo scelto di impiegare.

7.1. Mestieri e professioni più frequenti

In primo luogo, osserviamo le occorrenze dei 20 lemmi riferiti a mestieri e professioni più frequenti in tutto il corpus (Tabella 1).

Lemma	Occorrenze
capo	6032
re	3276
guardia	1856
medico	1500
cavaliere	1278
maestro	1276
poeta	1220
avvocato	1056
giudice	686
lettore	651
contadino	617
stella	508
ingegnere	454
servitore	450
regina	438
cameriere	413
serva	406
artista	402
cacciatore	393
segretario	376

Tabella 1 – Lista dei 20 lemmi più frequenti del corpus

La tabella non dà informazioni sui cambiamenti temporali, ma restituisce già l'immagine preliminare dei lavori menzionati più spesso nel corpus. I risultati sono piuttosto prevedibili: si trovano infatti mestieri comuni, come *medico, maestro, avvocato, ingegnere, segretario, servitore, contadino* e *giudice*.

Dall'altra parte, i lemmi nella tabella evidenziano problemi dovuti alla composizione alla lista da cui sono estratti. Emergono infatti parole come *stella* e *capo*: è ovvio però che si tratta di casi non pertinenti a mestieri e professioni. Nel primo caso (esempi 1, 2, 3 e 4), i lemmi osservati in gran parte dei contesti fanno riferimento agli astri e non alle dive dello spettacolo; nel secondo (esempi 5, 6, 7 e 8), alla parte anatomica e non alla persona che comanda qualcosa.

1. "Il fiume portava seco la notte con tutte le sue stelle." (D'annunzio, *Il Fuoco*)
2. "[…] e mentre il profumo dolce della violacciocca si mesce all'odore acre dell'euforbia, le prime stelle salgono sopra il Monte." (Deledda, *Canne al vento*)
3. "E Forestina chiedèa: - Babbo, e lassù, di là dalle stelle, che c'è?" (Dossi, *La colonia felice*)
4. "Si fermò sull'ultimo limite di questo, quando non vide più dinanzi a sé che il mare bruno ed immenso, su cui scintillavano le stelle." (Verga, *Una peccatrice*)

5. "Egli si fece rosso, e scosse energicamente il capo." (Castellani Fantoni Benaglio, *Mia*)
6. "E il burattino fece col capo e colle mani un segno come dire: «Non ne ho»." (Collodi, *Pinocchio*)
7. "Ella rimane a capo chino." (Rovetta, *Casta diva*)
8. "L'orgogliosa e la grande Inghilterra, dominatrice di mezzo mondo, ha piegato talvolta il capo dinanzi a noi, tigri di Mompracem." (Salgari, *Alla conquista di un impero*)

Inoltre, occorre una riflessione su cosa siano un mestiere o una professione:[23] *stella* e *capo*, anche nel caso del calco semantico di *star* o del ruolo, sono considerabili lavori? Altri casi dubbi dello stesso tipo riguardano titoli come *re* e *regina*, che, benché rappresentino attività continuative, non sono comunemente associati a lavori, ma forse piuttosto a status sociali, oltre che a titoli nobiliari. Infine, sebbene il *lettore* possa essere "l'insegnante incaricato di esercitare praticamente gli studenti in una lingua straniera" (Vocabolario Treccani), in realtà le opere del corpus si riferiscono piuttosto al loro pubblico (esempi 9. 10, 11 e 12).

9. "C'era una volta… – Un re! – diranno subito i miei piccoli lettori" (Collodi, *Pinocchio*)
10. "Che importa a me d'avere, per esempio, cento lettori nell'isola dei Sardi ed anche dieci ad Empoli e cinque, mettiamo, ad Orvieto?" (D'Annunzio, *Il piacere*)
11. "Ricorderanno i lettori che siamo nella seconda quindicina di febbraio" (Garibaldi, *Clelia ovvero il governo dei preti*)
12. "E ora i nostri giovani lettori si trasportino con la fantasia a Collebianco […]" (Baccini, *Il principino*)

23 Si tratta di una questione complessa che può essere affrontata da diversi punti di vista. Per esempio, per la definizione a fini sociologici e statistici, rimandiamo a Gallo e Scalisi 2013: 45-48.

Per avere un quadro[24] su come sono cambiati i mestieri nel corso del tempo, è possibile estrarre dati riguardanti periodi delimitati grazie ai filtri che abbiamo descritto nel § 4. I grafici in Figura 2, Figura 3 e Figura 4, che inseriamo qui a titolo esemplificativo,[25] contengono una rappresentazione dei 20 lemmi relativi a mestieri e professioni che occorrono più spesso nei decenni 1840-1849, 1880-1889[26] e 1910-1920 (e oltre), cioè il primo rappresentativo,[27] l'ultimo periodo, e il decennio a metà tra i due.

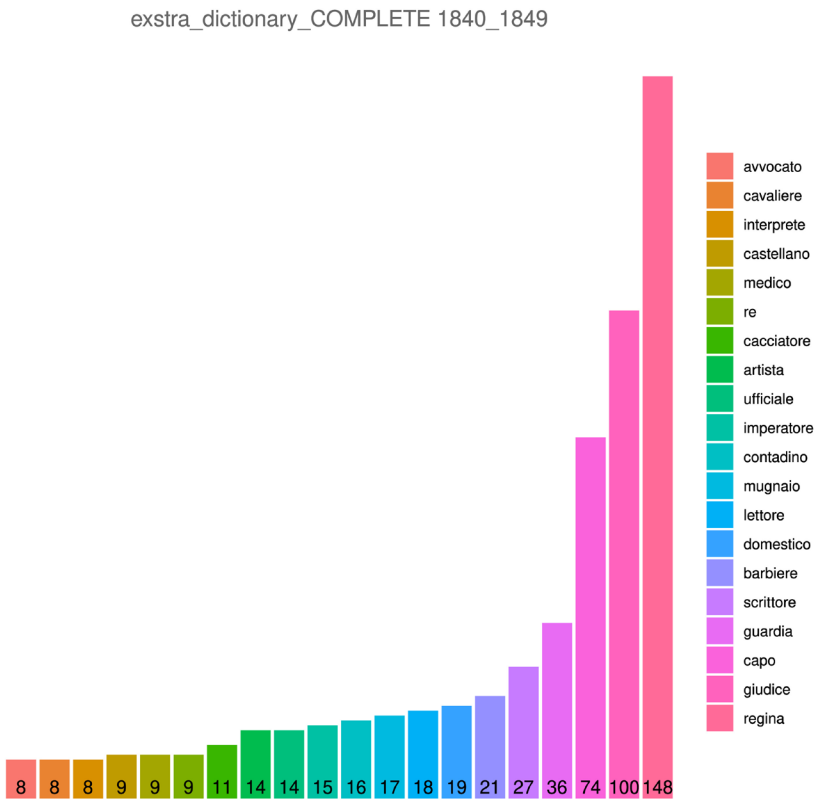

Figura 2 – Occorrenze dei 20 lemmi più frequenti nel decennio 1840-1849

24 Le rappresentazioni che seguono non sono significative dal punto di vista statistico perché non tengono conto della dimensione dei subcorpora, nella distribuzione dei lemmi nei singoli testi eccetera.

25 I grafici di tutti gli altri decenni sono disponibili sul repository di ExSTRA all'indirizzo https://github.com/flometis/ExSTRA (consultato il 10/12/2021).

26 Si tratta anche del decennio con il maggior numero di opere (21).

27 Nel corpus sono presenti anche opere pubblicate in precedenza, ma si tratta solo di due casi: l'analisi non sarebbe quindi rappresentativa.

Escludendo i risultati poco pertinenti di *capo* e *lettore*, che abbiamo già commentato e che compariranno anche nei due grafici successivi, gran parte dei risultati della Figura 2 rientrano sostanzialmente in tre categorie: quella lavoratrice più legata a mestieri manuali (*barbiere, domestico, mugnaio, contadino, cacciatore*), quella borghese (*giudice, medico, avvocato*), e infine compaiono titoli nobiliari (*regina, re, imperatore*) o altri titoli in qualche modo a essi connessi (*cavaliere, castellano*). *Interprete* è invece un altro caso di parola polisemica che non ha a che vedere con la professione legata all'intermediazione tra lingue diverse. Ciò emerge osservando i contesti d'uso del lemma (tutti contenuti in *Storia di una colonna infame* di Manzoni), in cui appare chiaro che il significato è piuttosto legato a colui o colei che spiega qualcosa, per lo più in ambito giuridico, come negli esempi 13 e 14:

13. "La questione dev'esser dunque, se i criminalisti interpreti (così li chiameremo, per distinguerli da quelli ch'ebbero il merito e la fortuna di sbandirli per sempre) sian venuti a render la tortura più o meno atroce di quel che fosse in mano dell'arbitrio […]"
14. "Anzi la ragione di quelle precauzioni, la ricavavano gl'interpreti dalla legge medesima […]".

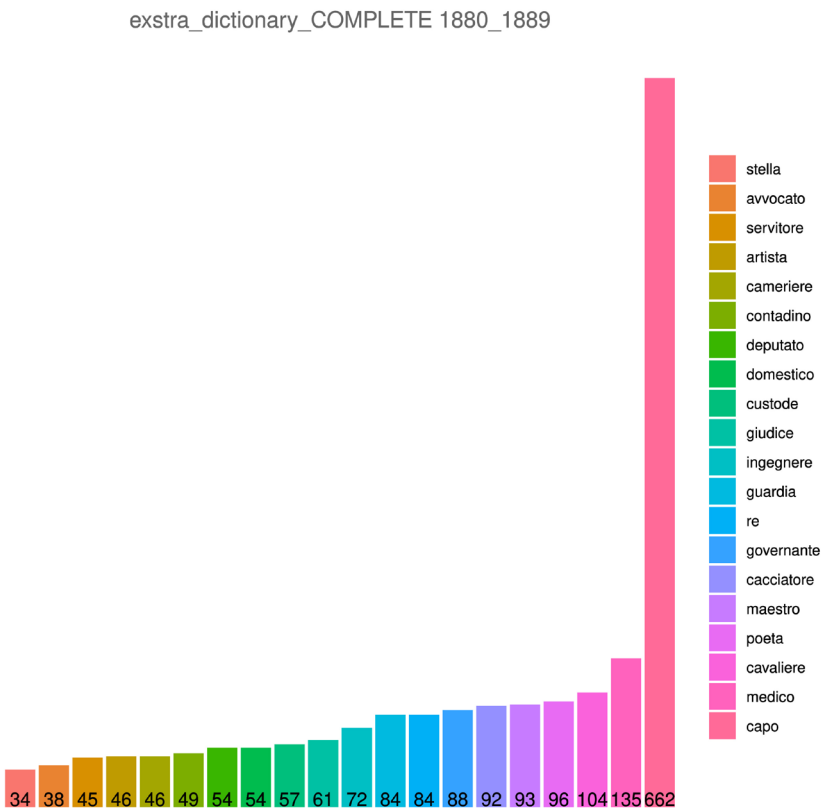

Figura 3 – Occorrenze dei 20 lemmi più frequenti nel decennio 1880-1889

La Figura 3 non evidenzia particolari cambiamenti nel decennio 1880-1889 rispetto al decennio 1840-1849: anche in questo caso si distinguono tre classi principali e, al di là della scomparsa di *regina* e della presenza di altri mestieri considerati più "umili" (*domestico, cameriere*) e borghesi (*medico, ingegnere, giudice*), si nota l'ingresso di mestieri collegati all'ambito artistico e culturale: oltre a *maestro*, compaiono *artista* e *poeta*.

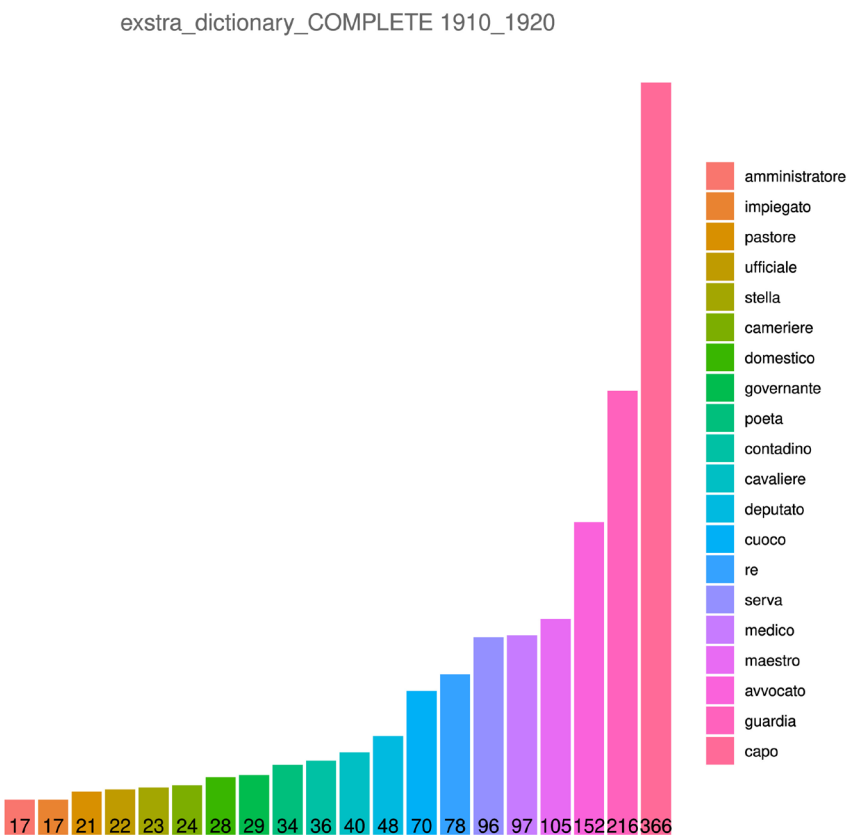

Figura 4 – Occorrenze dei 20 lemmi più frequenti nel periodo 1910-1920 (e oltre)

Anche l'ultimo periodo, rappresentato nella Figura 4, non vede grandi cambiamenti, a parte l'ascesa di *avvocato*. Lo scenario di mestieri e professioni nella narrativa italiana sembra quindi piuttosto costante nel tempo per quanto riguarda i mestieri descritti, con qualche variazione nel numero di occorrenze e nell'alternanza tra classi sociali.

Un altro filtro permette di confrontare i mestieri citati più spesso da uomini e donne (Figura 5 e Figura 6).

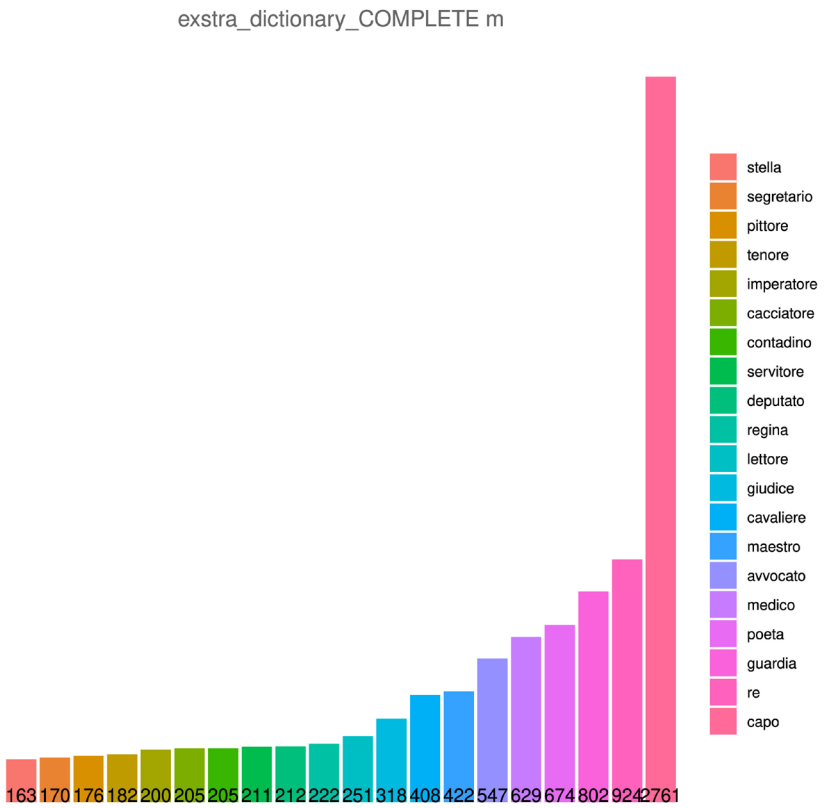

Figura 5 – Occorrenze dei 20 lemmi più frequenti nelle opere scritte da uomini

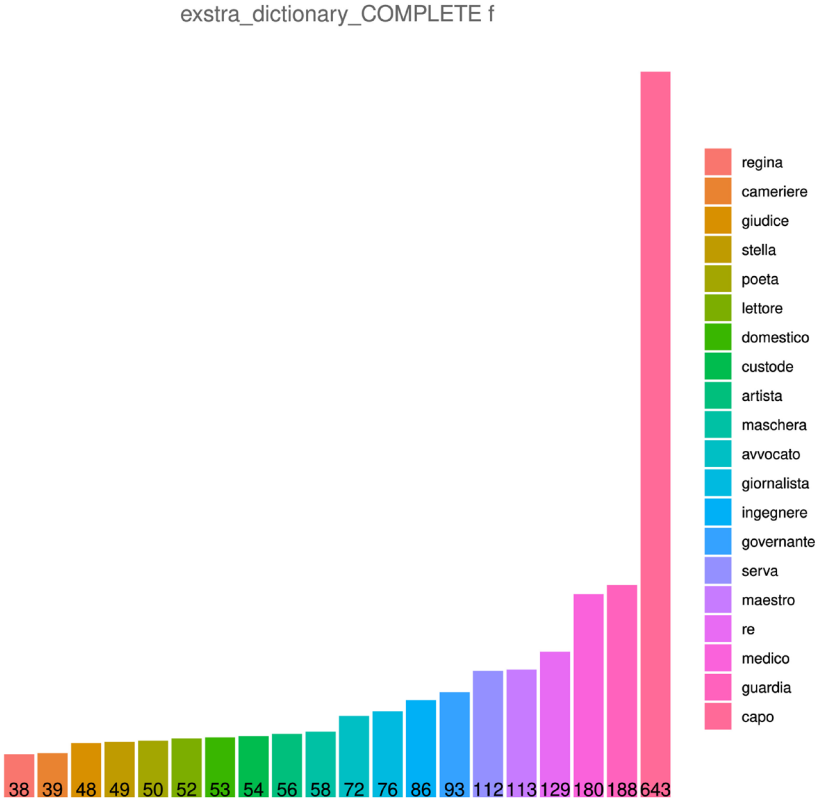

Figura 6 – Occorrenze dei 20 lemmi più frequenti nelle opere scritte da donne

Anche in questo caso non si notano particolari divergenze tra le due categorie, benché si possa osservare una maggiore tendenza nelle opere di donne a citare professioni più legate all'ambito domestico (*serva, governante, domestico, cameriere*). Non si tratta tuttavia di divergenze sostanziali. La parola *maschera* è evidentemente un falso positivo, poiché i lemmi non sono collegati alla figura che accompagna il pubblico in cinema e teatri, ma all'accessorio che copre il viso o ai travestimenti in generale (esempi 15, 16, 17 e 18).

15. "Lo stesso giovane che l'ultima notte di carnevale si è qui ricoverato in costume da maschera." (Invernizio, *La trovatella di Milano*)
16. "[…] Nora aveva seguito il veemente ed appassionato sfogo del banchiere: in quel momento le appariva trasformato come se l'enorme mucchio d'oro smosso dalle sue parole lo coprisse di una prodigiosa maschera abbagliante di riflessi gialli […]" (Di Luanto, *Per il lusso*)
17. "[…] il piccolo domino nero si voltò ai due giovani e li guardò a traverso i fori della maschera." (Serao, *La mano tagliata*)
18. "Così nera, col viso scuro come una maschera, aveva davvero qualche cosa di diabolico […]" (Deledda, *La madre*)

Confrontiamo infine due autori a ulteriore verifica della flessibilità dei filtri. Vediamo le differenze tra i lavori di Neera (Figura 7) e quelli di Emilio Salgari (Figura 8), un'autrice e un autore entrambi vissuti a cavallo di Otto e Novecento, ma che si occupavano di generi diversi: la prima rappresentava la condizione femminile, il secondo è invece celebre scrittore di romanzi di avventura ambientati in luoghi esotici. Nel corpus che abbiamo analizzato si trovano tre romanzi di ciascuno.

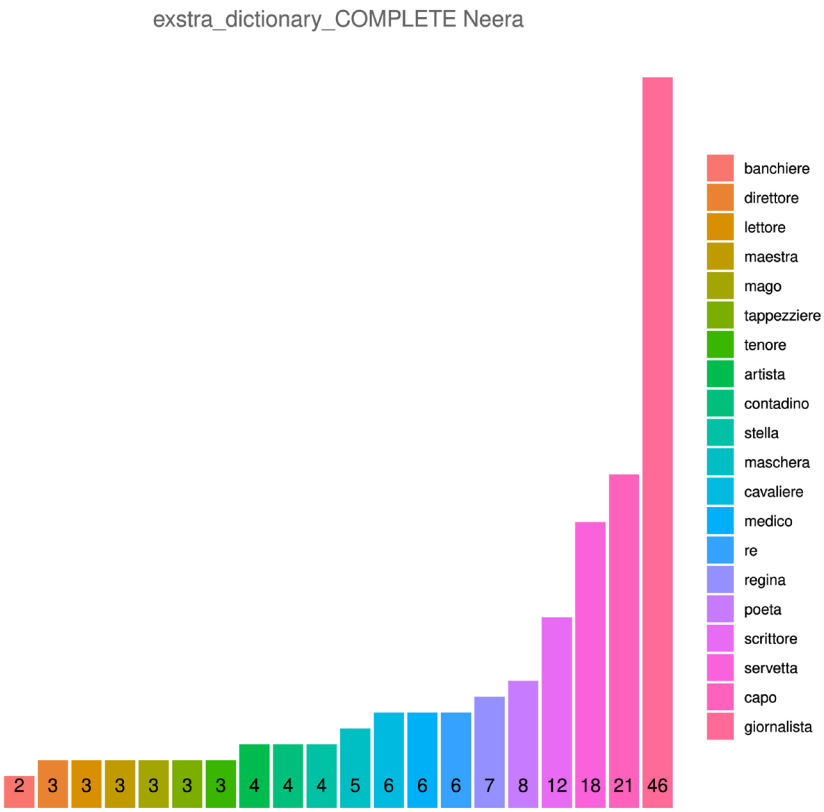

Figura 7 – Occorrenze dei 20 lemmi più frequenti nei romanzi di Neera

I mestieri citati da Neera si rifanno a diversi ambiti: quello di corte (*re, regina*); quello artistico (*scrittore, poeta, artista, tenore*); quello borghese (*giornalista, medico*); e quello legato alla classe lavoratrice (*servetta, contadino, tappezziere*). Si notano anche mestieri chiaramente al femminile, come *servetta*, *regina* e *maestra*.[28]

28 Rimandiamo al § 8 per una riflessione sui mestieri al femminile nella lista che abbiamo utilizzato nel nostro studio.

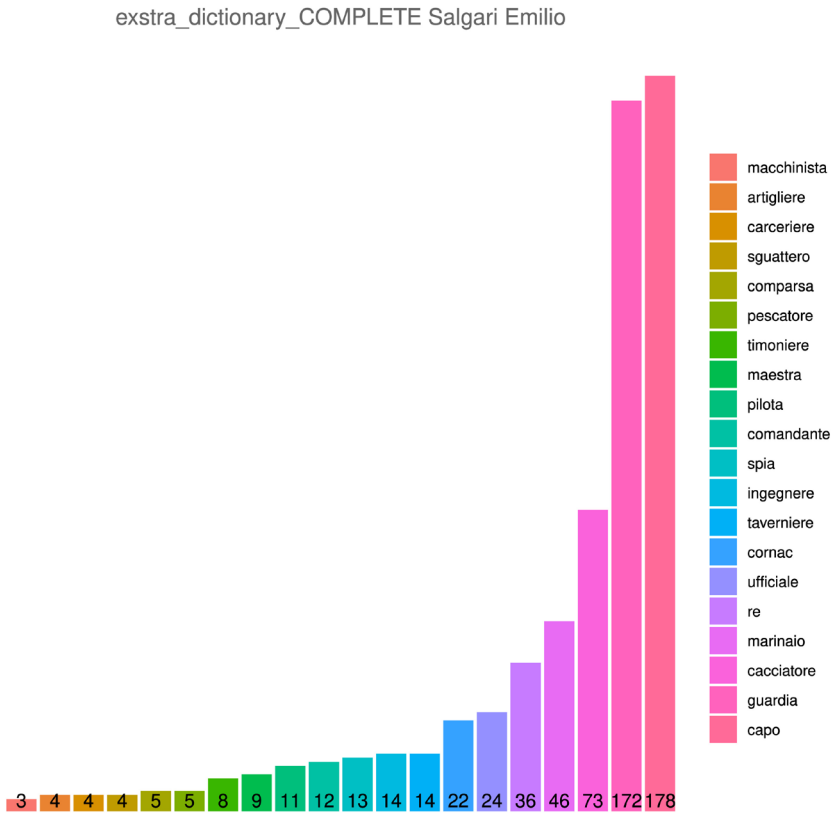

Figura 8 – Occorrenze dei 20 lemmi più frequenti nei romanzi di Emilio Salgari

In Salgari, invece, spiccano mestieri legati al mondo del mare e all'avventura (*marinaio*, *ufficiale*, *spia*, *comandante*, *pilota*, *timoniere*, *pescatore*) che ci si aspetterebbe di trovare nei romanzi dell'autore. L'ambientazione esotica si collega invece a *cornac*, parola che designa il conducente indiano di elefanti.

7.2. Analisi multivariata

L'analisi delle corrispondenze consente di individuare dei fattori impliciti che sottendono a tutti i corpus potenzialmente coinvolti nel corpus, andando perciò, ad es. a evidenziare i mestieri comuni a più testi e i mestieri che invece sono tipici unicamente di un testo solo. Intuitivamente, come si può vedere da Figura 9, il mestiere "cocchiere" è comune a *Il principino* e *La città del prossimo*, ma non compare ne *Il passaggio*. Il mestiere "poeta", invece, compare in *Il passaggio* e *La carità del prossimo*, ma non in *Il principino*. L'analisi delle corrispondenze considera le frequenze standardizzate, e dunque non è sensibile alla potenziale lunghezza dei testi (sempre intuitivamente: se un testo è molto lungo, potenzialmente

un mestiere è citato più spesso). Dunque, facendo sempre riferimento a Figura 9, possiamo ragionevolmente dire che il mestiere "cocchiere" sia citato più frequentemente in *Il principino*, rispetto quanto sia in *La carità del prossimo*, anche ipotizzando che il secondo sia (molto) più lungo del primo.

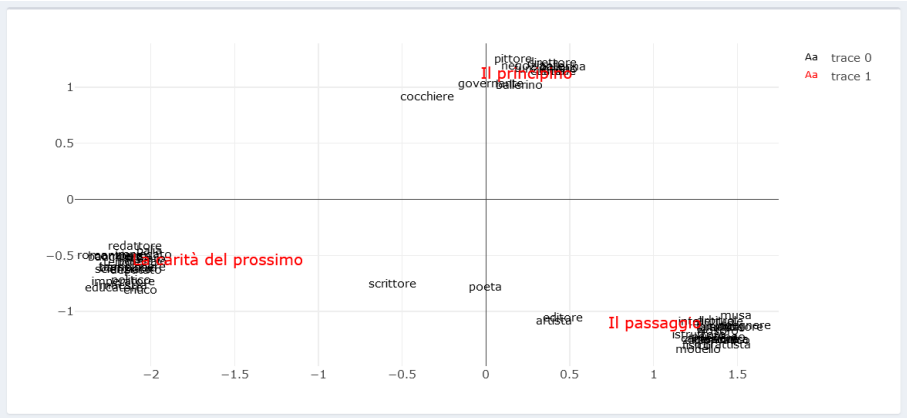

Figura 9 – Analisi delle corrispondenze per i romanzi Il passaggio, Il principino e La carità del prossimo

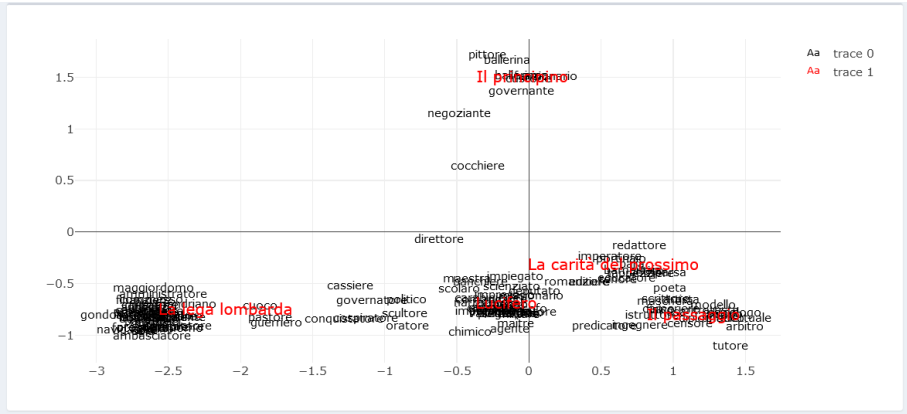

Figura 10 – Grafico delle analisi delle corrispondenze, con un numero più alto di romanzi

L'analisi delle corrispondenze consente anche di analizzare e comparare tra loro contemporaneamente più opere o subcorpora, superando i tradizionali limiti imposti da analisi tabellari o diagrammi a barre o diagrammi a dispersione. Questo consente anche di individuare opere che sono più "simili" tra loro nel nominare i vari mestieri. Come possiamo vedere in Figura 10, *Lucifero*, *La carità del*

prossimo e *Il passaggio* risultano essere molto simili tra loro, perché condividono un set di mestieri. La similarità tra testi può essere sempre dedotta dalla vicinanza grafica delle loro etichette: più due opere si trovano vicine tra loro nel grafico, più essi condividono un insieme di mestieri.

Nell'interpretare questi grafici bisogna sempre tenere a mente che lemmi identificati erroneamente come mestieri possono influenzare le analisi. È per questo motivo che il grafico consente di eliminare manualmente questi casi, come illustrato in Figura 11. Nel prossimo paragrafo si illustrano i comandi che consentono di interagire con i grafici.

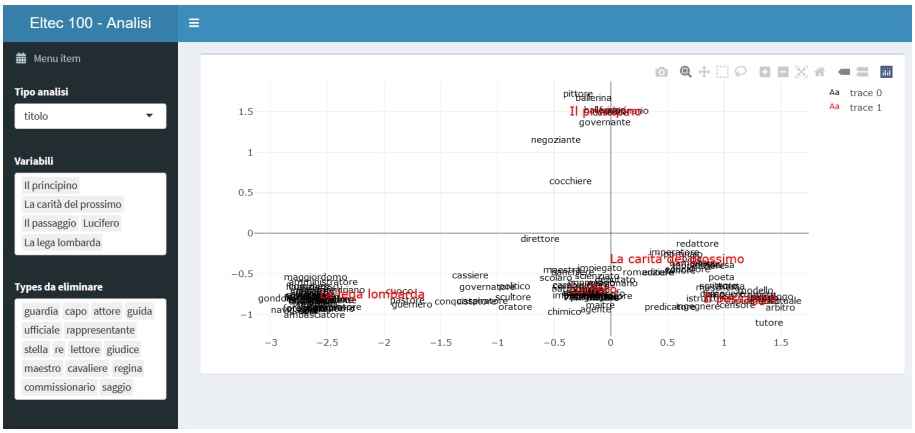

Figura 11 – Illustrazione dell'interfaccia grafica

L'interfaccia grafica consente agli utenti di interagire con diverse caratteristiche del corpus. Dal menù di navigazione a sinistra della pagina si possono selezionare diverse opzioni di visualizzazione, nella fattispecie:

- il subcorpus da analizzare (titolo, sesso, anno, decennio, autore);
- elementi da considerare (variabili, ovvero i singoli elementi che compongono il subcorpus selezionato al passo precedente);
- *types* da eliminare (quali *types* sono da eliminare dalle analisi).

Ad esempio, volendo confrontare tra loro diversi decenni, in modo da isolare i mestieri che caratterizzano gli anni tra il 1840 e 1849 dagli anni che vanno dal 1870 al 1879 e quelli che vanno dal 1900 e il 1909, dalla voce "Tipo di analisi" si dovrà selezionare "decennio". Dalla voce "Variabili" si dovrà selezionare "1840-1849", "1870-1879" e "1900-1910". La voce "Types da eliminare" consente di togliere dai grafici i mestieri identificabili come falsi positivi.

Il risultato di questa selezione è visibile in figura 12.

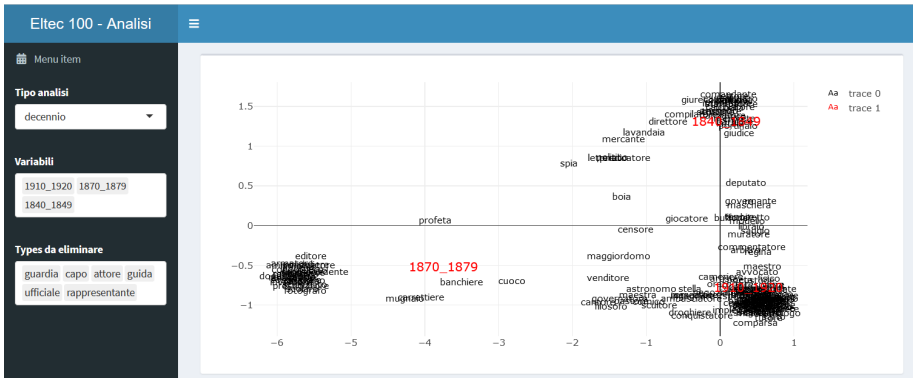

Figura 12 – Analisi per decennio

Volendo invece confrontare tra loro gli scrittori in base al loro sesso, dalla voce "Tipo di analisi" si dovrà selezionare "sesso", mentre dalla voce "Variabili" si dovranno selezionare le voci "m" e "f". Si veda l'esempio di figura 13. Si noti anche come, data la presenza di due uniche variabili, il grafico non rappresenti più l'esito di un'analisi delle corrispondenze, ma un semplice diagramma a dispersione delle frequenze standardizzate dei *types* nei due subcorpora. Sempre facendo riferimento a Figura 13, ad es. il *type* "avvocato" compare con una frequenza relativa di 0,004 nel subcorpus maschile, e una frequenza relativa di 0,002 nel subcorpus femminile.

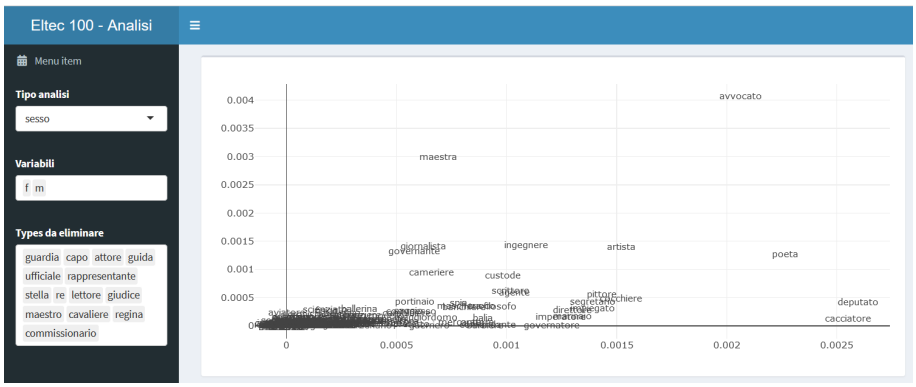

Figura 13 – Analisi per sesso

Come si può notare dalle immagini, a volte si individuano delle "nuvole di parole" così dense da risultare poco leggibili. Le "nuvole di *types*" sono caratterizzate da profili di frequenza estremamente simili tra loro, verosimilmente frequenze basse. Tali *types* a volte sarebbero perfettamente sovrapposti tra loro (si pensi agli

hapax) – cosa che renderebbe di fatto illeggibili i grafici. In questa applicazione le coordinate dei punti sovrapposti sono state modificate in modo impercettibile e casuale, evitando così le sovrapposizioni perfette delle etichette. Infine, l'interfaccia consente di aumentare la risoluzione per esplorare più in dettaglio il contenuto delle "nuvole di *types*", come illustrato in Figura 14.

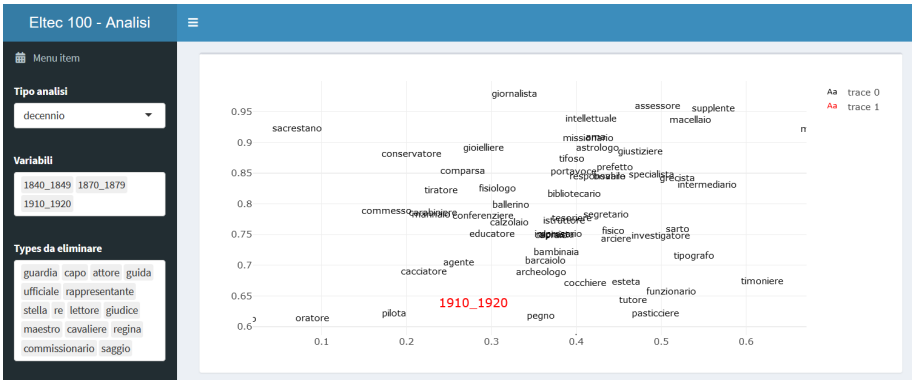

Figura 14 – La "nuvola di types" che caratterizza il decennio 1910-1920 rispetto ai decenni 1840-1849 e 1870-1879

Riprendiamo infine il confronto tra i due autori Neera e Salgari (Figura 15 e Figura 16).

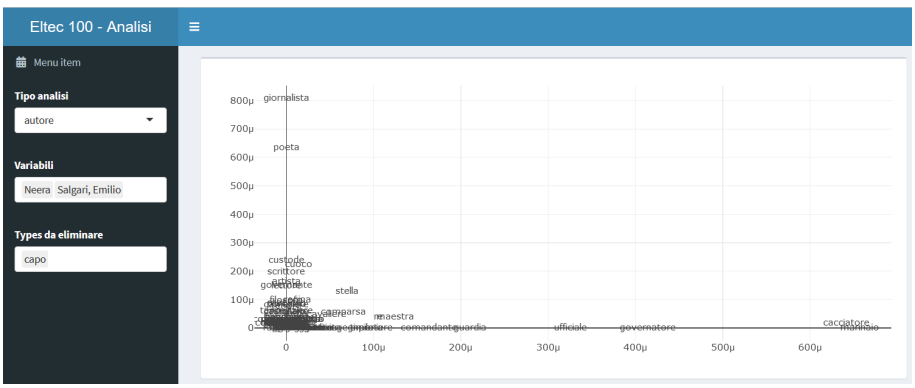

Figura 15 – Analisi per autore dei romanzi di Neera e Salgari

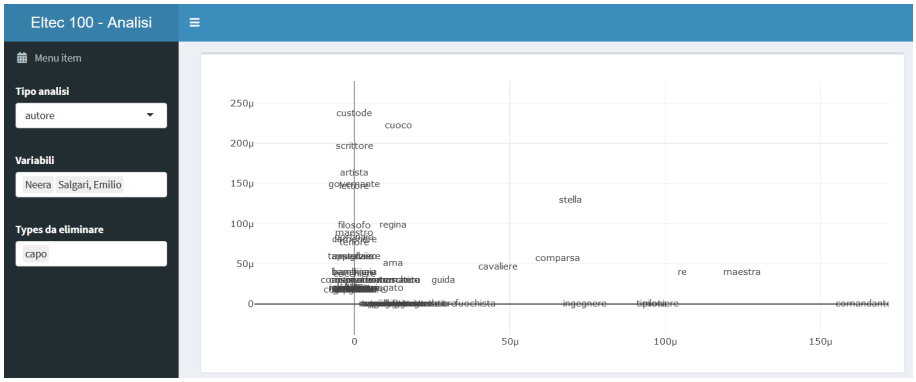

Figura 16 –Zoom sull'analisi per autore dei romanzi di Neera e Salgari

Rispetto ai dati riassunti nella Figura 8, sono ancora più chiare le differenze tra i due autori e trovano conferma alcune ipotesi già avanzate nel § 7.1. In Neera, infatti, si ripropongono i mestieri domestici (a cui si affiancano anche *cuoco* e *custode*, assenti nella Figura 8), e quelli del settore culturale, ai quali si aggiunge *filosofo*. In Salgari si affermano l'ambito navale e militare, con un'ulteriore conferma in *fuochista*, non presente nella Figura 8. Potrebbe invece sorprendere la maggiore presenza di *maestra*, che però fa riferimento all'ambito navale e non al mestiere (esempi 19 e 20):

19. "[…] attraversò quatto quatto il ponte, urtando forte col gomito un indiano che stava chiudendo il boccaporto di maestra." (Salgari, *I misteri della giungla nera*)
20. "Si slanciarono verso i paterazzi, s'inerpicarono fino alla estremità dell'albero di maestra." (Salgari, *I pirati della Malesia*)

8. Conclusioni: aspetti, problemi, soluzioni e prospettive di ricerca

Dai dati preliminari su mestieri e professioni, che abbiamo individuato nel corpus per testare il nostro sistema, si può desumere che i riferimenti al lavoro presenti nelle opere prese in analisi sono prevedibili, cioè sono citati mestieri comuni, alcuni dei quali diffusi ancora oggi. Questi sono tutto sommato costanti nel periodo rappresentato dal corpus, con alcune leggere variazioni nel numero delle occorrenze. Alcune differenze emergono dal confronto tra scrittori e scrittrici, che vede la prevalenza di mestieri legati all'ambito domestico nelle opere femminili. La comparazione di singoli autori mette invece in luce divergenze più marcate: il paragone tra Neera e Salgari mostra che i lavori menzionati dall'autrice e dall'autore sono piuttosto diversi, ed evidenziano anche i diversi generi letterari in cui si inseriscono i due scrittori. Si tratta tuttavia di dati ancora parziali e grezzi che saranno ulteriormente filtrati e approfonditi, soprattutto dal punto di vista statistico, con metodi che permettano di valutare la distribuzione

nel tempo, come ad esempio il *curve clustering* utilizzato da Sciandra, Trevisani e Tuzzi (2021) in questo stesso volume.[29]

La trattazione e i dati che abbiamo analizzato suggeriscono che il sistema che abbiamo proposto funziona e presenta diversi vantaggi: permette infatti di estrarre e incrociare diversi dati, a partire da liste consistenti, in *corpora* anche di grandi dimensioni, impossibili da spogliare a mano, una prerogativa dei metodi automatici di estrazione dei dati testuali. I metodi che abbiamo proposto sono inoltre replicabili e le diverse fasi potranno, in futuro, essere applicate a progetti di ricerca con finalità simili ma basati su diverse tipologie testuali, per ottenere dati utili ad ampliare le conoscenze linguistiche, storiche, sociologiche, statistiche e letterarie attualmente disponibili sia in prospettiva diacronica, come nel caso di questo studio, sia in prospettiva sincronica. La struttura flessibile del sistema, il ricorso a una fonte multilingue come *WikiData* e l'uso di uno strumento come *UdPipe*, che è disponibile per un'ampia gamma di lingue per le quali sono usati sempre gli stessi sistemi di annotazione, predispongono inoltre l'applicazione a progetti che riguardano lingue diverse dall'italiano, anche a scopo di confronto, per esempio in ambito traduttivo.

Per contro, nonostante gli aspetti positivi, occorre fare alcune considerazioni e vagliare alcuni problemi metodologici e legati agli strumenti che abbiamo impiegato, che sono emersi durante la fase di progettazione del sistema e durante l'analisi dei dati. Tali problemi sono utili a contestualizzare e interpretare meglio le informazioni ottenute, ma soprattutto a migliorare il sistema in vista di prossimi utilizzi.

In primo luogo, sebbene la lista di professioni e mestieri ottenuta dalla fusione dei dati di *Sapere* e *WikiData* sia piuttosto nutrita, anche grazie all'integrazione delle due fonti, presenta senza dubbio alcuni difetti. Un problema è sicuramente rappresentato dalla presenza di falsi positivi, come *stella* o *capo*, piuttosto ricorrenti nel nostro corpus. Tuttavia, si tratta di un problema di facile soluzione: sarebbe sufficiente uno scarto manuale delle forme indesiderate (già realizzabile tramite il sistema proposto nel § 7.2), che compaiono nei risultati preliminari.

La lacuna più evidente della lista è dovuta alla prevalenza di voci al maschile singolare, che limita i lemmi al femminile ad alcuni lavori, in alcuni casi quelli storicamente condotti da donne (es. *asolaia, balia, cardatrice, indossatrice, levatrice*). La mancanza di nomi al femminile è resa ancora più problematica dal processo di lemmatizzazione. Infatti, quest'ultimo solitamente riduce i *token* alla forma base maschile e singolare, che dovrebbe corrispondere alla forma presente anche nella lista. Tuttavia, la lemmatizzazione è incostante per la natura dell'intelligenza artificiale su cui si basa: prendiamo di nuovo il caso di *domestico*, i *token* al femminile *domestica* e *domestiche* vengono lemmatizzati come *domestico* in alcuni casi, mentre in altri come *domestica*. Quest'ultima forma non figura però nella nostra lista di mestieri e professioni, quindi il lemma non viene conteggiato corretta-

29 Cfr. anche Trevisani e Tuzzi 2013; 2015; 2018.

mente. Casi simili si presentano con *mugnaia*, *sguattera*, *pastora*, *ortolana*, presenti in diversi testi, ma che il lemmatizzatore lascia alla forma femminile, assente nella lista.

L'attuale sistema di individuazione non rende inoltre possibile la distinzione tra nomi di mestieri maschili e femminili, che può però essere interessante da analizzare per individuare il momento in cui sono emersi i nomi di determinati mestieri al femminile, la loro eventuale scomparsa e la loro distribuzione del tempo. Una soluzione, che sarà oggetto di nostri studi futuri, può prevedere la creazione di un dizionario di tutte le forme del lemma (maschile, femminile, singolare e plurale) tramite sistemi automatici, come la progettazione di un sistema che interviene sulla morfologia in base alle desinenze della parola e/o che consulta dizionari online strutturati in campi, per esempio il *Wikizionario*,[30] un altro progetto targato *WikiMedia*, che contiene le forme flesse delle voci.[31] Il *Wikizionario* contiene anche una sezione di sinonimi che possono essere facilmente estratti in modo automatico: potremmo valutarne l'aggiunta alla nostra lista di lemmi. Occorre tuttavia ponderare attentamente i risultati e valutare un'attenta pulizia manuale, poiché si potrebbe aumentare l'incidenza di falsi positivi. Un approccio basato su un dizionario delle forme flesse permetterebbe anche di evitare la lemmatizzazione per ridurre gli errori a essa correlati dovuti a un corpus di *training* inadatto per quantità e tipologia dei testi.

La lista di professioni e mestieri e gli errori del lemmatizzatore causano anche un altro problema: la difficoltà di individuazione di forme marcate, innanzitutto in diacronia, particolarmente rilevanti nel nostro studio. Sebbene la lista di *Sapere* riporti molte forme antiquate, è probabile che siano assenti molti lavori non più praticati. Lo stesso vale per *WikiData*, che, come già spiegato, tende a focalizzarsi su personaggi famosi, quindi può mancare di mestieri meno recenti e più umili. Questi potrebbero eventualmente essere integrati tramite la digitalizzazione di repertori di mestieri più antichi, che, come abbiamo visto, al momento non sono disponibili in formato elettronico. Inoltre, al momento non abbiamo ancora inserito la sezione "Professioni abbandonate" di *WikiData*, che potrebbe ampliare la gamma di lavori ormai desueti attualmente a nostra disposizione.

Un ulteriore problema dovuto alla variazione nel tempo si collega alle diverse regole ortografiche dell'italiano, ancora instabili nel periodo preso in analisi. Il corpus presenta infatti grafie antiquate: un esempio è la forma *lampionajo* che si trova in *Il fu Mattia Pascal* di Pirandello, che non è stata individuata perché nella lista si trova solo la versione che rispetta le regole di scrittura attuali (*lampionaio*). Questo problema può essere risolto con un'operazione semplice (benché poco filologica) come la normalizzazione[32] a monte della grafia dei testi, che permet-

30 https://it.wiktionary.org/wiki/Pagina__principale (consultato il 10/12/2021).

31 Si veda ad esempio la pagina relativa alla parola "scienziato" https://it.wiktionary.org/wiki/scienziato (consultato il 10/12/2021).

32 Si vedano anche le considerazioni di Cortelazzo (2021) in questo volume.

terebbe di ottenere un numero maggiore di risultati pertinenti. Anche le variazioni diatopiche, particolarmente comuni in italiano e tuttora vitali,[33] possono generare forme difficilmente individuabili. Pensiamo al caso dei diversi nomi usati in Italia per il fruttivendolo: *fruttarolo, fruttaiolo, verduraio, verduriere vendifrutta, ortolano, besagnin(o), erbaiolo, erbivendolo, venditore di frutta, quello della frutta, ortofrutticolo, ortofrutta, frutta e verdura, alimentari* (D'Achille e Gassman 2017: 147-148). Infine, le forme marcate a causa di variazioni diafasiche possono influenzare il numero di risultati: un caso emblematico può essere individuato nelle decine di sinonimi – determinati, oltre che dal registro linguistico, anche dal ricorso a gerghi, dalla variazione diatopica o dalla diacronia – relative al mestiere di prostituta, assenti nella nostra lista di lavori. Anche in questi casi, potrebbe essere utile attingere a diversi repertori già esistenti e a dizionari strutturati come il *Wikizionario* per l'integrazione di nuove forme.

Infine, la lista che abbiamo proposto per questo lavoro può essere la fonte per studi con metodi diversi, soprattutto in ambito informatico, come ad esempio il *word embedding*,[34] un metodo che permette di individuare in automatico parole con significato simile grazie al contesto in cui appaiono e che sarebbe quindi utile a individuare nel corpus nuovi mestieri non previsti nella lista. Anche in questo caso, tuttavia, visto il numero elevato di elementi già in nostro possesso, il rischio è di incorrere in un alto numero di risultati poco pertinenti. Sarebbe tuttavia possibile impostare dei filtri, per esempio basati su dizionari strutturati, per ottenere dati più puliti.

33 Cfr. D'Achille e Grossmann 2017: 147-148.

34 Cfr. anche Sciandra, Trevisani e Tuzzi (2021).

Bibliografia

Barbieri M. (1874) *Nomenclatura italiana figurata e corredata da un appendice di 750 nomi d arti e mestieri ad uso della gioventù e delle scuole primarie d Italia. Terza edizione riveduta ed aumentata*, Torino, Paravia.

Burr E. (1995) "Agentivi e sessi in un corpus di giornali italiani", in *Dialettologia al femminile. Atti del Convegno Internazionale di Studi, Sappada/Plodn (Belluno), 26.-30.06.1995*. A cura di G. Marcato, Padova, CLUEB, pp. 349-365.

Cardinaletti A. e Giusti G. (1991) "Il sessismo nella lingua italiana. Riflessioni sui lavori di Alma Sabatini (Sexism in the Italian Language: Reflections on the Works of Alma Sabatini)", *Rassegna Italiana di Linguistica Applicata*, XXIII, pp. 169-189.

Cortelazzo M. A. (2021) "Corpora e storia della lingua", *RITT, Rivista Internazionale di Tecnica della Traduzione*, 23, pp. 179-186.

D'Achille P. e Grossmann M. (2017) "I nomi dei mestieri in italiano tra diacronia e sincronia", in *Per la storia della formazione delle parole in italiano: Un nuovo* corpus *in rete (MIDIA) e nuove prospettive di studio*. A cura di P. D'Achille e M. Grossmann, Firenze, Franco Cesati, pp. 145-181.

Enciclopedia *Sapere* DeAgostini, https://www.sapere.it/ (consultato il 30/11/2021).

Enciclopedia Treccani, https://www.treccani.it/enciclopedia/ (consultato il 30/11/2021).

Fissi A. (1983) "I nomi di mestiere a Firenze fra '500 e '600", *Studi di lessicografia italiana*, 5, pp. 53-192.

Gallo F. e Scalisi P. (2013) "Rappresentare il lavoro che cambia. Una lettura diacronica dell'osservazione statistica delle professioni", *Sociologia del lavoro*, 129, 1, pp. 40-62.

Garzoni T. (1584) *Piazza universale di tutte le professioni del mondo*, Venezia, Somasco.

Greenacre M. J. e Blasius J. (1994) *Correspondence analysis in the social sciences: Recent developments and applications*, Londra-San Diego, Academic Press.

Il Nuovo Dizionario De Mauro, https://dizionario.internazionale.it/ (consultato il 30/11/2021).

Lo Duca M. G. (1990) *Creatività e regole. Studio sull'acquisizione della morfologia derivativa dell'italiano*, Bologna, Il Mulino.

Medici M. (1967) *Nuovi mestieri e nuove professioni*, Roma, Armando.

Proietti D. (1991) *Nuovi mestieri, nuove professioni: Come orientarsi per scegliere la propria occupazione*, Roma, Sovera Multimedia.

Robustelli C. (2012) "L'uso del genere femminile nell'italiano contemporaneo: Teoria, prassi e proposte", in *Atti della X Giornata REI (Roma, 29.11.2010) 'Politicamente o linguisticamente corretto?' Maschile e femminile: usi correnti della denominazione di cariche e professioni*. A cura di M. A. Cortelazzo, Bruxelles, DG Traduzione, pp. 1-18.

Sabatini A. (1987) *Il sessismo nella lingua italiana*, Presidenza del Consiglio dei Ministri, Direzione generale delle informazioni della editoria e della proprietà letteraria, artistica e scientifica.

Sciandra A., Trevisani M. e Tuzzi A. (2021) "Sulle tracce dell'espressione dell'interiorità: analisi diacronica di un corpus di narrativa italiana del XIX-XX secolo", *RITT, Rivista Internazionale di Tecnica della Traduzione*, 23, pp. 219-233.

Straka M., Hajič J. & Straková J. (2016) "*UdPipe*: Trainable Pipeline for Processing CoNLL-U Files Performing *Token*ization, Morphological Analysis, POS Tagging and Parsing", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4290-4297. https://aclanthology.org/L16-1680.

Trevisani M. e Tuzzi A. (2013) "Shaping the history of words", in *Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO)*. Ed. by I. Obradović, E. Kelih & R. Köhler, Belgrade, Akademska Misao, pp. 84-95.

Trevisani M. e Tuzzi A. (2015) "A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal", *Quality and Quantity*, 49, pp. 1287-1304.

Trevisani M. e Tuzzi A. (2018) "Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories", *Knowledge-based systems*, 146, pp. 129-141.

Vocabolario Treccani, https://www.treccani.it/vocabolario/ (consultato il 30/11/2021).

Wikizionario, https://it.wiktionary.org/wiki/Pagina__principale (consultato il 10/12/2021).