IncluInstIT: Un nuovo corpus per lo studio di linguaggio inclusivo su Instagram

Irene Caiazzo[1], Giovanna Maria Dimitri[2], Liana Tronci[3]
[1] Università per Stranieri di Siena, Italia – i.caiazzo@dottorandi.unistrasi.it
[2] Università degli Studi di Siena, Italia - giovanna.dimitri@unisi.it
[3] Università per Stranieri di Siena, Italia - tronci@unistrasi.it

ABSTRACT (ITALIANO)
Il contributo presenta una ricerca, tuttora in corso, di raccolta e analisi del corpus IncluInstIT, composto da forme di linguaggio inclusivo di genere presenti nella piattaforma social Instagram. Il corpus è stato elaborato attraverso metodologie semiautomatiche sia durante la fase di iniziale di raccolta dei post per la creazione del dataset, che nel successivo trattamento dei dati scaricati per selezionare i post pertinenti alla ricerca e rendere tutti i contenuti testuali - sia delle didascalie sia quelli contenuti nelle immagini - analizzabili da software di linguistica dei corpora.
Lo scopo del progetto è, da un lato, indagare come un luogo di dialogo digitale quale è Instagram possa essere un contenitore di usi linguistici innovativi, che si suppone non si trovino in tale quantità e varietà in altre piattaforme né in altre tipologie testuali (per esempio articoli di giornale o produzioni letterarie, ma neanche in produzioni orali) e, dall'altro, creare una metodologia di raccolta dati per costruire dataset relativi a casi di studio di linguaggio inclusivo che possa essere replicabile, anche per altre piattaforme e per altre lingue.
Parole chiave: linguaggio inclusivo; sessismo linguistico; linguistica dei corpora; italiano digitale

ABSTRACT (ENGLISH)
IncluInstIT: A New Corpus for Studying Inclusive Language on Instagram
The paper presents an ongoing research project focused on the collection and analysis of the IncluInstIT corpus, which comprises instances of gender-inclusive language found on the social media platform Instagram. The corpus was developed using semi-automatic methodologies both during the initial phase of post collection for dataset creation and in the subsequent processing of downloaded data to select posts relevant to the research. These methods also ensured that all textual content—whether captions or text embedded in images—was rendered analyzable by corpus linguistics software.
The project purses two primary goals: first, to investigate how a digital space for dialogue like Instagram can serve as a repository for innovative linguistic practices, which are hypothesized to occur in a quantity and variety not typically found on other platforms or in other textual genres (e.g., newspaper articles, literary productions, or even oral discourse); and second, to develop a data collection methodology for constructing datasets on instances of inclusive language that can be replicated for other platforms and languages.
Keywords: inclusive language; linguistic sexism; corpus linguistics; digital Italian

1. STATO DELL'ARTE
Il linguaggio inclusivo di genere - che può essere definito come l'uso di strategie linguistiche per rappresentare, attraverso la lingua, le diverse identità di genere - è ormai stato ampiamente trattato nei suoi aspetti teorici, sia per quanto riguarda la rappresentazione del femminile (a partire almeno da Sabatini 1986, 1987 e Violi 1986 fino a lavori più recenti come Giusti 2022, Fusco 2024, Robustelli 2024), sia con riferimento al tema del superamento del binarismo di genere e alla rappresentazione delle identità non binarie (per una panoramica sulla teorizzazione del linguaggio inclusivo in italiano cfr. Gheno 2022, Formato e Somma 2023, Formato 2024). Accanto alla riflessione teorica, quest'ultimo aspetto ha avuto anche dei risvolti pratici, con le proposte di implementazione delle forme di espressione linguistica finalizzate alla neutralizzazione del genere attraverso l'uso di segni quali asterisco (*), chiocciola (@), trattino basso (_), *-x*, o di possibili suoni della lingua quali *-u* o schwa, anche distinto per singolare (*-ə*) e plurale (*-3*) (cf. Comandini 2021 per una panoramica). Al momento, però, sono ancora pochi gli studi che analizzano come queste discussioni si riflettano nella pratica quotidiana della lingua. Comandini (2021) ha creato il primo corpus italiano (CoGeNSI) annotando manualmente testi prodotti su pagine Facebook queer al fine di analizzare le strategie di neutralizzazione del genere nelle loro regolarità e irregolarità e dare una prima tassonomia delle modalità d'uso di queste strategie dividendole in (a) riferimento alla prima

persona, (b) riferimento ad altre persone, (c) uso politico, ovvero l'uso che alcune persone o organizzazioni fanno di queste strategie per posizionarsi rispetto alle tematiche di inclusione.
Facendo tesoro dello studio di Comandini (2021) abbiamo deciso di creare un corpus coerente dal punto di vista del tipo di interrogazione ma diverso nelle forme e nei metodi di raccolta. Innanzitutto, è diversa la scelta del social media di riferimento: abbiamo scelto Instagram invece di Facebook in quanto, sebbene sia nato come piattaforma di condivisione di immagine, presenta anch'esso una componente linguistica - la didascalia, da sempre presente nella struttura del post - e ad oggi presenta contenuti non solo fotografici e autobiografici, ma anche di carattere divulgativo, anche nell'ambito del linguaggio inclusivo.[1] Un aspetto che rende di interesse la piattaforma per il nostro argomento è la sua demografia, più giovane rispetto a Facebook, e tendenzialmente più interessata e aperta a queste tematiche.[2] L'ultima ragione che ha spinto a questa scelta sono le modalità di raccolta dei post offerte dall'API utilizzata, *instaloader*[3]: la libreria infatti permetteva di non dover concentrare l'attenzione su pagine specifiche - come fatto nello studio di Comandini (2021) ma ha dato la possibilità di avere un ampio numero di autori di post, avendo scelto come filtro di selezione la presenza di hashtag e quindi potendo fare un'analisi che raccogliesse il maggior numero di utenti possibile.

2. MODALITÀ DI RACCOLTA

La raccolta dei post è stata eseguita seguendo le indicazioni contenute in Di Cristofaro (2023), dove si individua l'API *instaloader* come strumento di download dei post. Questa API consente di ottenere post attraverso diversi *target* di ricerca che nella loro forma base sono: *profile*, *hashtag* e *location* a cui si possono poi aggiungere ulteriori specificatori come *stories, tagged, comments* etc. (Di Cristofaro 2023, 208). Dunque è stato creato un profilo Instagram apposito per la raccolta dati (@accountphd), in quanto per utilizzare l'API era necessario essere loggati in un profilo e la modalità selezionata è stata la ricerca per hashtag e sono stati scaricati post che contenessero uno dei 21 hashtag selezionati inerenti all'argomento in esame: #femminileinclusivo, #femminilesovraesteso, #femminiliprofessionali, #generelinguistico, #italianoinclusivo, #linguaggioampio, #linguaggiodigenere, #linguaggioinclusivo, #linguaggioneutro, #linguaggiononsessista, #linguaggiosessista, #linguainclusiva, #linguaneutra, #linguanonbinaria, #linguasessista, #maschileinclusivo, #maschileneutro, #maschilenonmarcato, #maschilesovraesteso, #schwa, #sessismolinguistico.
La scelta degli hashtag da usare è stata fatta tenendo conto dei possibili hashtag legati alla tematica presa in esame e verificando l'esistenza di post che li contenessero sulla piattaforma. Oltre ai 21 hashtag elencati, ne sono stati testati altri che non hanno portato al download di alcun post ad esempio #linguaggiogenderizzato, #linguanonsessista, #linguaggiononbinario. Come principio generale per ogni hashtag scaricato che conteneva la parola 'maschile' si è cercato anche il corrispondente con 'femminile' allo stesso modo in cui si sono cercati sia hashtag con 'lingua' che con 'linguaggio'. Per necessità date dalle possibilità materiali e dalle modalità di processing che si sarebbero andate a utilizzare, sono stati esclusi hashtag che avrebbero previsto un numero particolarmente alto di post e dei quali un'alta percentuale non sarebbe stata strettamente attinente alla tematica (ad es. #asterisco, #chiocciola, #ministra, #femminile). Si può vedere l'hashtag #schwa come un'eccezione a questo principio, visto che presenta circa 8400 post, si è tuttavia deciso di conservarlo, considerando che l'uso di questo termine in italiano è strettamente legato al dibattito sul linguaggio inclusivo. Di conseguenza, il successivo filtraggio per lingua avrebbe eliminato i post non pertinenti, mentre quelli rimanenti sarebbero stati rilevanti per il corpus.
Per ogni post la libreria ha scaricato diversi file contenenti tutti i dati ottenibili di default dai post: un file JSON contenente i metadati (tra cui codice univoco del post, il suo url, profilo che ha pubblicato il post, data e ora, numero di "mi piace" e di commenti), un file .txt contenente la didascalia del post, un numero di file JPG o PNG corrispondente al numero di immagini presenti nel post e un file MP4 per i post contenenti video.

---

[1] La presenza di account che si occupano di divulgazione è stata già notata in diversi lavori, tra cui Bagaglini 2022 e 2024, quest'ultimo con riferimento diretto alla divulgazione di contenuti inerenti alla linguistica.

[2] Dati sull'utenza di Instagram e Facebook si possono trovare ai link: https://www.statista.com/statistics/248769/age-distribution-of-worldwide-instagram-users/ (cons: 10/01/2025)
https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/ (cons: 10/01/2025)
Invece, riguardo alle tendenze delle diverse generazioni rispetto alle tematiche inerenti alle comunità queer: https://www.ipsos.com/it-it/pride-month-2024-genz-propensa-identificarsi-lgbt (cons: 10/01/2025)

[3] https://github.com/instaloader/instaloader (cons: 10/01/2025)

Questo download è durato dal 27/05/2024 al 13/07/2024 ha raccolto file senza limiti cronologici, quindi dal 2011, anno di creazione del social, al giorno stesso dell'ultimo download. La fase di raccolta si è conclusa con l'ottenimento di 12905 post scaricati che sono andati a comporre il dataset su cui si costruirà il corpus. Il numero di post è stato determinato in base al numero di file JSON, mentre i numeri degli altri tipi di file variano; sono stati scaricati 12928 documenti in formato .txt (la differenza tra numero di post e di file .txt, quindi didascalie, si spiega tenendo conto della presenza di didascalie modificate dagli autori dei post successivamente alla loro pubblicazione e che sono state ugualmente scaricate dal programma con la dicitura .old), 25105 in formato JPG, 61 in formato PNG, 2 immagini in formato WEBP (in seguito convertito in JPG per essere aperto dal computer) e 1207 file video in formato MP4.

3. PRE-PROCESSING DEL DATASET

Le modalità della prima fase presentavano dei limiti che hanno fatto sì che fosse necessario svolgere diverse fasi di ripulitura dei dati prima che questi potessero essere analizzati con software di linguistica dei corpora. Questi limiti consistevano principalmente nella presenza di post duplicati, fatto causato dalla presenza di due o più hashtag tra i selezionati e quindi che l'API ha scaricato più di una volta, e di post in lingue diverse dall'italiano.
In questa fase sono stati utilizzati solo i file .txt dei post, i quali sono stati sottoposti a processi sia automatici che manuali per la rimozione dei file non desiderati all'interno del dataset (ad esempio utilizzo del programma *remove duplicate lines*[4] e della API *langdetect*[5]) e che hanno fatto diminuire il numero di post a 4834. Di conseguenza sono diminuiti anche gli altri file a: 11628 JPG, 40 PNG, 331 MP4 e 1 WEBP. La presenza di post duplicati si spiega per la tendenza di inserire più di un hashtag nel post per segnalare l'argomento di cui si sta parlando, caratteristica tipica della piattaforma come si evince da una lettura dei metadati dei post scaricati in cui ci sono anche informazioni sugli hashtag utilizzati. La presenza di lingue straniere invece ha più di una ragione: in alcuni casi è dovuta alla presenza delle stesse parole anche in altre lingue ad esempio con gli hashtag #schwa e #linguaneutra (in questo secondo caso si è rilevato che tutti i post scaricati erano in lingua portoghese e non italiana e quindi sono stati rimossi), in altri invece il post tratta di lingua italiana ma lo fa in una lingua straniera a scopo spesso divulgativo, sono state trovate infatti post di alcune pagine di italiano a stranieri o che diffondono più in generale la cultura italiana all'estero.

4. FASE ATTUALE: GESTIRE LA MULTIMODALITÀ

La fase in corso ha lo scopo di recuperare il contenuto testuale presente all'interno delle immagini che, insieme alle didascalie, compongono i post che comporranno il corpus. Sebbene la ricerca non voglia prendere una svolta multimodale, e quindi più semiotica che strettamente linguistica, si ritiene che tenere conto dei contenuti linguistici nelle immagini sia necessario per tenere conto della caratteristica fondante della piattaforma di riferimento, ovvero il ruolo centrale che ha l'immagine - che sia una foto o che riporti una scritta - nella totalità del post.
Il primo passo per ottenere questo contenuto testuale è stato l'estrazione attraverso l'OCR *tesseract*[6] che ha permesso di creare un file .txt per ogni immagine che gli è stata sottoposta. Questi file sono stati successivamente divisi in file vuoti e file che presentavano contenuto ed è seguito il controllo manuale dei testi creati, per controllare fossero corretti.
Le principali problematiche di questa fase sono state: la difficoltà del software di recuperare il testo in immagini che spesso presentavano molto rumore di sottofondo (ad es. testo sovrapposto ad altre immagini) oppure con font non convenzionali o scrittura a mano e creare delle categorie per distinguere quali contenuti testuali fossero di interesse per la ricerca e quali no.

5. FASI FUTURE E OBIETTIVI DI ANALISI

Le fasi successive della ricerca si svilupperanno in diverse direzioni al fine di indagare in modo completo e in funzione dei vari parametri il corpus creato.
In primo luogo, si svolgerà un'analisi quantitativa dei testi presenti nel corpus, per verificare le tendenze in funzione della frequenza delle forme contenute e delle loro possibili correlazioni, al fine di comprendere quali temi entrano in gioco su Instagram quando si discute di linguaggio inclusivo e in quali modi se ne

[4] https://codebeautify.org/remove-duplicate-lines (cons: 10/01/2025)
[5] https://github.com/Mimino666/langdetect (cons: 10/01/2025)
[6] https://github.com/tesseract-ocr/tesseract (cons: 10/01/2025)

discute. Successivamente, avverrà una marcatura delle forme che presentano strategie di linguaggio inclusivo, al fine di indagarne gli usi e metterli in relazione con altri studi (per es. Comandini 2021). La disponibilità di post che coprono un arco temporale ampio permetterà anche di studiare se e come queste strategie sono state utilizzate nel passato e come vengono usate oggi, se ci sono stati mutamenti nel loro uso e, ad esempio, se le preferenze nell'uso di una o di un'altra forma sono cambiate.
A fianco a questo lavoro diretto sui testi, avverrà anche un'analisi dei metadati raccolti, così da studiare dati come la cronologia dei post, le tendenze nell'uso di hashtag, il numero e il tipo di utenti che hanno pubblicato questi post.
Un'ulteriore linea di ricerca, per ora solo ipotizzata, è l'analisi dei commenti presenti in questi post. Infatti grazie ai metadati già scaricati è possibile vedere quali post hanno ricevuto commenti e in che quantità, tramite la libreria già utilizzata nella raccolta dei post, *instaloader,* si potrebbero scaricare anche i relativi commenti e quindi vedere come anche in questo genere di testo web si usano strategie inclusive e, soprattutto, un tipo di analisi qualitativa potrebbe far emergere le reazioni e le opinioni di un maggior numero di utenti rispetto alla tematica.

BIBLIOGRAFIA

Bagaglini, V. (2022). «LA COMUNICAZIONE SCIENTIFICA SUI SOCIAL NETWORK: UN'ANALISI DELLA SCRITTURA DIVULGATIVA SU TWITTER, FACEBOOK E INSTAGRAM». Italiano LinguaDue 13 (2):310-35. https://doi.org/10.54103/2037-3597/17141.

- 2024. «La Divulgazione Della Linguistica Su Instagram: "La Polifonia Del Discorso Specialistico"». Lingue E Culture Dei Media 7 (1-2):38-67. https://doi.org/10.54103/2532-1803/22392.

Comandini, G. (2021). «Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. : Indagine su un corpus di italiano scritto informale sul web.» Testo e Senso, fasc. 23 (23 dicembre 2021): 43–64.

Di Cristofaro, M. (2023). Corpus Approaches to Language in Social Media. 1ª ed. New York: Routledge,. https://doi.org/10.4324/9781003225218.

Formato, F., Somma A. L. (2023). «Gender Inclusive Language in Italy: A Sociolinguistic Overview» The Journal of Mediterranean and European Linguistic Anthropology, Vol. 5(1) (2023): 22–40. DOI: 10.47298/jomela/v5-i1-a3 jomela.pub

Formato, F. (2024). Feminism, corpus-assisted research, and language inclusivity. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009236379

Gheno, V. (2022) Femminili singolari: il femminismo è nelle parole*.* Quarta edizione. Saggi pop 60. Firenze: Effequ.

Giusti, G. (2022). «Inclusività nella lingua italiana: come e perché. Fondamenti teorici e proposte operative», DEP*.* Deportate, esuli, profughe n. 48 / 2022: 1-19.

Ipsos. 2024. Pride Month 2024: la GenZ è la più propensa a identificarsi come LGBT+
https://www.ipsos.com/it-it/pride-month-2024-genz-propensa-identificarsi-lgbt

instaloader
https://github.com/instaloader/instaloader

langdetect
https://github.com/Mimino666/langdetect

remove duplicate lines
https://codebeautify.org/remove-duplicate-lines

Robustelli, C. (2024) «Sessismo Linguistico». Tindara A. et al. (Eds.), Promuovere la parità di genere nelle istituzioni pubbliche e private, 2024, 171–91.

Sabatini, A. (1986). Raccomandazioni per un uso non sessista della lingua italiana. Per la scuola e l'editoria scolastica. Roma: Presidenza del Consiglio dei ministri.

- (1987) Il sessismo nella lingua italiana. Roma: Presidenza del Consiglio dei ministri.

Statista (2024a). Distribution of Facebook users worldwide as of April 2024, by age and gender
https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/

- (2024b). Distribution of Instagram users worldwide as of April 2024, by age and gender
https://www.statista.com/statistics/248769/age-distribution-of-worldwide-instagram-users/

tesseract
https://github.com/tesseract-ocr/tesseract

Violi, P. (1986). L’infinito singolare. Considerazioni sulle differenze sessuali nel linguaggio. Verona: Essedue, 1986.