Relazione di Introduzione alla Programmazione

Chiara Mancuso

1 Introduzione

In questo progetto si intende analizzare il rapporto tra linguaggio e realtà occupazionale in Italia, con particolare attenzione alle disparità di genere. L'obiettivo è di confrontare i dati reali forniti dall'ISTAT sull'occupazione di alcuni settori lavorativi con la frequenza delle forme linguistiche associate a queste professioni all'interno di un corpus di testi (corpus Paisà), confrontando i gender gap.
La domanda di ricerca è: *la rappresentazione linguistica delle professioni riflette la distribuzione reale dell'occupazione tra uomini e donne, oppure mostra dei bias di genere?*
Il progetto si articola in 3 fasi:

1. **Analisi dei dati ISTAT:** si analizza la distribuzione occupazionale per genere in sette settori selezionati: *Sanità e assistenza sociale; Estrazione di minerali da cave e miniere; Ricerca scientifica e sviluppo; Attività editoriali; Attività legali e di contabilità; Istruzione; Attività immobiliari*, in diverse fasi: Italia-2017; macroaree (Nord-ovest, Nord-est, Centro, Sud e Isole) nel 2017; analisi dell'evoluzione temporale (2012-2017) per tutte le macroaree.

2. **Analisi linguistica del corpus Paisà:** si sono identificate e quantificate le occorrenze dei termini legati alle professioni e settori nel corpus, con una distinzione del genere grammaticale.
   Inoltre, si è analizzata l'occorrenza delle professioni in base alle categorie di url.

3. **Comparazione tra dati reali e rappresentazioni linguistiche:** si confrontano valori di gender gap ottenuto dai dati ISTAT e dal corpus, evidenziando le eventuali discrepanze tra le due analisi.

L'obiettivo è di offrire una riflessione sulla coerenza (o non coerenza) tra linguaggio scritto e realtà sociale e, sulla presenza di eventuali bias di genere linguistici che influenzano la rappresentazione delle professioni nel linguaggio.

2 Organizzazione delle cartelle di lavoro

Per questo progetto, si è organizzato il lavoro in questo modo:

1. `Progetto`: notebook con le analisi, le operazioni sui dati e, le visualizzazioni grafiche.
2. *output_progetto*: cartella per gli output salvati, che si articola in questa maniera:
    - **df_puliti**: in cui verranno salvati entrambi i file puliti (paisa e istat), per ritrovare gli 'originali' puliti e leggibili in ogni momento.
    - **analisi_istat**: qui verranno salvati i nuovi dataset ottenuti dalle le analisi. Contiene due sottocartelle: *percentuali* e *gender gap*. Quest'ultima contiene le analisi più rilevanti, perché concernenti con i valori di gender gap.
    - **analisi_paisa**: qui verranno salvate le analisi dal corpus: sia inerenti agli url, che all'analisi testuale.
    - **analisi_comparative**: comprendente i risultati delle analisi comparative tra i dati istat e del corpus paisa.
3. **Moduli**: in questa cartella ho salvato tre miei moduli, comprendenti:
    - **euristiche**: contenente le euristiche e le espressioni regolari utilizzate per estrarre correttamente le parole dal corpus e analizzarle.
    - **mie_funzioni**: contenente le funzioni che si utilizzeranno per le analisi.
    - **classe**:contenente una classe formata da funzioni per estrarre le parole dal corpus.

Per le analisi si utilizzaranno prevalentemente funzioni della libreria Pandas, espressioni regolari, cicli for, condizioni if e Seaborn/Matplotlib.

3 Analisi istat

L"analisi dei dati istat si compone di 3 parti, in ognuna delle quali si analizzano diverse combinazioni di dati.

3.1 Preparazione all'analisi

Prima di poter iniziare l'analisi, è stato necessario integrare e ripulire i dati Istat.
I dati grezzi erano suddivisi in più file, in quanto il dataset originale aveva etichette poco leggibili. Perciò si è unito (con *.merge()* con i file che contenevano i formati leggibili inerenti a: *sesso*, *attività economica ATECO*, *area geografica* e *tipologia contrattuale*, rinominando le colonne. Si sono selezionate solo le colonne e le righe utili a questa analisi; infine, sono state create 2 colonne disaggragate per il sesso (in modo da escludere *Non specificato* e *Totale*).

In un secondo dataset, si sono infine accorpati gli anni in un'unica colonna, per facilitare analisi e visualizzazioni successive. Si sono anche calcolati le somme e le percentuali per maschi e femmine nei settori per ogni anno (usando il metodo *groupby()*). Il risultato è stato salvato nel nuovo dataset **'istat_analisi_macroaree_anni'**.

Questo nuovo dataset pulito e leggibile, sarà utilizzato come base per la maggior parte dei nuovi dataset per le analisi successive.

3.2 Analisi istat: dati inerenti all'Italia per l'anno 2017

Per questa analisi è stato filtrato il dataset 'istat_analisi_macroaree_anni' per considerare esclusivamente l'anno 2017 e la macroarea Italia.
Successivamente sono stati creati due DataFrame separati per i dati relativi ai maschi e alle femmine, poi cancatenati, con un'unica colonna per il *sesso*, in modo da facilitare la visualizzazione grafica e il confronto di genere.
Si è usato il metodo **.assign()**,per aggiungere la nuova colonna. Così si è evitato di dover effettuare due operazioni di 'melt()' distinte, per le colonne relative al sesso e alle percentuali ('.melt()' verrà utilizzato successivamente).

Per comprendere meglio la distribuzione occupazionale, si è realizzato un grafico a barre affiancate `graf_ita_2017` che mostra il numero assoluto di occupati per settori, diviso per genere. In questo modo si possono osservare situazioni di equilibrio e disparità. Infine, sono stati riportati i dati numerici di occupazione assoluta e percentuale per genere per ogni settore.

3.2.1 Analisi percentuale degli occupati per sesso e settore, rapportata al totale nazionale (2017, Italia

In questa fase si è calcolata la percentuale di lavoratori maschi e femmine impiegati in ciascun settore, rapportata con il totale nazionale degli occupati dello stesso sesso nel 2017. L'obiettivo è quantificare il peso specifico di ciascun settore all'interno dell'occupazione maschile e femminile.
Dal dataset complessivo `istat_pulito` si sono selezionate le righe relative al totale nazionale, corrispondenti a *Tutte le professioni* o *Totale*. Poi si sono raggruppati e sommati per sesso gli occupati nel 2017 (con metodo 'groupby()') per ottenere il totale nazionale maschile e femminile (per tutte le professioni). Successivamente un *merge*, in un nuovo dataset, tra i totali nazionali e il dataset `df_sesso_conc_2017_italia`, contenente le percentuali relative.

Infine, il calcolo della percentuale di occupati per settore rispetto al totale nazionale dello stesso sesso.

Infine è stato creato un nuovo dataset `df_percentuali_2017_italia`, che contiene entrmabe le percentuali (generali e settori), ordinato per valore percentuale generale (genere femminile).

Infine, si è creato un grafico a barre orizzontali affiancate (*graf_bar_perc_gen*), utilizzando il nuovo dataset, consentendo una lettura immediata delle differenza

di genere nel mercato del lavoro italiano.

3.2.2 Calcolo del Gender Gap per settore, nel 2017 in Italia

Come ultima analisi, per questo primo sub-task, si calcola il **Gender Gap** nei settori lavorativi, calcolato come differenza percentuale tra la quota maschile e quella femminile di occupati, all'interno di ciascun settore (percentuali interne al settore).
Per ciascun settore sono state selezionate le percentuali di occupazione maschile e femminile dal dataset `df_percentuali_2017_italia`. I dati sono stati inizialmente organizzati in un dizionario e, poi convertiti in un dataframe riepilogativo (usando il metodo `pd.DataFrame.from_dict`, che permette di convertire un dizionario in DataFrame). Un **valore positivo** del Gender Gap indica una prevalenza maschile all'interno del settore, mentre un **valore negativo** una prevalenza femminile.
Successivamente si visualizzano questi dati nel grafico `graf_barh_gender_gap_settore_sorted` a barre orizzontali, utilizzando il nuovo dataframe (df_gender_gap_2017_italia), ordinato per gender gap. Le barre a destra della linea dello 0 (azzurre) indicano una prevalenza maschile; al contrario, rosse, indicano una prevalenza femminile, per visualizzare i settori con maggior gap.

3.3 Analisi dei dati inerenti alle Macroaree per l'anno 2017

Come secondo sub-task (segue molto la logica del primo), si sono esaminati i dati relativi agli occupati suddivisi per macroarea geografica, sesso e settore di lavoro, per l'anno 2017.
Le macroaree considerate sono: *Nord-ovest, Nord-est, Centro, Sud e Isole*. Creato un dataset adatto alle analisi,`df_sesso_conc_2017_macroaree`, in cui ho una sola colonna per il sesso e per le percentuali, per ogni macroarea (concatenato).
Successivamente per ogni macroarea, sono stati prodotti grafici a barre (`graf_istat_macroaree_2017`) che rappresentano le percentuali di occupati maschili e femminili in ciascun settore lavorativo, per evidenziare possibili differenze di genere nelle macroaree. *Per realizzarlo, si è iterato su tutto il dataset, con un ciclo for: un grafico per ogni macroarea.*

3.3.1 Analisi dei valori percentuali di occupazione rapportati al totale nazionale

Seguendo la logica di Italia-2017, si è costruito un dataframe riassuntivo, con i dati inerenti alle percentuali per macroarea e settore (`df_percentuali_2017_macroaree`), contenente le percentuali interne ai settori e quelle generali. Infine il grafico a barre orizzontali, per rappresentare visivamente i risultati ottenuti (graf_istat_macroaree_2017).

3.3.2 Calcolo del Gender Gap (macroaree, 2017)

Infine, si è calcolato il Gender Gap come divario percentuale di occupazione tra maschi e femmine, per ciascun settore lavorativo, in ogni macroarea. Un valore positivo indica una prevalenza maschile, mentre un valore negativo indica una predominanza femminile. Creando il nuovo dataset `df_gender_gap_2017_macroaree`, ordinato per grandezza del gender gap. Successivamente si sono visualizzati i risultati ottenuti, in un grafico a barre orizzontali, per cui se un valore è positivo sarà a destra della linea dello 0, altrimenti sarà a sinistra (ad ogni macroarea è stato assegnato un colore diverso).

3.4 Evoluzione del Gender Gap per Settore e Macroarea (2012-2017)

In questa ultima parte di analisi, si esamina l'andamento occupazionale dal 2012 al 2017 distinguendo per *Macroarea, Settore di lavoro e Genere*. L'obiettivo è osservare l'andamento (aumento o riduzione) del gender gap nel tempo per queste macroaree e settori.
Si utilizza una funzione che si trova nel modulo `mie_funzioni`.

3.4.1 La funzione

La funzione `calcola_gender_gap_settore`, ha lo scopo di confrontare le percentuali di occupazione maschile e femminile tra due anni di riferimento (2012 e 2017).

Prende in input un sottoinsieme del DataFrame contenente i dati relativi a un singolo settore (filtrato per macroarea) e restituisce un dizionario con: *le percentuali di occupazione maschile e femminile nei due anni*; *i rispettivi Gender Gap*; *la variazione del Gap tra i due anni (punti %)*; *indicatore sulla direzione della variazione del gap: aumentato oppure diminuito.*

La funzione effettua prima un controllo sulla disponibilità dei dati per entrambi gli anni. Se sono disponibili, procede con il calcolo dei Gender Gap nei due anni e della variazione. Il risultato viene restituito come dizionario.

3.4.2 Evoluzione del gender gap per le macroaree (Italia esclusa)

Inizialmente l'analisi considera soltanto le macroaree, utilizzando la funzione `calcola_gender_gap_settore`. Infine, i risultati sono stati ordinati per maggiore variazione del gap.
Si sono poi rappresentati graficamente i risultati, in 2 grafici:

1. **Grafico a barre**: per evidenziare la variazione in punti percentuali del gap: più la barra è lunga, più la variazione sarà ampia.

2. **Grafico a linee**: per mostrare l'andamento del gap nei settori nel tempo in ogni macroarea.

3.4.3 Confronto con l'Italia

Infine, è stato effettuato un confronto tra questi dati e quelli per l'Italia (utilizzando la funzione anche qui). Successivamente si è creato un nuovo dataset, concatenando i due (per macroarea e Italia): `df_gender_gap_macroaree_italia_anni`.

Infine è stato creato un altro dataset `confronto_finale_gap_italia_macroaree`, risultato di un *.merge()* tra dataset minori, in cui sono presenti i settori, e i valori di gap, divisi per anno, macroarea, Italia e variazione dall'Italia, per poter confrontare i dati ottenuti. In questo modo, sarà presente l'Italia solo come confronto.
Infine si sono visualizzati graficamente i risultati con un **grafico a linee** (uno per ogni settore di lavoro), in cui si avrà: 1 linea per macroarea e, una per l'Italia (nera tratteggiata), per osservare se i trend delle macroaree sono in linea o meno con quelli dell'Italia.

4 Analisi Paisa

4.1 Normalizzazione del corpus

Come prima operazione si procede con la normalizzazione del testo del corpus originale (utilizzando la mia_funzione `normalize`).

Funzione `Normalize`: conversione in minuscolo; sostituzione degli apostrofi con spazi; rimozione della punteggiatura e dei caratteri speciali sostituiti con spazi; rimozione degli spazi superflui all'inizio e alla fine; tokenizzazione base (*.split()*); rimozione delle stopwords per focalizzarsi su termini significativi. Infine, si ricompongono i token (*.join()*) per avere un testo in formato stringa.
Il nuovo testo normalizzato viene salvato nel file csv (`paisa_pulito`).

4.1.1 Assegnazione delle categorie tematiche agli URL

Durante l'analisi preliminare del corpus, si è riscontrata la presenza di url contenenti messaggi di errore nei testi, come *"wwwoffle"*. Si sono identificati questi casi, definendo un'espressione regolare che filtra le righe con questi messaggi. Gli URL errati sono stati etichettate come sospetti e utilizzati per creare un filtro *booleano* e, una colonna apposita.

Ogni URL viene poi associato a una categoria, che si trova nel modulo *euristiche*. Sono state assegnate con la funzione ***assegna_categoria***, che riceve in input un URL e un dizionario di categorie con espressioni regolari e, controlla se l'URL corrisponde a uno dei pattern specificati nel dizionario e, in tal caso, restituisce la categoria corrispondente. Se nessun pattern corrisponde, viene restituita la categoria generica `'altro'`.

Infine, si crea il nuovo dataframe `paisa_analisi`, che sarà utilizzato per tutte le analisi del corpus.

4.2 Estrazione di professioni, settori e contesti dal corpus PAISA

Dal corpus normalizzato, è stata condotta un'analisi lessicale per l'individuazione delle professioni, settori di lavoro e dei rispettivi contesti d'uso.
L'estrazione viene effettuata con l'uso della classe `EstrazioneProfessioniSettori`; le informazioni estratte sono state organizzate in tre nuove colonne, rispettivamente per professioni, settori e contesti.

4.2.1 La classe `EstrazioneProfessioniSettori`

La classe `EstrazioneProfessioniSettori` si basa su tre funzioni fondamentali:

- **estrai_professioni_e_settori**: riceve in input un testo e restituisce, tramite espressioni regolari, tutte le professioni e i settori individuati, dopo aver applicato filtri sulle parole proibite (contenute nel modulo `euristiche`).
- **filtra_match_per_contesto**: verifica se una determinata parola si trova in un contesto ritenuto valido. Analizza una finestra di tre parole prima e dopo la parola target, escludendo i match in cui compaiono elementi proibiti.
- **estrai_contesto**: estrae una finestra contestuale più ampia (5 parole prima e dopo) attorno alle professioni trovate (occorrenze), partendo dai token e, restituendo la stringa ricompattata (utile per la raffinazione delle espressioni regolari).

L'architettura della classe permette un controllo flessibile sulle procedure di estrazione, in maniera da non dover richiamare ogni volta tutti gli argomenti necessari alle funzioni.

4.2.2 Euristiche per l'estrazione

Per ogni settore di lavoro è stata costruita una lista di espressioni regolari che tengono conto del genere e numero dei termini, evitando però forme ambigue (in formato dizionario Settore:[professioni]) Queste espressioni sono state compilate e organizzate in un dizionario che associa ciascun settore a una serie di pattern linguistici. Si è utilizzato il metacarattere `\b` per assicurarsi che i match corrispondessero a parole intere.

Oltre ai pattern, sono state definite per ciascun settore delle **parole proibite**, per escludere falsi positivi derivanti dal contesto della parola. Ad esempio, per la professione "medico" si escludono contesti come 'visita medica' o 'esame medico'.

Queste euristiche rappresentano una componente fondamentale del sistema, poiché permettono di aumentare la precisione dell'analisi linguistica (inizialmente non avevo filtrato *casa editrice* e, risultavano circa 1000 occorrenze, dopo aver filtrato grazie alle 'parole proibite', sono diminuite sensibilmente).

4.2.3 Conteggio delle professioni nel corpus PAISA

Terminata l'estrazione, si sono contate le professioni nel corpus. Sono stati aggregati tutti i risultati contenuti nella colonna `professioni_estratte` e, utilizzando *Counter*, è stata calcolata la frequenza di ciascuna professione.

Questa operazione ha permesso di identificare i termini più ricorrenti nel corpus in riferimento alle professioni, successivamente visualizzati con un grafico (che discrimina anche per femminili e maschili).

Infine, si sono esplorati i contesti linguistici in cui le professioni compaiono. Per ogni occorrenza nella colonna `contesti_estratti`, si è visualizzata una selezione dei contesti associati, in modo da verificarne la coerenza e l'adeguatezza.

4.3 Stima del genere delle professioni e analisi delle frequenze

Per valutare la distribuzione di genere delle professioni estratte, si è stimato il genere grammaticale basandosi sui suffissi delle parole rappresentanti le professioni, tramite la funzione `stima_genere_solo_suffisso`.

I generi stimati sono stati quindi aggiunti come nuova colonna, per mantenere l'associazione tra professione e genere. Poi si aggregano le professioni raggruppandole in base al genere, grazie alla funzione `conta_professioni_per_genere`.

I risultati sono stati visualizzati con un confronto grafico tramite istogrammi a barre orizzontali che mostra le professioni più frequenti per ciascun genere; e, un grafico a barre combinato che confronta la frequenza assoluta delle professioni in ambito maschile e femminile.

4.3.1 Le funzioni

- `conta_professioni_per_genere`: la funzione prende in input un DataFrame che contiene i dati estratti dal corpus, in particolare le colonne con i settori associati alle professioni e i generi stimati delle stesse. Esegue un ciclo su ogni riga e verifica che le colonne interessate contengano effettivamente valori. Per ogni elemento, estrae la professione e il genere corrispondenti, e aggiorna un dizionario per i conteggi che tiene traccia del numero di occorrenze di professioni maschili e femminili. Se la professione non è ancora presente nel dizionario, viene inizializzata con due contatori a zero (uno per genere), che vengono poi incrementati ad ogni occorrenza. Infine, la funzione restituisce questo dizionario, permettendo di analizzare la distribuzione per genere nelle diverse professioni.

- `stima_genere_solo_suffisso`: funzione che stima il genere grammaticale di una parola basandosi esclusivamente sul suo suffisso. La funzione prende in input una parola e verifica se questa termina con uno dei suffissi tipici del genere grammaticale, definiti in due liste separate. Se il suffisso corrisponde a uno di quelli femminili, la parola viene classificata come *femminile*; se corrisponde a uno dei suffissi maschili, è classificata come *maschile*. Nel caso in cui il suffisso non corrisponda a nessuna delle liste, la funzione restituisce `None`.

4.4 Estrazione e analisi dei settori lavorativi per genere

Avendo già estratto inizialmente i settori, si effettua direttamente il conteggio delle occorrenze divise per genere: cioè si distinguono quante professioni di genere maschile e femminile sono associate a ciascun settore, grazie alla funzione `conta_settori_per_genere()`.

Per ogni settore vengono estratti i conteggi assoluti per il genere maschile e femminile e, successivamente, viene calcolata la percentuale di occorrenza dei settori.

Viene poi creato il dataset `df_paisa_analisi_settori` riassuntivo, usato per rappresentare i valori delle percentuali tramite un grafico a barre orizzontali sovrapposte, che mostra la distribuzione percentuale di genere delle parole associate ai diversi settori.

4.4.1 Funzione `conta_settori_per_genere`

La funzione `conta_settori_per_genere()` ha lo scopo di contare quante volte, all'interno del dataframe di analisi, compaiono professioni associate a ciascun *settore lavorativo*, suddividendo il conteggio in base al genere grammaticale (maschile o femminile).

Questa funzione è molto simile, per struttura e logica, alla precedente funzione `conta_professioni_per_genere()`.

Viene quindi creato un dizionario vuoto chiamato, in cui per ogni settore incontrato verranno il numero di occorrenze maschili e quelle femminili. Si itera su tutte le righe del dataframe in input, accedendo per ciascuna riga alle liste di (`settori_estratti`) e ai generi stimati delle professioni associate. Dopo un controllo dei valori all'interno delle liste (se dati mancanti o non corretti, la riga viene saltata). Se il settore è definito e il genere è maschile o femminile, si aggiorna il dizionario incrementando il contatore corrispondente. Alla fine, la funzione restituisce il dizionario contenente per ogni settore i totali delle occorrenze divise per genere.

4.5 Calcolo del Gender Gap per settore

Per concludere l'analisi della distribuzione di genere nel corpus, si è calcolato il *gender gap*, definito come la differenza tra la percentuale di occorrenze maschili e quella femminili per ciascun settore: valori positivi indicano una prevalenza

maschile, mentre valori negativi un predominio femminile. Creata una nuova colonna chiamata `Gender Gap`, il dataframe è stato ordinato in base a questo valore per facilitare l'interpretazione e la visualizzazione, tramite un grafico a barre orizzontali, con valori di gap decrescenti.

4.6 Analisi degli URL associati alle professioni

Alla fine dell'analisi del corpus, si è esplorata la distribuzione delle professioni estratte all'interno delle diverse categorie di siti web. Per ogni riga del dataset, sono state recuperate la lista delle professioni e la categoria dell'URL associato al testo. È stato quindi costruito un dizionario che, per ogni professione, registra il numero di occorrenze nelle varie categorie di URL. Al termine, per ogni professione, sono state stampate le distribuzioni delle occorrenze per categoria.

5 Analisi Comparativa ISTAT-PAISA

In questo ultimo task, si confrontano i risultati ottenuti dall'analisi del corpus **PAISA** con i dati dell'**ISTAT** per l'anno 2017.

Per verificare se e in quale misura il linguaggio usato nel corpus rifletta la distribuzione occupazionale reale tra uomini e donne nei diversi settori lavorativi.

5.1 Confronto tra Gender Gap ISTAT e PAISA (Italia)

Per condurre il confronto è stato necessario armonizzare i due dataset, rinominando innanzitutto le colonne, per renderle coerenti.

Dopo questa fase di preparazione, i due dataset sono stati uniti tramite un'operazione di `merge` sulla colonna `Settore`. Il risultato è il DataFrame `comparazione_gender_gap_istat_paisa_italia`,che verrà utilizzato per analisi successive.

Poi, per realizzare un grafico a barre comparativo efficace, è stato applicato un `melt` al DataFrame, per appaiare i due valori di Gender Gap in un'unica colonna con l'indicazione della fonte (`ISTAT` o `PAISA`).

5.2 Analisi per Macroaree: confronto ISTAT - PAISA

Dopo aver effettuato l'analisi a livello nazionale, si è passati ad un'analisi per macroaree geografiche.

Per confrontare questi dati con le percentuali linguistiche del corpus PAISÀ, è stato eseguito un `merge` sui settori, senza tenere conto della macroarea, poiché PAISÀ non fornisce distinzione geografica.

Il DataFrame ottenuto, `comparazione_gender_gap_macroaree`, è stato utilizzato per produrre una serie di grafici comparativi. Per ciascuna macroarea è stato generato un grafico a barre in cui, per ogni settore, sono visualizzati il `Gender Gap ISTAT` e il `Gender Gap PAISÀ`.

5.3 Analisi complessiva: Macroaree e Italia

Per ottenere una visione più completa, sono stati uniti i dati relativi all'intera Italia con quelli delle singole macroaree, in modo da poter confrontare il Gender Gap sia a livello nazionale sia regionale.

Infine, è stata calcolata la media del Gender Gap per ciascun settore, considerando i dati ISTAT e quelli derivanti dal corpus linguistico PAISA. Questo consente di evidenziare le divergenze tra la realtà occupazionale ufficiale e la rappresentazione linguistica.

5.4 Conclusioni

Alla fine di questa analisi, emerge che il corpus PAISA tende a sovrastimare la presenza maschile rispetto ai dati ufficiali ISTAT, soprattutto nei settori dove le donne sono maggioranza. Solo in settori tradizionalmente maschili, come l'estrazione mineraria, si osserva una buona corrispondenza tra dati reali e linguaggio. Questo fenomeno potrebbe riflettere la persistenza di bias culturali e linguistici che privilegiano le forme maschili nel discorso scritto. Sarebbe interessante un approfondimento, ma già questo è un risultato più che interessante!